

KATHOLIEKE UNIVERSITEIT LEUVEN
FACULTEIT TOEGEPASTE WETENSCHAPPEN
DEPARTEMENT ELEKTROTECHNIEK
Kasteelpark Arenberg 10, 3001 Leuven (Heverlee)

INTEGRATIVE ANALYSIS OF DATA, LITERATURE, AND EXPERT KNOWLEDGE BY BAYESIAN NETWORKS

Promotor:
Prof. dr. ir. B. De Moor
Prof. dr. Y. Moreau

Proefschrift voorgedragen tot
het behalen van het doctoraat
in de toegepaste wetenschappen

door

Péter Antal

28 January 2008

KATHOLIEKE UNIVERSITEIT LEUVEN
FACULTEIT TOEGEPASTE WETENSCHAPPEN
DEPARTEMENT ELEKTROTECHNIEK
Kasteelpark Arenberg 10, 3001 Leuven (Heverlee)

INTEGRATIVE ANALYSIS OF DATA, LITERATURE, AND EXPERT KNOWLEDGE BY BAYESIAN NETWORKS

Jury:
Prof. dr. ir. Y. Willems, voorzitter
Prof. dr. ir. B. De Moor, promotor
Prof. dr. Y. Moreau, promotor
Prof. dr. ir. S. Van Huffel
Prof. dr. D. Timmerman
Prof. dr. ir. J. Vandewalle
Prof. dr. T. Dobrowiecki (TUB)

Proefschrift voorgedragen tot
het behalen van het doctoraat
in de toegepaste wetenschappen

door

Péter Antal

U.D.C. 681.3*I2, 681.3*I21 28 January 2008

© Katholieke Universiteit Leuven – Faculteit Toegepaste Wetenschappen
Arenbergkasteel, B-3001 Heverlee (Belgium)

Alle rechten voorbehouden. Niets uit deze uitgave mag vermenigvuldigd en/of openbaar gemaakt worden door middel van druk, fotocopie, microfilm, elektronisch of op welke andere wijze ook zonder voorafgaande schriftelijke toestemming van de uitgever.

All rights reserved. No part of the publication may be reproduced in any form by print, photoprint, microfilm or any other means without written permission from the publisher.

D/2007/7515/99

ISBN 978-90-5682-865-3

Foreword

In my research I crossed many borders between systems, countries, disciplines, and between the industry and the academy. Therefore, I am in debt and would like to thank the people who helped me in my results presented in thesis.

First, I thank Herman Verrelst for inviting me to the Department of Electrical Engineering at the Katholieke Universiteit Leuven, helping my first steps in Leuven and sharing his ideas about the spin-off activity he followed.

I would like to express my gratitude to Prof. Bart De Moor for his support of my research (planned for a half year, extended to four years), for the possibility of participating in the stimulating environment of the emerging bioinformatics group, and for his trust that the page count of my Ph.D. manuscript will ever increase, then that it can be cut to a manageable level. I thank Prof. Yves Moreau for his patient, parsimonious, and accurate advices on the content and the style of our papers and the Ph.D. manuscript. I greatly appreciate the professional and personal support of Prof. Dirk Timmermann in the IOTA project, and his belief in Bayesian Bayesian networks. I am also in debt to Prof. Sabine Van Huffel and Prof. Joos Vandewalle for their advices on ROC methodology and Bayesian neural networks.

I would like to thank Stein Aerts, Janick Mathys, Gert Thijs, Frank De Smet and Kathleen Marchal for their biomedical crash-courses. I thank Patrick Glenisson for his professionalism to nurture our ideas on integrated analysis of genomical text and data (from ATAGC to TextGate). I am very grateful to Geert Fannes for his trust and work, because many of these concepts would never have been finished without his propensity, fluency and perseverance w.r.t. probability theory, Bayesianism, C-MATLAB coding and debugging (the trouble is that we cannot grasp multitemporal causality. . . ;-) Dank u voor uw hulp!

At the Budapest University of Technology and Economics, I am in debt to Prof. László Györfi for firmly securing a probabilistic approach to machine learning in his courses. I would like to express my thanks to Prof. Tadeusz Dobrowiecki at the Department of Measurement and Information systems for sharing his broad vision on artificial intelligence and painting red (in each June ;-) my Ph.D. manuscript with his comments. I thank András Millinghoffer and Gábor Hullám for their work and diligent "reports" about bugs in the software.

Finally, I thank with love to my family, particularly to my wife (sometimes an epidemiologist colleague ;-), for ensuring conservative steps in our "random walk" on the ever changing landscape of Hungary and Europe.

Abstract

We developed methods to incorporate expert knowledge and electronic literature into Bayesian inference over domain models and conditional models. Particularly, we investigated the relations between and the joint usage of three types of probabilistic models: the "literature" model corresponding to free-text electronic literature, the "causal" domain model and a particular conditional model. These models were applied to the preoperative classification of ovarian masses.

First, we collected and elicited textual, qualitative and quantitative information about ovarian cancer, such as electronic resources, the qualitative and quantitative characterization of the associative pairwise relations between variables, the causal and multivariate aspects of the relations, and complete probabilistic, causal domain models as Bayesian networks annotated with free-text and links to the electronic literature. This "annotated" Bayesian Network was the precursor of our proposal for probabilistic logical knowledge bases incorporating complex distributions and free-text information.

Second, we characterized and investigated a model-based method for statistical text analysis that uses Bayesian networks to support knowledge extraction and discovery from biomedical publications.

Third, we performed a cross-comparison and evaluation of the elicited expert priors and the posteriors for the models based on literature and clinical data. We devised methods to perform Bayesian inference about classification oriented, complex structural features of a causal model, such as sets of relevant features or classification subgraphs, incorporating heterogeneous information sources.

Finally, we evaluated the classification performance of Bayesian classifiers including logistic regression, multilayer perceptrons and various Bayesian networks. For Bayesian network classifiers we analyzed the induced joint posterior over various structural features and performance measures. For logistic regression and multilayer perceptrons we proposed and investigated methods to derive structural and parametric priors from priors over Bayesian networks.

The system, which we implemented performs personalized, domain-specific Bayesian inferences over the optionally linked "literature" model, causal domain model and conditional model by fusing expertise, electronic literature and observational data. Specifically, it performs a Bayesian, four-level, sequential analysis of relevance — at the levels of pairs of variables, sets of variables, submodels, and models — incorporating diverse priors; thus facilitating knowledge-rich statistical data analysis.

Notation*

List of symbols

x,$\underline{x}$,$\underline{\underline{x}}$	scalar, (column)vector or set, matrix
$X, x, p(X)$	random variable X, value x, probability mass function/density of X
$\mathrm{E}_{X,p(X)}[f(X)]$	expectation of $f(X)$ w.r.t. $p(X)$
$\mathrm{var}_{p(X)}[f(X)]$	variance of X w.r.t. $p(X)$
$I_p(\underline{X}\|\underline{Z}\|\underline{Y})$	observational conditional independence of $\underline{X}$ and $\underline{Y}$ given $\underline{Z}$ w.r.t. p
$(X \perp\!\!\!\perp Y\|Z)_p$	$I_p(\underline{X}\|\underline{Z}\|\underline{Y})$
$(X \not\perp\!\!\!\perp Y\|Z)_p)$	$\neg I_p(\underline{X}\|\underline{Z}\|\underline{Y})$
$CI_p(\underline{X};\underline{Y}\|\underline{Z})$	interventional conditional independence of $\underline{X}$ and $\underline{Y}$ given $\underline{Z}$ w.r.t. p
$\prec$	(partial) ordering
$\prec^c$	a complete reference ordering of the domain variables
$G, \underline{\theta}$	Bayesian network (BN) structure, BN parameters
$G^{\sim}$	essential graph of Directed Acyclic Graph (DAG) G
$\hat{G}_C^{\prec}(D)$	an optimal graph w.r.t. ordering $\prec$, data set D, and score/method C
$\mathcal{G}(n)/\mathcal{G}^k(n)$	set of DAGs over n nodes/with maximum k parents
$\mathcal{G}^{\prec}$	set of DAGs compatible with ordering $\prec$
$\sim$	compatibility relation
$\mathrm{pa}(X_i, G) \sim \prec$	$\mathrm{pa}(X_i, G)$ parental set is compatible with ordering $\prec$
$F, \mathcal{F}, f$	feature function, its range, a feature value
$\mathcal{F}^{\prec}$	set of values f compatible with $\prec$
$S_i(f, \prec)$	the set of valid parental sets of X_i in feature f given ordering $\prec$
$C_i(f, \prec, \mathrm{pa})$	a clause expressing $\mathrm{pa} \in S_i(f, \prec)$
$\mathrm{MB}_p(X_i)$	a Markov Blanket of X_i in p
S^{MLP}/S, $\underline{\underline{\omega}}$	Multilayer perceptron (MLP) structure, MLP parameters
$\mathrm{pa}, \mathrm{pa}(X_i, G)$	set of parental variables, set of parents of X_i in G
pa_{ij}	the jth configuration of the values of the parents of X_i in an ordering
$\mathrm{bd}(X_i, G)$	set of parents, children and the children's other parents of X_i in G
$\mathrm{MBG}(X_i, G)$	the Markov Blanket/Mechanism Boundary Graph of X_i in G
$\mathrm{MB}(X_i, G)$	Markov Blanket of X_i defined by $\mathrm{bd}(X_i, G)$ in p compatible with G
$\mathrm{MBM}(X_i, X_j, G)$	the binary Markov Blanket membership
n	number of random variables
k	maximum number of parents in DAGs

*See also the remarks about style and notation in Section 2.2

N	number of observed samples
$N_+/N_{\ldots,+,\ldots}$	the appropriate sum of $N_i/N_{\ldots,i,\ldots}$
D_N/D_N^L	real/literature data set with N complete observations
$D\|X$	data set D restricted to the set of variables X
$D^{\mathrm{IO}_1/\mathrm{IO}_2}$	clinical data sets
$D^{\mathrm{ME}^{HMR}_{\mathrm{O/R}}}, D^{\mathrm{PM}^{HMR}_{\mathrm{O/R}}}$	literature data sets based on a Medline (ME) and Pubmed (PM) corpus with $H/M/R$ filters binarized with Occurence/Relevance
D^*/D'	artifical data set by bootstrap or Monte Carlo methods
$\|\|$	cardinality
$1()$	indicator function
$S_i^{h/m/r/n}$	set of undirected edges with node i with high, medium, reasonable, and negligible pairwise relevance
$G^{H/M/R}$	the three prior DAG structures with high, medium, and reasonable relevance
$S_i^{H/M/R}$	the set of parents of node i in DAGs $G^{H/M/R}$
f', f''	first and second derivatives of function f
A^T	transpose of the matrix A
$\mathcal{A}()$	free-text annotation for an object
ξ^+/ξ^-	informative/noninformative background knowledge
KB	knowledge base (axioms)
$KB \models \alpha$	the entailment of sentence α w.r.t. knowledge base KB
$\mathcal{M}(KB)$	the set of models of a knowledge base KB
$\neg, \wedge, \vee, \not\equiv, \rightarrow$	operations of negation, and, or, exclusive or, implication
$\cap, \cup, \setminus, \Delta$	the intersection, union, difference, and symmetric difference
$KB \vdash_i \alpha$	the provability of sentence α by method $\vdash_i$ from axioms KB
Γ	the Gamma function
$\mathrm{Beta}(x\|\alpha,\beta)$	the probability density function (pdf) of the Beta distribution
$\mathrm{Dir}(x\|\underline{\alpha})$	the pdf of the Dirichlet distribution
$\mathrm{N}(x\|\mu,\sigma)$, $\mathrm{N}(x\|\underline{\mu},\underline{\underline{\Sigma}})$	the pdf of the normal distribution
BD,BD_e	Bayesian Dirichlet prior, observationally equivalent BD prior
BD_{CH}	a Bayesian Dirichlet (BD) prior with hyperparameters 1
$\mathrm{BD}_{\mathrm{eu}}$	a BD prior, where the hyperparameters are the converse of the number of parameters in the local dependency model
$L(\underline{\theta}; D_N)$	the likelihood function $p(D_N\|\underline{\theta})$
$H(X,Y)$, $I(X;Y)$	the entropy and the mutual information of X and Y
$\mathrm{KL}(X\|Y)$	the Kullback-Leibler divergence of X and Y
$H(X\|Y)$	the cross-entropy of X and Y
$L_1(,)$, $L_2(,)$	the Manhattan and the Euclidean distances the absolute and the quadratic losses
$L_0(,)$	the 0-1 1oss
$\mathcal{O}()/\Theta()$	asymptotic, proportional upper/upper and lower bound
$\mathrm{maxK}^{\mathrm{th}}(s)$	the Kth value in decreasing ordering in the set of scalars s

Acronyms

ABN	Annotated Bayesian Network
AUC	Area Under the ROC curve
BAN-BN/BAN	Bayesian Network Augmented Naive Bayesian Network
BMA	Bayesian Model Averaging
BN	Bayesian Network
BNC	Bayesian Network Classifier
DAG	Directed Acyclic Graph
FSS	Feature Subset Selection (problem)
FGS	Feature (sub)Graph Selection (problem)
HPD	High Probability Density (region)
IDO	IDO/99/03 project (K.U.Leuven) entitled "Predictive computer models for medical classification problems using patient data and expert knowledge"
IOTA	a study by the "International Ovarian Tumor Analysis" consortium
IR	Information Retrieval
LR	Logistic Regression
KE	Knowledge Engineering
KB	Knowledge Base
MAP	Maximum A Posteriori
MD	MEDLINE
MI	mutual information
ML	Maximum Likelihood
MLP	Multilayer perceptron
MBG	Markov Blanket/Mechanism Boundary Graph (a.k.a. classification or feature subgraph)
MB	Markov Blanket/Boundary set
MBM	Markov Blanket/Boundary Membership
(MC)MC	(Markov Chain) Monte Carlo
MPFs	Most Probable Features (problem)
Naive-BN/N-BN	Naive Bayesian network
OC	Ovarian Cancer
pABN-KB	Probabilistic Annotated Bayesian Network Knowledge Base
PM	PUBMED
ROC	Receiver Operating Characteristic (ROC) Curve
TAN-BN/TAN	Tree Augmented Naive Bayesian Network

Publication list

[10] P. Antal. Applicability of prior domain knowledge formalised as Bayesian network in the process of construction of a classifier. In *Proc. of the 24th Annual Conf. of the IEEE Industrial Electronic Society (IECON '98)*, pages 2527–2531, 1998.

[27] P. Antal, H. Verrelst, D. Timmerman, S. Van Huffel, B. De Moor, and I. Vergote. How might we combine the information we know about a mass better? The use of mathematical models to handle medical data. 1st Monte Carlo Conf. on Updates in Gynaecology, 2000, Internal Report 00-145, ESAT-SISTA, K.U.Leuven (Leuven, Belgium), 2001.

[28] P. Antal, H. Verrelst, D. Timmerman, Y. Moreau, S. Van Huffel, B. De Moor, and I. Vergote. Bayesian networks in ovarian cancer diagnosis: Potential and limitations. In *Proc. of the 13th IEEE Symp. on Comp.-Based Med. Sys. (CBMS-2000)*, pages 103–109, 2000.

[18] P. Antal, G. Fannes, H. Verrelst, B. De Moor, and J. Vandewalle. Incorporation of prior knowledge in black-box models: Comparison of transformation methods from Bayesian network to multilayer perceptrons. In *Workshop on Fusion of Domain Knowledge with Data for Decision Support, 16th Uncertainty in Artificial Intelligence Conference*, pages 42–48, 2000.

[11] P. Antal, G. Fannes, S. Van Huffel, B. De Moor, J. Vandewalle, and Dirk Timmerman. Bayesian predictive models for ovarian cancer classification: evaluation of logistic regression, multi-layer preceptron and belief network models in the Bayesian context. In *Proc. of the 10th Belgian-Dutch Conference on Machine Learning, BENELEARN 2000*, pages 125–132, 2000.

[22] P. Antal, T. Meszaros, B. De Moor, and T. Dobrowiecki. Annotated Bayesian networks: a tool to integrate textual and probabilistic medical knowledge. In *Proc. of the 13th IEEE Symp. on Comp.-Based Med. Sys. (CBMS-2001)*, pages 177–182, 2001.

[15] P. Antal, G. Fannes, Y. Moreau, B. De Moor, J. Vandewalle, and D. Timmerman. Extended Bayesian regression models: a symbiotic application

of belief networks and multilayer perceptrons for the classification of ovarian tumors. In *Lecture Notes in Artificial Intelligence (AIME 2001)*, pages 177–187, 2001.

[17] P. Antal, G. Fannes, F. De Smet, and B. De Moor. Ovarian cancer classification with rejection by Bayesian belief networks. In *Workshop notes on Bayesian Models in Medicine, European Conference on Artificial Intelligence in Medicine (AIME'01)*, pages 23–27, 2001.

[19] P. Antal, P. Glenisson, T. Boonefaes, P. Rottiers, and Y. Moreau. Towards an integrated usage of expression data and domain literature in gene clustering: representations and methods. Internal Report 01-69, ESAT-SISTA, K.U.Leuven (Leuven, Belgium), 2001.

[14] P. Antal, G. Fannes, Y. Moreau, and B. De Moor. Bayesian applications of belief networks and multilayer perceptrons for ovarian tumor classification with rejection. *Artificial Intelligence in Medicine*, vol. 29, pages 39–60, 2003.

[20] P. Antal, P. Glenisson, G. Fannes, J. Mathijs, Y. Moreau, and B. De Moor. On the potential of domain literature for clustering and Bayesian network learning. In *Proc. of the 8th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (ACM-KDD-2002)*, pages 405–414, 2002.

[23] P. Antal, T. Meszaros, B. De Moor, and T. Dobrowiecki. Domain knowledge based information retrieval language: an application of annotated Bayesian networks in ovarian cancer domain. In *Proc. of the 15th IEEE Symp. on Comp.-Based Med. Sys. (CBMS-2002)*, pages 213–218, 2002.

[5] S. Aerts, P. Antal, B. De Moor, and Y. Moreau. Web-based data collection for ovarian cancer: a case study. In *Proc. of the 15th IEEE Symp. on Computer-Based Medical Sys. (CBMS-2002)*, pages 282–287, 2002.

[16] P. Antal, G. Fannes, Y. Moreau, D. Timmerman, and B. De Moor. Using literature and data to learn Bayesian networks as clinical models of ovarian tumors. *Artificial Intelligence in Medicine*, vol. 30, pages 257–281, 2004.

[13] P. Antal, G. Fannes, Y. Moreau, and B. De Moor. Using domain literature and data to annotate and learn Bayesian networks. In H. Blockeel and M. Denecker, editors, *Proc. of 14th Belgian-Dutch Conference on Artificial Intelligence (BNAIC'02)*, pages 3–10, 2002.

[191] Y. Moreau, P. Antal, G. Fannes, and B. De Moor. Probabilistic graphical models for computational biomedicine. *Methods of Information in Medicine*, vol. 42(4), pages 161–168, 2002.

[114] P. Glenisson, P. Antal, J. Mathys, Y. Moreau, and B. De Moor. Evaluation of the vector space representation in text-based gene clustering. In *Proc. of the Pacific Symposium on Biocomputing (PSB03)*, pages 391–402, 2003.

[24] P. Antal and A. Millinghoffer. Learning causal bayesian networks from literature data. *Proceedings of the 3rd International Conference on Global Research and Education, Inter-Academia'04*, pages 149–160, 2004.

[188] P. Antal and A. Millinghoffer. Statisztikai adat- és szövegelemzés Bayeshálókkal: a valószínűségektől a függetlenségi és oksági viszonyokig. *Híradástechnika*, vol. 60, pages 40–49, 2005 (in Hungarian).

[25] P. Antal and A. Millinghoffer. A probabilistic knowledge base using annotated bayesian network features. In *Proceedings of the 6th International Symposium of Hungarian Researchers on Computational Intelligence*, pages 1–12, 2005.

[26] P. Antal and A. Millinghoffer. Literature mining using bayesian networks. In *Proc. of third European Workshop on Probabilistic Graphical Models*, pages 17–24, 2006.

[21] P. Antal, G. Hullám, A. Gézsi, and A. Millinghoffer. Learning complex bayesian network features for classification. In *Proc. of third European Workshop on Probabilistic Graphical Models*, pages 9–16, 2006.

[189] A. Millinghoffer, G. Hullám, and P. Antal. On inferring the most probable sentences in bayesian logic. In *Workshop notes on Intelligent Data Analysis in bioMedicine And Pharmacology (IDAMAP-2007), 11th Conference on Artificial Intelligence in Medicine (AIME 07)*, pages 13–18, 2007.

List of Figures

1.1 An artificial Bayesian network structure showing also the Markov Blanket and the Markov Blanket Graph of a target variable. . . . 5
1.2 The temporal evolution of the collective belief — inferred from the literature — that a given variable is relevant for the preoperative diagnostics of ovarian cancer. 6
1.3 The reconstruction of prior knowledge in a biomedical domain from literature data and its incorporation in learning causal domain models. 7
1.4 The temporal evolution of the belief — inferred from growing amount of clinical data — that a given set of variables is relevant for the preoperative diagnostics of ovarian cancer. 9
1.5 The temporal evolution of the belief — inferred from growing amount of clinical data — that a given subgraph over the subset of the variables is (exactly) relevant for the preoperative diagnostics of ovarian cancer. 10
1.6 The two-step methodology for the fusion of knowledge and data for classification. 11
1.7 The learning curves for the multilayer perceptron and various Bayesian network models (Naive, TAN, GTAN, BN) using an informative and noninformative parameter priors. 12

3.1 The Markov Blanket and the Markov Blanket Graph of a target variable in a Markov chain. 38
3.2 The sets of observationally equivalent Bayesian network structures. 40

4.1 An early BN for the ovarian cancer problem. 63
4.2 Three prior BN structures for the thirty-five IOTA variables. . . 64
4.3 The annual number of papers in ovarian cancer. 66

5.1 The text-based hierarchical cluster tree and similarity network of the domain variables. 74
5.2 The separated and integrated IR in knowledge engineering. . . . 75

6.1 The derivation of the transitive publication model. 82

6.2 The maximum a posteriori Bayesian network given the $D^{PM_R^R}$ literature data set of the thirty-five variables. 84

8.1 The histogram of the sample sizes of "strong divergence" for the expert's probability estimates and the percentages of the refuted estimates with a given credibility. 121
8.2 The hyperposterior of the virtual sample size for the naive, best inductive and elicited structures. 122
8.3 The advantage of the expert's estimates per variable in a prequential evaluation. 123
8.4 The advantage of the expert's estimates per variable with virtual sample size 150 in a prequential evaluation. 124
8.5 The advantage of the transformed prior in the naive model and the parental sets in the naive model. 125
8.6 The evaluation of the combinations of (1) the naive, best-inductive and elicited structures and (2) noninformative and informative parameter priors. 126
8.7 The sequential evaluation of the parental sets in the expert's G^H Bayesian network given the expert's total causal ordering. 130
8.8 The temporal evolution of the posteriors of more than one variable difference in the parental sets of the variables in the expert's G^M model given the expert's total causal ordering. 131
8.9 The temporal evolution of the edge-differences between clinical data-based maximum a posteriori Bayesian networks and the expert's G^M Bayesian network. 132
8.10 The rate of decrease of the posteriors of the most probable MB sets, MBGs, and BN structures. 140
8.11 The temporal evolution of the belief — inferred from growing amount of clinical data — that a given variable is relevant for the preoperative diagnostics of ovarian cancer. 142
8.12 The temporal evolution of the belief — inferred from growing amount of clinical data — that a given variable is not relevant for the preoperative diagnostics of ovarian cancer. 143
8.13 The temporal evolution of the collective belief — inferred from the literature — that a given variable is not relevant for the preoperative diagnostics of ovarian cancer. 144
8.14 The MBM-based approximations of the posteriors and ranks of the 20 most probable MB($Pathology$) sets. 145
8.15 The effect of various structure priors from expert and literature on learning MAP Bayesian network for varying sample sizes. . . 148

9.1 The Bayesian network representation of the independence assumptions of the Bayesian conditional modeling. 151

10.1 The estimated posterior distribution of the performance measure and the model complexity w.r.t. classification. 177

10.2 The estimated posterior of the number of parameters and inputs for the MAP MBGs. 178
10.3 The estimated mean and conditional distribution of the AUC variable conditioned on the ratio of the number of parameters and the number of inputs. 179
10.4 The effect of informative parameter prior on classification performance in case of small sample size. 181
10.5 The effect of informative parameter prior on classification performance in case of large sample size. 182
10.6 The AUC performance for BNs using different text-based prior and expert priors. 183
10.7 The effect of BMA using the small set of variables. 184
10.8 The effect of BMA using the medium set of variables. 185
10.9 The effect of BMA using the large set of variables. 186

A.1 The biplot of the domain variables and 604 cases of the IOTA-1.1 data set. 200
A.2 The biplot of the domain variables and 782 cases of the IOTA-1.2 data set. 201
A.3 The sorted eigenvalues of the covariance matrix of the IOTA-1.2 data set. 202
A.4 The BN structure used in the parameter elicitation and the maximum a posteriori PDAG over the same set of variables. 202
A.5 The maximum a posteriori Bayesian network compatible with the expert's total ordering of the thirty-five variables. 203
A.6 The maximum a posteriori essential graph over the thirty-five variables. 203
A.7 The maximum a posteriori Bayesian network. 204
A.8 The maximum a posteriori Bayesian network given the $D^{PM_R^R}$ literature data set (compatible with the expert's total ordering of the thirty-five variables). 204
A.9 The maximum a posteriori Bayesian network given the $D^{PM_R^H}$ literature data set, ($\mathrm{BD_{eu}}$ parameter priors, compatible with the expert's total ordering of the thirty-five variables). 205
A.10 The maximum a posteriori Bayesian network given the $D^{PM_R^H}$ literature data set (CH parameter priors, compatible with the expert's total ordering of the thirty-five variables). 205

List of Tables

8.1 The comparison of the expert's relevance ranks for pairwise relations and various pairwise text- and data-scores (the AUC values of univariate discriminators). 127
8.2 The Spearman rank correlation coefficients for the cross-comparison of the expert score, the text scores, and the data scores. 128
8.3 Detailed causal comparison of prior and data based BNs. 132
8.4 Typed and causal differences between the prior structures and an ordering-specific clinical data based MAP BN. 133
8.5 Typed and causal differences between the prior structures and a clinical data based MAP BN. 133
8.6 Typed and causal differences between the prior structures and the clinical data based MAP BN. 134
8.7 Typed and causal comparison of literature based BNs against a prior and a clinical data based structure. 135
8.8 The learnability of the expert's opinion that a given variable is relevant for the preoperative diagnostics of ovarian cancer. 141
8.9 The sensitivity, specificity and misclassification rate of the most probable MB sets of the Pathology variable. 146

10.1 Expert agreement with the prior domain model in discriminating benign and malignant adnexal masses. 179

11.1 Main types of the elicited prior knowledge and their relation to constructs and methods. 192

A.1 The abbreviations and the short description of the domain variables. 198
A.2 Univariate statistics based on the IOTA-1.1 data set for the thirty-one variables containing 604 cases. 199
A.3 The properties of the forward selected LR models over the elicited, medium and complete variable sets. 206
A.4 The ordering conditional posteriors of the sets of parental sets in the expert's total ordering 207
A.5 The posteriors of the MBM(Pathology, X_i) features for single and unconstrained orderings. 208

A.6 Convergence score values and the standard error of the MCMC estimates of the posterior of the MBM(Pathology,.) features. . . 209
A.7 The most probable MB sets of the Pathology variable. 210
A.8 The estimated posteriors with convergence and confidence values of the most probable MB sets of the Pathology variable. 211
A.9 The most probable MBGs given the reference ordering. 211
A.10 The most probable MBGs of the Pathology variable (unconstrained case). 212
A.11 The estimated posteriors with convergence and confidence values of the most probable MBGs of the Pathology variable. 212

Contents

Foreword **i**

Abstract **iii**

Notation **v**

Publication list **ix**

1 Introduction **1**
1.1 A tour of the thesis 3
1.2 Chronology of doctoral activities 9
1.3 Chapter-by-chapter overview 13

2 A Bayesian primer **17**
2.1 The subjective interpretation of probability 18
2.2 The general scheme of Bayesian inference 18
2.2.1 Setting up the model 19
2.2.2 Predictive inference 20
2.2.3 Parametric inference with Bayes' rule 21
2.2.4 Reporting the posterior 21
2.2.4.1 Reporting the posterior distribution 22
2.2.4.2 Reporting posterior quantities 22
2.2.5 Model transformation and reparameterization 23
2.3 Inference with Monte Carlo methods 24
2.3.1 Markov Chain Monte Carlo methods 24
2.3.1.1 Markov chains 25
2.3.1.2 MCMC with the Metropolis-Hastings scheme . . 27
2.3.1.3 Convergence and confidence issues 28
2.3.2 The hybrid Markov Chain Monte Carlo method 29
2.4 Model evaluation and selection 29
2.4.1 The prequential framework 29
2.4.2 Maximum a posteriori analysis 31

3 Bayesian networks primer **33**
3.1 Representational issues . . . 34
3.1.1 Three aspects: belief, relevance and causation . . . 34
3.1.1.1 The model of observational independencies . . . 34
3.1.1.2 The model of causal (in)dependencies . . . 35
3.1.2 Probabilistic Bayesian networks . . . 35
3.1.2.1 Markov conditions . . . 35
3.1.2.2 Definitions of Bayesian networks . . . 37
3.1.2.3 Stability . . . 38
3.1.2.4 Equivalence classes of Bayesian networks . . . 39
3.1.3 Causal Bayesian networks . . . 41
3.1.3.1 On the possibility of causal interpretation . . . 41
3.1.3.2 The Causal Markov Condition . . . 42
3.1.3.3 The interventionist and mechanistic views . . . 42
3.1.3.4 Pairwise causal relations . . . 43
3.1.4 On the relativity of the interpretations . . . 43
3.1.5 Bayesian networks in the Bayesian framework . . . 44
3.1.5.1 Parameter priors for Bayesian network models . 44
3.1.5.2 Structure priors for Bayesian network models . . 46
3.2 Inference methods . . . 49
3.2.1 Inference over values with observations . . . 49
3.2.1.1 Fixed parameter and fixed structure . . . 50
3.2.1.2 Bayesian parameter and fixed structure . . . 50
3.2.1.3 Bayesian parameter and structure . . . 51
3.2.2 Inference over domain values with interventions . . . 51
3.2.3 Inference over model parameters . . . 51
3.2.4 Inference over model structures . . . 52
3.3 Knowledge engineering . . . 53
3.4 Prequential analysis by Bayesian networks . . . 54
3.5 Learning Bayesian networks . . . 55
3.5.1 Score functions and their properties . . . 56
3.5.2 Search algorithms for finding high-scoring BNs . . . 57

4 Prior knowledge and data about ovarian cancer **59**
4.1 The biomedical background, the IDO, and the IOTA projects . . 59
4.1.1 The domain and domain concepts . . . 60
4.1.2 Previous predictive models . . . 60
4.2 The data sets . . . 61
4.2.1 The IDO data set . . . 61
4.2.2 The IOTA data sets . . . 61
4.3 Knowledge engineering BNs . . . 62
4.3.1 An early Bayesian network for ovarian cancer . . . 62
4.3.2 Parameter priors for a small-scale model . . . 63
4.3.3 Elicitation of structural priors . . . 63
4.3.3.1 Prior structures from a model-based approach . 64
4.3.3.2 Priors from a pairwise relevance approach . . . 65

4.3.3.3 The causal ordering of variables 65
4.3.4 Electronic resources for knowledge engineering 65
4.3.4.1 Text kernels 65
4.3.4.2 Document collections 66
4.3.4.3 Domain vocabularies 66

5 Fusing BNs and logical knowledge bases 67
5.1 Bayesian knowledge engineering 68
5.2 Probabilistic knowledge bases by embedded Bayesian networks . 69
5.3 Keyword profiles of ABN-KB objects 72
5.4 Explorations by keyword-based profiles 74
5.5 An ABN-based information retrieval language 75
5.5.1 Informational relevance expressed by ABN sentences . . . 75
5.5.2 An IR language for contextual relevance 76

6 Text mining with BNs 77
6.1 The literature data 78
6.2 Concepts, associations, and causation 79
6.3 Literature mining 79
6.4 BN models of publications 80
6.5 Local scores for pairwise relationships 83
6.6 Results 84

7 Inference over BN features 85
7.1 Bayesian network features 87
7.1.1 Edges: direct pairwise dependencies 88
7.1.2 Ordering of the variables 88
7.1.3 Relevant variables 89
7.1.4 MBG subnetworks 93
7.1.5 Learning of subnetworks 93
7.1.6 The properties and taxonomy of features 94
7.2 The Markov Blanket (sub)Graph feature 95
7.3 The bootstrap confidence measure 99
7.4 On the advantage of feature posteriors 103
7.5 MC methods for a feature posterior 105
7.5.1 The DAG-based MCMC methods 106
7.5.2 The ordering-based MCMC methods 106
7.5.2.1 The ordering-conditional feature posteriors . . . 106
7.5.2.2 Advantages of ordering-based MCMC 108
7.5.2.3 Estimating edge and pairwise relevance 109
7.6 Decision over features using MC estimates 110
7.6.1 The Most Probable Features problem 111
7.6.2 Effect of feature cardinality in MPFs 111
7.7 Integrating estimation and search of MBGs 113

8 Analysis and fusion **117**
8.1 Fusion of expertise, literature, and data . . . 118
8.1.1 Fusion through linked models . . . 118
8.1.2 Fusion through linked features . . . 119
8.1.3 Fusion of pairwise text-based scores and models . . . 120
8.2 Data-based evaluation of the small BN . . . 120
8.2.1 From prior parameters to hyperposteriors . . . 120
8.2.2 Evaluation of parental sets and configurations . . . 123
8.2.3 Evaluation of models and transformed priors . . . 123
8.3 Analysis of local scores . . . 124
8.4 Analysis at the model level . . . 129
8.4.1 Structure priors vs. clinical data . . . 130
8.4.2 Evaluating literature models . . . 134
8.5 Feature learning . . . 135
8.5.1 An estimation and search method for MBGs . . . 136
8.5.2 The exact treatment of the orderings . . . 139
8.5.2.1 Posteriors of Markov blanket memberships . . . 140
8.5.2.2 Posteriors of MB sets and MB graphs . . . 143
8.5.3 Applying MCMC methods over the orderings . . . 147
8.6 Effect of fusion . . . 147

9 Bayesian classification **149**
9.1 On the validity of the conditional approach . . . 150
9.2 The Bayesian modeling of class probabilities . . . 151
9.3 Reporting as decision in Bayesian classification . . . 152
9.3.1 Reporting the class label . . . 152
9.3.2 Reporting the class probability . . . 153
9.4 Bayesian network classifiers . . . 153
9.4.1 Domain models as classifiers . . . 153
9.4.2 The naive Bayesian network and its extensions . . . 155
9.5 Logistic regression and its extensions . . . 156
9.5.1 Logistic regression . . . 156
9.5.2 The relation between MBG and LR models . . . 158
9.5.3 The multilayer perceptron extension . . . 160

10 Bayesian classifiers with a prior domain model **161**
10.1 Reasons for the dual representation . . . 162
10.2 Parameter priors for Bayesian classifiers . . . 164
10.2.1 Prior transformation between BNs . . . 165
10.2.2 Noninformative priors for LRs and MLPs . . . 166
10.2.3 Informative MLP prior from a Bayesian network . . . 166
10.2.3.1 Using a prior data set . . . 167
10.2.3.2 Using a prior over data sets . . . 168
10.2.3.3 Using conditional distance minimization transformation . . . 169
10.2.3.4 Discrete-continuous transformations . . . 172

10.2.3.5 Analytic approximation of the transformed informative prior . . . 173
10.3 Structure priors for Bayesian classifiers . . . 173
10.4 Joint probabilities of conditional features . . . 176
10.5 The frequentist LR modeling . . . 177
10.6 Effect of parameter priors on classification . . . 178
10.7 Effect of structure priors on classification . . . 180
10.8 Effect of model averaging on classification . . . 182
10.9 Discussion . . . 184

11 Conclusion **187**
11.1 Contributions of this dissertation . . . 187
11.2 The developed software platform . . . 190
11.3 Applicability in the postgenomic era . . . 191
11.3.1 Main constructs and methods . . . 191
11.3.2 Main types of the prior knowledge . . . 192
11.3.3 From current results to proposed uses . . . 192
11.4 Challenges . . . 194

A **197**

Appendix **196**

Bibliography **213**

Chapter 1

Introduction

The recent technological developments in life sciences enabling the sequencing of genomes and high-throughout genomic, proteomic, metabolic techniques have redefined biology and medicine and opened the genomic and post-genomic era. The rapidly accumulating scientific knowledge and data, combined with the effect of the developing semantic web have expanded and redefined human cognition by creating the long sought "world brain" in the "e-science" context [103]. An important factor behind this development has been the sheer volume of knowledge as even the narrowly interpreted "domain knowledge" increasingly exceeds the limits of individual cognition. The semantic web offers a potential solution for this new growth of human knowledge, consequently biomedical knowledge is becoming more and more "external" (i.e., distributed, collectively shared and maintained in knowledge bases, databases and electronically accessible repositories of natural language publications). These trends suggest that further development of life sciences depends equally on efficient externalization and fusion of knowledge as on further technological breakthroughs.

An important and inherent feature of this new voluminous knowledge is uncertainty. Various forms of uncertainty may arise because of the multilevel and multiple approaches in biomedicine, beside incompleteness and inherent uncertainty, but many of these can be managed within the single framework of probability theory using a subjectivist interpretation. The corresponding Bayesian framework offers a normative method for representing knowledge, learning from observations and, with utility theory, reaching optimal decisions. In short, the Bayesian approach provides a normative and unified framework for knowledge engineering, statistical machine learning and decision support. Its ability to incorporate consistently the voluminous and heterogeneous prior knowledge in statistical learning connects statistics and knowledge engineering, leading to the concept of adaptive knowledge bases or "knowledge intensive" statistics. The Bayesian framework also offers a computational framework for learning and using complex probabilistic models, mainly by various stochastic simulations to perform Bayesian inference, leading to computationally-intensive statistics. Actually, the exponential increase in computational power in the last fifty years

was the main condition for the sudden widespread of Bayesian techniques in the nineties. As the complexity of the priors, the models and the queries can be expected to grow further, new advances supporting the use of background domain knowledge in prior incorporation and in posterior analysis are essential in applied Bayesian statistics.

The vast biomedical domain knowledge, which is a mixture of human expertise, knowledge bases, databases and literature repositories has posed a new, practical challenge for applied Bayesian data analysis: how to use heterogeneous domain knowledge and data efficiently in knowledge engineering, machine learning and decision support. This challenge is particularly acute in the complex and rapidly changing fields of medicine and genomics where much of the voluminous knowledge is only available as free-text scattered throughout the literature. Here the proper interpretation of the results of data analysis became an important bottleneck. That is beside the technology of measurements and the statistical aspects of data analysis, the support for understanding and revealing the biomedical relevance of the results became essential.

This thesis investigates the integrative* analysis and fusion of heterogeneous sources, such as expert knowledge, literature and statistical data with special emphasis on classification, on the usage of domain literature and on multiple models. Roughly speaking, our goal was to work out a theoretical framework and implement a system for the formulation and inference of probabilistic queries in a special domain as a prototype for a general view of the semantic web as a probabilistic knowledge base. The topic of the thesis also contributes to knowledge intensive and computation intensive Bayesian statistics by (1) investigating the role of voluminous, heterogeneous, partly electronic a priori knowledge, involving also beliefs arising from domain literature and knowledge bases and (2) performing Bayesian statistical inferences over knowledge-based, multivariate properties of complex models.

In our investigation of incorporating complex, heterogeneous priors in Bayesian data analysis, the Bayesian network was the main model class. The Bayesian network representation became an important tool in many disciplines related to the engineering and induction of knowledge, such as in the overlapping fields of decision theory, statistics, artificial intelligence, causality research, machine learning and data mining. In the thesis we used Bayesian networks for knowledge acquisition and representation, for statistical text mining, for inferring complex, multivariate properties of the domain, and for performing prediction.

Whereas a pure prediction and classification task permits more specialized solutions (such as various kernel methods for classification), frequently it is equally important to understand the effects and interrelations of the domain variables. We therefore investigated the applicability of Bayesian networks in statistical text mining and in the integrative analysis. From the point of view of conditional modeling this work supports the process of construction of a classifier providing a methodology and a probabilistic framework to (1) collect

*We use the term "integrated" to indicate joint usage of multiple sources, "integral" to indicate the complete treatment of a domain and "integrative" to indicate the existence of underlying overall models.

domain knowledge manually, semi-automatically and automatically, (2) formalize various priors for black-box classifiers or for hybrid systems and (3) support the interpretation and understanding of the classifier and its predictions.

In the thesis the modeling and classification of ovarian tumors served as a real world application domain. In the first part of the thesis, the derivation of various priors related to clinical and partly biological models of ovarian cancer (OC) are presented both with manual knowledge engineering and with automated knowledge discovery and information extraction methods. The next topic of the thesis is the fusion of the sources to perform inferences on model properties, particularly related to classification such as the set of relevant variables and the structure of their effect on a target variable. Finally, we present a method that derives an informative distribution for black-box parametric classifiers from the formalized priors for Bayesian networks and we investigate the role of such priors in a classification problem.

1.1 A tour of the thesis

The general goal of the thesis was to develop an overall probabilistic framework that incorporates the textual prior knowledge such as publications, various forms of expert knowledge ranging from free-text comments to quantitative estimates, domain models such as Bayesian networks and conditional models such as logistic regression, because such an overall probabilistic framework allows the formulation and inference of complex, integrative queries. From an engineering point of view this goal corresponds to the integrated treatment of the phases of data analysis, such as preprocessing or interpretation. From a conceptual point of view it means the development of new probabilistic models for publications and the fusion of publication models, domain models and conditional models.

The idea of an integrative probabilistic framework led to the development of the following concepts, methods and systems

1. *Annotated Bayesian network based information retrieval*, a model-based, personalized information retrieval method (see Chapter 5);

2. *Bayesian network based text-mining*, literature mining with causal, probabilistic publication models (see Chapter 6);

3. *First-order probabilistic knowledge bases based on Annotated Bayesian network*, the concept of embedding complex posteriors in logical knowledge bases (see Section 5.2);

4. *Complex Bayesian network features for classification*, the concept of the Markov Blanket subgraph (MBG) feature[†] and the Bayesian inference method for Markov blankets (MB) and MBG features, which is an integrated estimation and search method using the sorted (ordering conditional) MBG space (see Section 7.2 and Alg. 1);

[†]We follow the general practice that the term feature is used as a descriptor of the domain (i.e., a domain variable) and as a property of a model as well.

5. *Bayesian, four-level, sequential analysis of relevance*, the analysis of relevant variables at the levels of pairs of variables, sets of variables, submodels, and models (i.e., at the levels of Markov Blanket Memberships, Markov Blanket sets, Markov Blanket graphs, and Bayesian networks);

6. *Prior transformation methods*, a Bayesian network based method to induce informative priors for parametric black-box classifiers, and the evaluation of the advantages of priors in classification (see Chapter 10);

7. *Probabilistically linked model spaces*, the concept of literature based "posterior priors" for domain models and the concept of induced priors for conditional models from domain models (see Section 8.1 and Chapter 10).

These concepts and methods were responses to the following challenges in biomedicine such as the availability of electronic prior knowledge, the flourishment of Bayesianism and the growing importance of data exploration and knowledge discovery beside hypothesis driven research.

1. *Expert knowledge, literature and data.* How can we support knowledge elicitation, information extraction, knowledge discovery and statistical data analysis in a joint manner?

2. *Probabilistic logic.* How can we fuse logic and probability theory, specifically publications, free-text annotations and the results of Bayesian inferences about complex models?

3. *Domain and conditional modeling.* How can we combine the advantages of domain and conditional modeling, such as interpretability and the existence of prior vs. lower computational complexity and better performance?

4. *Probabilistically linked models.* How can we use multiple models with heterogeneous data such as literature data and clinical data in a semantically transparent and computationally efficient way?

The developed results are illustrated with the following examples. Anticipating Chapter 3 about Bayesian networks, this model class uses directed acyclic graphs (DAGs) to represent a probability distribution and optionally the causal structure of the domain. In an intuitive causal interpretation, the nodes represent the uncertain quantities, the edges denote direct causal influences, defining the model structure. A local probabilistic model is attached to each node to quantify the stochastic effect of its parents (causes). Fig. 1.1 shows an artificial Bayesian network structure G using variables from the OC domain. It introduces also two central concepts of the thesis, the Markov Blanket set and the Markov Blanket Graph of a given target variable Y in DAG G. The Markov Blanket set of variable Y in DAG G denoted with $\mathrm{MB}(Y, G)$ is a sufficient set of variables to shield probabilistically Y from the rest of the variables. The Markov Blanket set $MB(Y, G)$ induces the pairwise Markov Blanket Memberhip relation denoted with $\mathrm{MBM}(Y, X, G)$, which corresponds to the general concept of

relevance/irrelevance (i.e., conditional probabilistic dependency/independency). The Markov Blanket Graph of variable Y in DAG G MBG(Y, G) includes also the incoming edges into Y and into its children in DAG G.

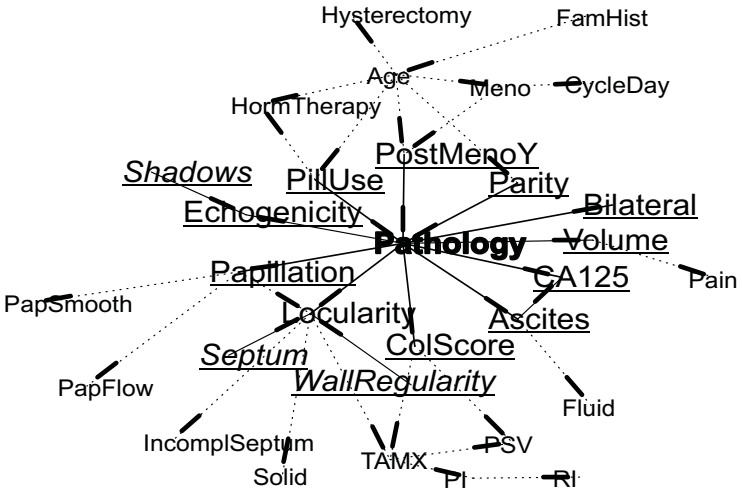


Figure 1.1: An artificial Bayesian network structure G showing also the Markov Blanket and the Markov Blanket Graph of a target variable Pathology. Underscore denotes the Markov Blanket set MB(Pathology, G) (i.e., the members of the Markov Blanket set MBM(Pathology, X, G)). Italic (with underscore) denotes conditionally relevant variables (i.e., if a variable is pairwise irrelevant, but it is relevant of another variable is known). Smaller font size denotes the irrelevant variables. Solid lines denote the edges of the Markov Blanket Graph MBG(Pathology, G).

Example 1.1.1. *Annotated Bayesian network based information retrieval.*

Let us assume that we are in the middle of a knowledge elicitation or a data analysis session with our domain experts using Bayesian networks. We have a partially specified probabilistic domain model, a pile of papers about the domain, a mass of notes about multiple aspects and levels and we try to find further related papers either to extend our prior model or to interpret and evaluate the inferred model. How can we formulate a model-based and personalized information retrieval query using our fragments, comments and papers collected about the model? Because of the separation of the information retrieval, knowledge engineering and inductive techniques, this task was dependent on the interplay of a domain expert and data analyst or knowledge engineer. To support the integration using the electronic literature we developed a query language and implemented an information retrieval system capable for incorporating annotated Bayesian network fragments into the query. The following query expresses the information need about a variable CA125 and its influencing factors (relevant variables) in the ovarian cancer context with emphasis on "Meigs-syndrome" (see Chapter 5). The relevant variables are referred as the Markov Blanket of

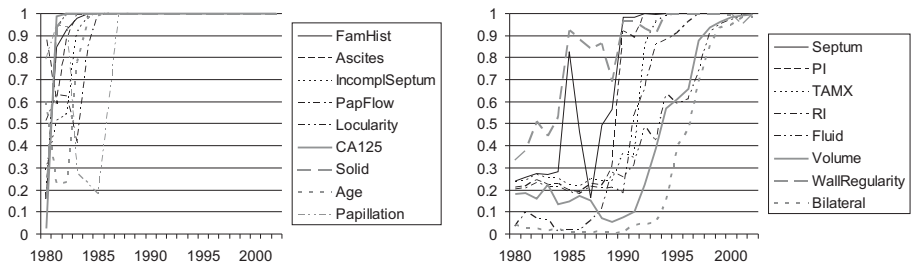


Figure 1.2: The temporal evolution of the collective belief — inferred from the literature — that a given variable is relevant for the preoperative diagnostics of ovarian cancer. Belief in (pairwise) relevance is represented by the posterior of the MBM feature, thus the figure shows the sequential posteriors of the MBM(Pathology,X_i,G) relations with variables X_i with fast/slow convergence to 1 using the temporal sequence of publications between 1980 and 2005 in the large PubMed corpus binarized with corelevance, $\mathrm{BD}_{\mathrm{eu}}$ priors, noninformative structure priors and conditionally on the expert's total causal ordering.

the variable CA125 in a given Bayesian network structure G (MB($CA125, G$)) and $\mathcal{A}$ denotes annotations attached to various parts of the model.

$$\text{“CA125}''\text{, } \mathcal{A}(\text{MB(CA125, G)}), \ \mathcal{A}(\text{IOTA}), \ \text{“Meigs}'\text{syndrome}''$$

Example 1.1.2. *Bayesian network based text-mining.*

The ABN-IR system can help us to find further related papers to extend our prior model for example with new structural aspects, but it is usually a time-consuming task to extract and weight structural relations. A variety of information extraction techniques can be applied for the automation of this step, with linguistic or statistical roots, but these methods by definition have a bottom-up characteristic: they assume explicit statements of the target relation under reconstruction and the domain experts integrate them into an overall prior domain model. First we experimented with such co-occurrence and co-relevance based information extraction methods, but later we proposed a top-down knowledge discovery method using Bayesian networks (see Chapter 6). This method infers a confidence for relevance relations by Bayesian averaging over generative publication models. It can discover prior causal information even if only associated domain entities are reported in the literature. Fig. 1.2 shows the sequential posteriors of the relevance of the variables w.r.t. the type of the ovarian tumor using the publications between 1980 and 2005 (see Chapter 6).

Example 1.1.3. *Probabilistically linked model spaces.*

The introduced probabilistic publication models allow the definition of an overall hierarchical metamodel including probabilistic models for corpora of the literature and for the real statistical data sets. We discussed this data level

fusion and proposed an approximation using probabilistically linked models at the level of model features (see Chapter 6 and Section 8.1). Basically it uses the transformed posteriors of model features given the literature as prior in a subsequent inference phase as shown in Fig. 1.3.

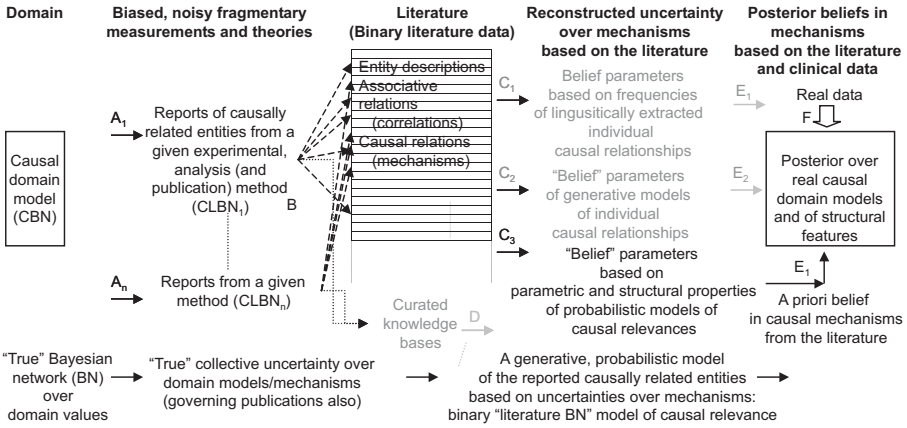


Figure 1.3: The reconstruction of prior knowledge in a biomedical domain from literature data and its incorporation in learning causal domain models. The steps show the sources of mechanism uncertainty, their generative function in publications, the discovery of mechanism uncertainty and the incorporation of the reconstructed mechanism uncertainty in Bayesian inference methods. Arrows $A_1, \ldots, A_n$ indicate generative models of causal relevances from various points of view, such as different experimental setup, analysis method and publication style. Arrow B denotes their publication. Arrow C indicates usage of the overall publications to integrate various fragments into a combined causal domain model. Arrow D indicates that the accepted domain theories are represented in the knowledge bases and are later transformed into a priori distribution for the subsequent Bayesian learning. Arrow E and F shows the Bayesian fusion of the reconstructed mechanism uncertainty as prior with real data.

Example 1.1.4. *Probabilistic Annotated Bayesian network knowledge bases.*

The development of a logical knowledge base for the prior knowledge including free-text annotations, references to standard knowledge bases and to publications raised the issue of the integration of complex posteriors over the publication models and domain models. For this problem, we proposed the use of such first-order probabilistic knowledge bases, in which complex distributions are embedded in a logical knowledge base (see Def. 5.2. We discussed a model based and syntactic interpretations of the induced probability over sentences of such a probabilistic annotated Bayesian network knowledge base and discussed the applicability of an ordering-based MCMC method for features having an order conditional conjunctive normal form (see Section 8.1 and 7.1.6). For example the probability of the following sentence expresses the posterior belief

that in domain G there is a causal link from variable *Age* to *Locularity* and the annotations ($\mathcal{A}$) of all its edges e are rated as relevant by the expert (see Section 5.2 for details).

$$DPath(G, Age, Locularity) \wedge \forall e DEdge(G, e) \Rightarrow Contain(\mathcal{A}(e), \text{“}relevant''\text{)}$$

Example 1.1.5. *Complex Bayesian network features for classification.*

The probabilistic annotated Bayesian network knowledge base allows the formulation of unrestricted first-order sentences including structural model properties, but the estimation of their truth value (i.e., their probability) poses a serious computational challenge. Because of our interest in classification, we tried to identify structural model properties sufficient for classification for which an efficient estimation method exist. We proposed the Markov Blanket Subgraph (MBG) feature as an ultimate feature from the point of view of conditional modeling, a.k.a. Mechanism boundary subgraph, and classification or feature subgraph (see Fig. 1.1 and Section 7.2). We generalized the feature subset selection (FSS) problem — which corresponds to the Markov Blanket set (MB) feature — by formulating its equivalent at the level of the MBG feature, as the feature (sub)Graph Selection (FGS) problem (see Def. 7.2.3). Then we formalized the *Most Probable Features* problem (MPFs) (Def. 7.6.1) and analyzed the effect of feature cardinality on estimating and selecting the optimal features (see Th. 7.6.1). We proposed an integrated Monte Carlo estimation and search method based on the truncated MBG-ordering space (see Alg.1). We demonstrated that a full Bayesian inference over the feature sets and feature subgraphs is feasible, which allows a new, separate level of data analysis. Based on this we developed a “*Bayesian, four-level, sequential analysis of relevance*” at the levels of Markov Blanket Memberships, Markov Blanket sets, Markov Blanket graphs, and complete Bayesian networks (see Section 8.5).

Example 1.1.6. *Prior transformation methods.*

Beside the structural aspects of the domain model, the numerical values of the model parameters were also investigated in the thesis. In this case the literature was processed only manually and the domain expert provided prior estimates taking into account the literature, so we had no distinct literature based parameter priors. Our primary interest was to transform such informative priors into priors for classification systems and investigate their effects. First we evaluated the value of parameter estimates in the original model class used for its elicitation. We used a hyperparameter to express a global confidence in the parameters, which has a counting interpretation as the number of complete cases incorporated into the estimates of the parameters. As the posterior of this hyperparameter shows in Fig. 8.2, the prior estimates correspond to approximately 150 cases with this data set, which agrees with our expectations (see Section 8.2 for details). The next challenge was to integrate this parameter prior for a particular domain model with a classification oriented model, which in our

MB_1={ Age, Bilateral, Volume, Ascites, Septum, IncomplSeptum, Papillation, PapFlow, WallRegularity, Shadows, Echogenicity, ColScore, CA125, PI, RI, PSV, TAMX, Solid}
MB_2=MB_1 \{PSV}
MB_3=MB_1 ∪{Meno}\{Age}
MB_4=MB_3 \{PSV}
MB_5=MB_1 ∪{FamHistBrCa}

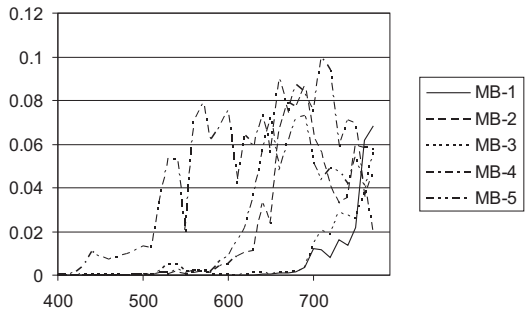


Figure 1.4: The temporal evolution of the belief — inferred from growing amount of clinical data — that a given set of variables is (exactly) relevant for the preoperative diagnostics of ovarian cancer. Belief in relevance is represented by the posterior of the MB feature, thus the figure shows the sequential posteriors of high-scoring MB(*Pathology*) feature values given the expert's total causal ordering and using the temporal sequence of the IOTA-1.2 data set, BD_{eu} priors and noninformative structure priors. These posteriors are less than 10^{-6} for sample size less than 400, so the x-axis starts from this value. The ten most probably MB sets are defined in Table A.7.

case was a multilayer perceptron. Again, as in the case of publication models and domain model, where we suggested the use of a two-step literature based posterior prior, we proposed an analogous approximation to an overall meta-model merging BNs and MLPs. We proposed transformation methods to induce an informative parameter prior for a given multilayer perceptron structure from the prior of a Bayesian network. Fig. 1.6 shows this two-step methodology using a hybrid BN-MLP representation for the fusion of knowledge and data in classification (see Chapter 10 for details).

Finally we evaluated the effect of parameter and structural priors on the predictive performance of domain models and classification models. Fig. 1.7 reports the detailed effect of the parameter prior incorporation for varying proportions of samples used in the training set, which shows that the induced informative prior is efficient in the small sample region and not restrictive in the large sample region (i.e., if the sample size is less or much larger than the number of free parameters, see Section 10.6 for details)

To describe the background and clarify the joint works with my colleagues, I summarize the chronological overview. The contributions of the thesis are enumerated in Section 11.1.

1.2 Chronology of doctoral activities

1. *Using prior domain knowledge formalized as a Bayesian network in classifier construction.* The proposal of using Bayesian networks to organize and formalize prior domain knowledge and to support the construction of a specific classifier was the starting point for the thesis [10]. Its central idea was to induce informative structure and parameter priors for a

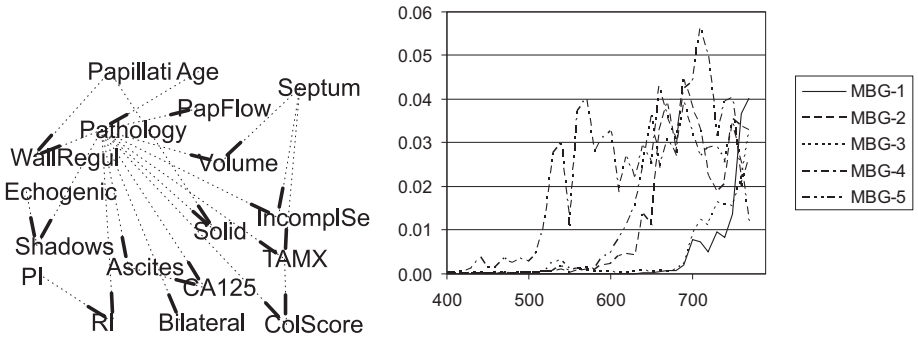


Figure 1.5: The temporal evolution of the belief — inferred from growing amount of clinical data — that a given subgraph over the subset of the variables is (exactly) relevant for the preoperative diagnostics of ovarian cancer. Belief in relevance is represented by the posterior of the MBG feature. (Left) The maximum a posteriori MBG subgraph (MBG-1). (Right) The sequential posteriors of high-scoring MBG(*Pathology*) feature values given the expert's total causal ordering and using the temporal sequence of the IOTA-1.2 data set, $\mathrm{BD}_{\mathrm{eu}}$ priors and noninformative structure priors. These posteriors are less than 10^{-6} for sample size less than 400, so the x-axis starts from this value. The reported MBGs are defined in Table A.9.

parametric conditional model by projecting a domain model.

2. *The transformation of Bayesian network parameter prior into a multilayer perceptron parameter prior using model projection and virtual sample.* The general proposal of deriving informative parametric priors for parametric black-box classifiers has been tested in the case of multilayer perceptrons [18, 11, 15, 14]. This work has been done mostly in 2000 in cooperation with Geert Fannes, who developed and implemented the proper treatment of parameter priors for multilayer perceptrons with respect to symmetries in the parameter space. These results can be found in his doctoral thesis, with many of his extensions, for example to use continuous Bayesian networks to represent the parameter prior [85].

3. *Web-based medical data collection, quality management and preprocessing.* The participation in the data collection of the IOTA project in $2000-2002$ provided an excellent opportunity to become familiar with the real world data set used in the thesis, particularly to have an overview of the process of the web-based medical data collection and quality checking [5].

4. *Integrated analysis of microarray data, gene annotations and literature with clustering.* The integrated usage of expert beliefs, expert annotation, domain literature and statistical data was investigated in case of clustering algorithms as well. The implemented text indexing and mining system has provided the foundation in 2001 to develop a prototype system for the automated textual analysis of gene clusters (TXTGate). On the one hand it

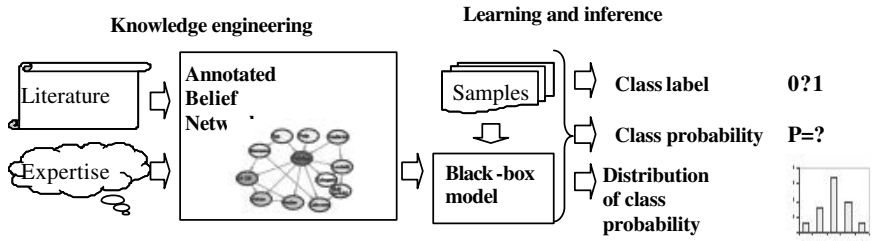


Figure 1.6: The two-step methodology covering the fusion of knowledge and data for classification. First, we formalized the prior domain knowledge in a Bayesian network. Second, we induced informative structure and parameter priors for parametric conditional models to support various Bayesian inferences. The Bayesian approach to classification can target three levels: the class label (discrimination), the class probability (regression) and the distribution of the class probability (right).

performed clustering in the "literature world" of gene annotations and domain literature and on the other hand it provided various textual profiling of the clusters to support clustering in the "data world" of microarrays. First results about its application were reported in [19, 20] in cooperation with Patrick Glenisson, who was responsible for clustering and evaluation. Related results can be found in his doctoral thesis "Integrating scientific literature with large scale gene expression analysis" [113], describing also the developed internet service TXTGate [115].

5. *Model and domain explorations by ABN-KB keyword profiles.* The construction of Bayesian network models annotated with expert textual comments and links to domain literature, together with the implemented text indexing and mining system has provided the foundation in 2001 to develop and implement an "Annotated Bayesian network"-based information retrieval language to support contextualized (personalized and domain-specific) information retrieval in cooperation with Tamás Mészáros from the Budapest University of Technology and Economics [22, 23].

6. *Bayesian network based statistical analysis of domain literature.* After investigation of the pairwise, associative statistical analysis of the literature in 2001, the next phase was the domain model based statistical analysis of the domain literature. The proposed model based approach is aimed at discovering latent causal knowledge in contrast to the individual relation based, associative text mining methods. Furthermore, the Bayesian network based statistical analysis of domain literature offers a causal, generative foundation for prior elicitation from the literature [20, 13, 16, 26].

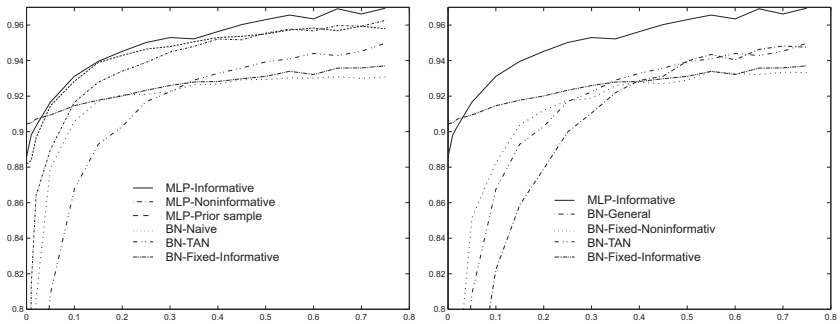


Figure 1.7: The learning curves for the multilayer perceptron models using an informative prior (MLP-Informative), a noninformative prior (MLP-Noninformative) or prior samples (MLP-Prior sample). For the Bayesian network models, the learning curves correspond to the Naive Bayes structure (BN-Naive) with noninformative prior, a search in the generalized tree-augmented networks (BN-TAN) with noninformative prior, and to the fixed prior structure in combination with the informative prior (BN-Fixed Informative) (left). The other figure shows the learning curves for the multilayer perceptron and Bayesian network models using an informative prior (MLP-Informative and BN-Fixed Informative) in comparison with three Bayesian network models using a noninformative prior in combination with a search over the generalized tree-augmented network space (BN-TAN), the fixed prior structure (BN-Fixed Noninformative) and a general Bayesian network structure learning algorithm (BN-General) (right). The x axis indicates the proportion of samples used for training while the y axis represents the corresponding area under the ROC curve.

7. *Integrated analysis of expert beliefs, expert annotation, domain literature and statistical data with Bayesian networks.* A pairwise, associative approach towards an integrated analysis of expert beliefs, expert annotation, domain literature and statistical data in Bayesian network learning was reported in 2002. In this case both the elicitation from a domain expert and the text mining method using the expert annotation and domain literature has produced prior beliefs over pairwise relations, which were cross-compared and evaluated against the corresponding data scores [16]. The multivariate extension of the analysis with complex features was devised in 2003, such as the Markov Blanket subgraphs [25, 21]. Additionally, since both the medical data set and the literature data is temporal, the Bayesian inference over complex structural features was expanded with a sequential analysis.

8. *Evaluation of new parameter priors and multivariable structure priors.* The last phase of elicitation of expert beliefs over structure priors and over parameterization for Bayesian networks with the new IOTA variables has been performed in 2003.

9. *Informative structural and parameter priors for parametric Bayesian clas-*

sifiers. We evaluated the classification performance of various Bayesian network classifiers, such as naive, tree-augmented and general Bayesian network classifiers, and of Bayesian logistic regression and multilayer perceptron models extensively in 2000 and 2001, particularly with respect to the effect of priors and in the Bayesian context with rejection [28, 18, 11, 15, 17, 12, 14]. The new classification oriented Markov Blanket spanning subgraph features allowed to accomplish the original goal from 1998 to derive priors also for the parameter structures of conditional classifiers.

1.3 Chapter-by-chapter overview

The structure of the dissertation follows the phases of the construction of a classification model with the dual goal of understanding the domain and of performing predictions. It starts with preparing domain resources, then exploring, extracting, formalizing and transforming priors, finally using it in Bayesian inference. Chapter 2 reviews the Bayesian framework, particularly the Markov Chain Monte Carlo methods and the sequential model evaluation. Chapter 3 summarizes the representation, inference and learning of Bayesian networks. In Chapter 4 we introduce the ovarian cancer domain. It contains the description of the clinical data sets from the IDO project at the K.U.Leuven) entitled "Predictive computer models for medical classification problems using patient data and expert knowledge" and from the IOTA project, which is a multicenter study by the "International Ovarian Tumor Analysis" consortium. It describes the original and the derived electronic resources, such as the literature data sets. It summarizes the results of knowledge engineering including the elicited expert knowledge and the results of various checks and evaluations. Chapter 5 first describes a fusion method of complex distributions and logical knowledge bases, specifically for the fusion of distributions specified by BNs or over BNs and textual knowledge bases. Then it presents a Bayesian network-based information retrieval language for annotated Bayesian network to support the knowledge engineering of complex Bayesian networks in the "e-science" era. Chapter 6 describes the statistical analysis of the domain literature with Bayesian networks. It characterizes the proposed Bayesian network based analysis by positioning it in the spectrum of text mining methods from shallow statistical approaches to linguistic approaches. Chapter 7 describes methods how to perform Bayesian inference over complex Bayesian network features, particularly over classification oriented features. It introduces a special feature called Markov Blanket spanning subgraph or Mechanism Boundary subgraph feature, discusses its relevance for conditional modeling and for causal modeling. Chapter 8 contains the results of the learning of Bayesian networks from heterogeneous sources, that is the integrated analysis and fusion of heterogeneous information resources. It contains results about comparing and combining expert prior knowledge, literature data, medical data on different levels, such as pairwise, higher order feature and complete domain model level. Chapter 9 is an overview of Bayesian classification, specifically the use of domain models as classifiers, the Bayesian

conditional modeling and particularly the multilayer perceptron model and its relation to the MBG feature. Finally, Chapter 10 discusses derivation of informative structure and parameter priors for parametric black-box classifiers. It demonstrates of the proposed methodologies in the ovarian cancer domain and the performance of various classification models with informative priors.

Chapter 2. A Bayesian primer

This chapter introduces the Bayesian framework including the interpretation of probability, the interpretation of models (parameters) and its link to decision theory. Secondly, it summarizes certain techniques applied in the thesis, particularly the Markov Chain Monte Carlo methods to perform Bayesian inference, the approaches to model averaging and the sequential model evaluation for model selection and data exploration.

Chapter 3. A Bayesian network primer

This review chapter starts with discussing the representational power of Bayesian networks, including the interpretation of its parameters and structure. It provides an overview about the forms of independencies such as conditional, observational, interventionist, or contextual, and about the structure of independencies of a probability distribution and its representation with directed acyclic graphs. It summarizes the statistical equivalence of such graphical representation and the possible causal interpretation of directed acyclic graphs. Continuing the Bayesian approach to Bayesian networks, various inferences in Bayesian networks are summarized as we can perform probabilistic inferences over domain values, over parametric and structural properties of the model as well. First, the three layers of probabilistic inferences about domain values are summarized, the case of fixed structure and parameterization, the case of fixed structure and a prior over the parameterization and the case of the full Bayesian approach with priors over the structures and parameterization. Next, the probabilistic inference over structural features of the Bayesian network model are outlined in the full Bayesian context (i.e., with prior over the structures), though the computational details are elaborated later in Chapter 7. The final topic in this chapter is the learning of Bayesian networks.

Chapter 4. Prior knowledge and data about ovarian cancer

The chapter starts with an overview of the ovarian cancer domain, then it documents the results of knowledge engineering. On the one hand, it presents the elicited expert knowledge such as the textually, qualitatively and quantitatively characterized domain variables, pairwise relations and complete domain models, partly with complete parameterization. On the other hand, it presents the description of the automatically collected original and the derived electronic resources, such as the literature data sets. Then it continues with the description of the medical statistical data sets from the IDO and the IOTA projects.

Chapter 5. Fusing BNs and logical knowledge bases

This chapter first discusses the role of knowledge engineering within the Bayesian data analysis as prior formulation when significant amount of electronic prior knowledge and statistical data are available (i.e., Bayesian knowledge engineering). Second it describes a fusion method of complex distributions and logical knowledge bases, specifically the fusion of distributions related to Bayesian networks and textual knowledge bases. Third it describes a numerical vector representation (i.e., statistical keyword profiles) of parts of ABN-KBs, which are formalized as elements of a language for ABN-KBs. The ABN-KB based keyword profiles have multiple roles. First, they allow the exploration of the knowledge base (e.g., the exploration of a complex Bayesian network) by direct browsing, by clustering of the profiles or by the visualization of the similarity of the profiles. As the profiles are part of the ABN-KB language, this extension allows more complex sentences with the standard probabilistic semantics of the ABN sentences (e.g., based on the posterior of BN structures). Second, if the knowledge base is expanded with a collection of domain publications, the keyword profiling relations and functions can be applied on the publications as well, which allows the integrated exploration of the knowledge-model and the domain literature. A simple example of this integrated exploration is the identification of relevant publications for a given aspect of the ABN-KB. We report this usage, which supports contextual information retrieval by providing a personal and domain-specific context through keyword profiles. Another kind of integrated exploration of the ABN-KB and the domain literature is reported in the next Chapter, in which certain ABN-profiles are used for text-mining.

Chapter 6. Statistical text mining with BNs

This chapter first provides an overview of various text mining methods from linguistic to shallow statistical approaches for knowledge discovery and information extraction. Then it proposes a Bayesian network based text mining method that is oriented towards underlying generative models of associative patterns (i.e., towards a consistent collection of relations forming a domain model). It characterizes the applicability of various Bayesian network structures for the statistical analysis of the occurrence patterns of domain concepts in the domain literature and discusses their interpretation.

Chapter 7. Bayesian inference over BN features

This chapter first overviews various methods for learning properties of a Bayesian network and to induce confidence measures for such properties, including bootstrap. Then we provide a taxonomy of Bayesian network features, including a structural feature called Markov Blanket spanning subgraph or Mechanism Boundary subgraph feature. This complex feature embodies a classificational submodel and it is on an intermediate level between simple features such as edges or Markov Blanket memberships and complete Bayesian networks. Next,

we describe general Bayesian methods to perform Bayesian inference over semantic propositions about structural features of Bayesian networks (i.e., methods to compute or approximate the posterior probabilities of sets of Bayesian networks with arbitrary properties, including textual conditions). Next, we describe methods to perform Bayesian inference over Markov Blanket spanning subgraph features and over simpler structural features defined over them.

Chapter 8. Analysis of heterogeneous information

This chapter is about the integrated analysis of heterogeneous information resources. It presents a unified probabilistic fusion of expert prior knowledge, literature data and real, statistical data at the level of data, model features and domain models. The chapter starts with the sequential evaluation of the expert priors. Then it describes the comparison of the expertise, literature and data at the pairwise level using visualization methods, rank statistics and classification correspondence. Next, we compared complete causal domain models from experts, literature and data using pairwise and multiparental, causal difference measures. Then the comparison is performed at the level of conditional features. Finally, we report the effect of incorporating expert priors and priors from text mining in Bayesian inference with medical data.

Chapter 9. Bayesian classification

This review chapter first outlines the Bayesian approach to classification, particularly the use of Bayesian networks and multilayer perceptrons. It overviews performance measures for discrimination and for prediction of probabilities, including a discussion of classification with rejection. Next, it discusses the application of Bayesian networks for classification, particularly the tree-augmented Bayesian networks. The chapter summarizes the logistic regression, its relation to Bayesian network classifiers and its extension to the multilayer perceptron.

Chapter 10. Bayesian classifiers with a prior domain model

This last chapter discusses the applicability of prior domain knowledge formalised as Bayesian network in the process of construction of a classifier. It describes the developed projection-based method and a virtual sample based prior transformation methods from Bayesian networks into parametric black-box classifiers, such as logistic regression and multilayer perceptron models. Then we present the joint posterior of various conditional features and performance measures, which allows the derivation of structure priors for such regression models. Finally, we evaluated the classification performance of Bayesian network classifiers and logistic regression models with informative priors.

Chapter 2

A Bayesian primer

We outline the Bayesian decision theoretic framework and define some of its concepts. We also summarize some practical aspects of the Bayesian framework such as performing Bayesian inference with Monte Carlo simulations and evaluating models in the prequential framework.

In this thesis, uncertainties are formalized exclusively within the framework of probability theory. There are numerous reasons for the probabilistic approach, particularly for its subjective interpretation, that is for the Bayesian approach. We will discuss some of these concepts in relation to an application in ovarian cancer diagnosis. We indicate the main points of an axiomatic argument based on decision theoretic considerations [34]. For a discussion of the advantages and disadvantages of the Bayesian approach in statistics, see [31, 214], in artificial intelligence, see [43]; for a historical outlook and recent trends, see [32, 183, 9].

Subsequently we will summarize Monte Carlo methods to estimate expectations and to provide confidence measures for the estimates as well, though the sampling methods can be equally used to explore the posterior. We use mainly the following works: [108, 111, 194, 176, 102, 120, 171]. For an overview of analytic approaches to evaluate expectations if analytic forms of the posterior are available, see [34, 108]. For a detailed treatment of Monte Carlo approaches to compute other quantities such as credible regions, and Bayes factors, see [45].

Next we discuss a method based on the sequential evaluation of the predictive performance of the model, called "prequential analysis". Because of its sequential nature, it is capable to provide a detailed sample-by-sample compatibility of the data and the model, which is particularly relevant if the data is ordered. Finally we discuss methods to support the analysis of the posterior, namely to find model classes with large posterior probability and modes of the posterior.

2.1 The subjective interpretation of probability

Regarding the variety of events present in knowledge and data analysis, it seems to be an oversimplification to express the uncertainty over these events with a scalar quantity subject to the axioms of probability theory (for the moment we discuss only finite and discrete events). We assume that the uncertainty is related to the occurrence of events and not to the relevance of the event system itself (i.e., not to the applicability of propositions representing events). The establishment of a system of complete and mutually exclusive events is particularly challenging in medical applications because of the multiple levels of analysis, the contextual (conditional) and ambiguous definitions (for an attempt to establish a terminology in ovarian cancer diagnosis by ultrasonography with well-defined meaning of quantitative measurements see [240]). Assuming a proper event system, the following interpretations were proposed for the probabilistic representation of uncertainties over these events. The physicalist or propensity approach relies on some inherent randomness of the events [206]. The frequentist approach recourse to the limiting relative frequencies of certain types of events (for an overview within a computational framework, see [251]). The axiomatic approaches formally deduce the existence and uniqueness of subjective (personal) probabilities corresponding to optimal decision in a decision theoretic framework (for a formalization and references, see [34]). The instrumentalist approach takes a pragmatic point of view evaluating indirectly the usage of subjective probabilities as a modeling tool [60, 108, 69].

The usage of probabilities with subjective interpretation to represent uncertainties over outcomes is only the first step towards the Bayesian framework. The so-called representation theorems show that the assumption of infinite exchangeability (i.e., that beliefs are independent of the ordering of the observations) implies *as if* the observables are conditionally independent random samples from a sampling distribution with parameter θ and θ itself have a probability distribution representing beliefs over its limiting values. Whereas this interpretation can be criticized on the ground of the asymptotic nature of these results (note the infinite exchangeability assumptions and that certain finitely exchangeable sequences have no mixture representation, see p226,[34]), this provide the second part of axiomatic foundations for the Bayesian framework.

Chaining these together, according to the adopted subjectivist interpretation of probability, the uncertainties over outcomes are represented with probabilities as beliefs over the outcomes, which can be represented as a mixture of parametric distributions with a probability distribution over its parameter expressing beliefs in its parameterization.

2.2 The general scheme of Bayesian inference

Irrespectively of whether the axiomatic approaches for the subjective interpretation of probability and the Bayesian approach are accepted as normative or suggestive, from an instrumentalist standpoint the *Bayesian framework* is con-

ceptually very simple. First, in a certain context ξ one can express his beliefs $p(x|\xi)$ in observable quantities x by specifying his belief $p(\theta|\xi)$ in (unobservable) parameter quantities θ of relevant parametric models $p(x|\theta)$ (ξ^+ and ξ^- denote the availability and lack of relevant background knowledge). Second, one can use the joint distribution $p(x, \theta|\xi)$ according to the rules of probability theory for performing any inference over observable and parameter quantities. The inference provides standard probabilistic conclusions $p(\alpha(x, \theta)|\xi)$ reflecting his personal belief in the proposition α with respect to this context, where the proposition about x and θ usually includes parts of the background knowledge ξ as well. This illustrates one of the main strengths of Bayesianism which is that throughout the process from setting up the model to the final inference uncertainties are expressed exclusively in a single coherent system of probabilities.

Before examining in detail various forms and properties of such inference, note that we follow a standard notation in probability theory: using capitals for random variables and possibly the same lower case letters for values $P_X(X = x)$, omitting the names of the random variables, the indication of their distributions and the range of summation or integration if it is unambiguous. The same notation $p(.)$ will be used for probability mass functions and densities, and we use the terms density and distribution mass function interchangeably. Furthermore, if possible X denotes the explanatory or independent variable, Y the outcome or dependent variables and D denotes the observed data set. In the sequel we assume that the data set consists of N complete cases $D_N = \{x^{(1)}, \ldots, x^{(N)}\}$ (i.e., each variable $X_i \in \underline{V}$ is observed). The parameters of Bayesian networks and multilayer perceptrons are differentiated with θ and ω. If necessary, vectors are differentiated with underline and double underline denotes matrix.

2.2.1 Setting up the model

In an idealistic Bayesian approach the family of the included models should be as broad as possible expressing beliefs in any potentially relevant model. However, three issues have to be considered: the potential violation of the principle of Ockham's razor, the computational difficulty to cope with such a large class of models and the practical difficulty of specifying a priori beliefs. The first objection against Bayesianism related to Ockham's razor's preference for simplicity can be rejected based on the explanation that, put it simply, more general models corroborated less than more specific models if they fit to the observations [146, 175, 194]. The second computational issue is treated in Section 2.3, basically relying on the increased availability of computational power to perform Bayesian inference with stochastic simulations. The third issue of specification of a priori beliefs for a wide range of models is a central theme of the thesis, for now we discuss only the concept of hierarchical modeling that we need for the exposition of the Bayesian framework.

The set up of *hierarchical models* involves exchangeability considerations as discussed in Section 2.1 (but now at the level of parameters θ) to validate a mixture representation and leads to the concept of *hierarchical priors* using *hyperparameters* ϕ:

$$p(\theta, \phi) = p(\phi)p(\theta|\phi). \tag{2.1}$$

A frequently occurring form in practice is that the specification is usually achieved by a structured specification of the relevant models using model classes $\mathcal{M}^i$, model structures $\mathcal{S}_k^i$ or M_k^i and parameters θ_k^i. Correspondingly the a priori belief in a given model from model class i with structure k and parameters θ_k^i can be expressed as a product

$$p(\theta_k^i, M_k^i, \mathcal{M}^i) = p(\mathcal{M}^i)p(M_k^i|\mathcal{M}^i)p(\theta_k^i|M_k^i). \tag{2.2}$$

These specifications together with the conditional probabilistic model of observable quantities $p(x|\theta, \phi)$ or $p(x|\theta_k^i, M_k^i)$ provides the joint distribution.

2.2.2 Predictive inference

The specification of the a priori beliefs over relevant models allows us to perform (prior) predictive inferences over the observable quantity x

$$p(x) = \sum_k p(M_k) \int p(x|\theta_k)p(\theta_k|M_k)\,\mathrm{d}\theta_k. \tag{2.3}$$

The operation of integration or summation over models and their parameterization implements marginalization and is termed in this context as *Bayesian model averaging* [177, 180, 178, 136]. Postponing momentarily the discussion of the a posteriori beliefs after observing a data set D, we can write the *posterior predictive distribution* conditioned on the data set D as

$$p(x|D) = \sum_k p(M_k|D) \int p(x|\theta_k)p(\theta_k|D, M_k)\,\mathrm{d}\theta_k. \tag{2.4}$$

These equations illustrate that prediction in the Bayesian framework has the following distinctive property w.r.t. the frequentist framework: it averages over models (i.e., there is no model selection). Whereas this a normative result to perform general predictive inference dictated by the axioms of probability theory, related results about the advantage of specialized model averaging and approximate model averaging have also been reported (for the case of binary classifiers, see [128]; for regression models in the committee framework, see [36]; for Bayesian networks, see [180]; for an overview, see [136]). As we will see in Chapter 3 and 9, frequently the integration and in special cases even the summation and the integration can be performed analytically, otherwise stochastic simulation methods discussed in this chapter and in Chapter 7 can be used to approximate the inference.

In a certain sense the predictive distributions are the target of the Bayesian framework and the models are secondary devices, particularly in an instrumentalist interpretation of the Bayesian framework. Consequently, the ideal result

of Bayesian analysis is the full report of the predictive distribution. If the report of the full distribution is not possible, other descriptors are discussed in Section 2.2.4.2. Now we continue with another type of probabilistic inference using the a priori belief $p(\theta)$ and the model $p(x|\theta)$.

2.2.3 Parametric inference with Bayes' rule

The specified joint distribution, in which observable quantities and parameters have equal status, allows inference over parameters (i.e., parametric inference conditioned on the observable quantity using the famous *Bayes' rule*):

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{\int p(x|\theta)p(\theta)\,\mathrm{d}\theta} \propto p(x|\theta)p(\theta). \tag{2.5}$$

In Eq. 2.5 $p(\theta)$ is the *prior distribution* or prior, $p(x|\theta)$ is the *sampling distribution* that also defines the *likelihood* and the *likelihood function* $L(\theta; x)$. $p(x)$ is the *marginal likelihood of the data* that defines a normalizing constant and $p(\theta|x)$ is the *a posteriori distribution* of the parameters or simply the posterior. Eq. 2.5 also shows that the posterior is a kind of equilibrium between the prior and the likelihood, and with an increasing number of observations, the posterior is more and more dominated by the likelihood and the effect of prior becomes negligible.

The posterior (parameter) distributions has already appeared in the posterior predictive distribution in Eq. 2.4 after observing the data set D as $p(\theta|D)$, $P(M_k|D)$ and $p(\theta_k|D, M_k)$ (see [34]). In the discrete case the posterior of the model $p(M_k|D)$ is given by

$$p(M_k|D) = \frac{p(D|M_k)p(M_k)}{p(D)} \tag{2.6}$$

where the *marginal model likelihood* or evidence for M_k is

$$p(D|M_k) = \int p(D|\theta_k, M_k)p(\theta_k|M_k)\,\mathrm{d}\theta_k \tag{2.7}$$

and the *marginal data likelihood* is

$$p(D) = \sum_k p(D|M_k)p(M_k). \tag{2.8}$$

2.2.4 Reporting the posterior

The results of an idealized Bayesian analysis can be divided into three categories. The first category includes the exact report of the predictive and parametric posteriors (e.g., by the report of its analytic closed form). The second includes the report of exact values of standard statistical descriptors, such as moments, modes, quantiles, etc. The third group includes various expectations over these posteriors, such as the posteriors of arbitrary domain-specific propositions or expected losses of certain actions, which in the most general case can include

any mixture of predictive and parametric random variables and domain-specific background knowledge.

2.2.4.1 Reporting the posterior distribution

The report of the posterior is easy if it has an analytic form. It can be especially informative if it is from the same parametric family as the prior, which property also shows certain compatibility of the prior and the sampling distribution. In fact, the easy specification of the prior and the tractability of the prior-to-posterior analysis lead to the concept of *conjugate prior*.

Definition 2.2.1 ([108]). *A family $\mathcal{F}$ of prior distributions $p(\theta)$ is said to be conjugate for a class of sampling distributions $p(x|\theta)$, if the posteriors $p(\theta|x)$ also belongs to $\mathcal{F}$.*

The conjugate priors for the broad class of the exponential family are updated to posteriors by the updating their so-called *hyperparameters* using only a summary statistics of the observations (see [108, 34]). In this case the hyperparameters frequently have an intuitive interpretation based on the observations and the prior specification for the parameters corresponds to the specification of summaries of real or virtual past observations (see Th. 3.1.6 for an application with BNs).

Another reason that the posterior frequently has an approximately analytic form is that according to the "Bayesian central limit theorem" under general conditions the posterior of the parameters has a Gaussian distribution [34]. For its application in case of MLP priors, see Section 10.2.3.5.

2.2.4.2 Reporting posterior quantities

If the full report of the posterior over observable quantities or model parameters is not adequate, as it is often the case with complex models or moderate sample size, we can report some standard statistical descriptors corresponding to the posterior or simply the posterior probabilities of arbitrary propositions including even semantic parts from our background knowledge.

If only a value $\hat{x}$ of the observable quantity can be reported in case of x, then the reporting can be interpreted as an action whose utility is specified by a utility or *loss function* $L(x, \hat{x})$. The optimal decision x^* based on the posterior predictive distribution is

$$x^* = \arg\min_{\hat{x}} \int L(x, \hat{x}) p(x|D) \, \mathrm{d}x. \tag{2.9}$$

In the case of parameter estimation with loss function $L(\theta, \hat{\theta})$ and observation x (and prior $p(\theta)$), the optimal estimate $\hat{\theta}$ minimizes the *posterior expected loss*

$$\varrho(p(\theta), \hat{\theta}|x) = \int L(\theta, \hat{\theta}) p(\theta|x) \, \mathrm{d}\theta. \tag{2.10}$$

If the reported and reference values $\hat{x}_i, x_i$ can be interpreted as discrete probability distributions $\hat{p}_i, p_i$, a frequent choice for loss function is the logarithmic loss, which leads to the *cross-entropy* H and to the *Kullback-Leibler* (semi)distance KL, which is always positive and can dominate L_1 and L_2 distances [99, 59]

$$\mathrm{H}(\underline{\underline{p}}\|\underline{\underline{\hat{p}}}) = -\sum_i p_i \log(\hat{p}_i), \tag{2.11}$$

$$\mathrm{KL}(\underline{\underline{p}}\|\underline{\underline{\hat{p}}}) = \sum_i p_i \log(p_i/\hat{p}_i). \tag{2.12}$$

If regions can be reported then the concept of *credible region* and the *highest probability density (HPD) region* minimizing the volume of the region are useful quantities (see Section 8.2.1).

Definition 2.2.2. *A region $C \subseteq Range(\Theta)$ is a 100(1-α)% credible region if*

$$\int_C p(\theta|.)\,\mathrm{d}\theta \geq 1-\alpha.$$

Furthermore, the region C is a highest probability density region if

$$p(\theta_1|.) \geq p(\theta_2|.)\ \forall \theta_1, \theta_2 : \theta_1 \in C, \theta_2 \notin C \text{ almost everywhere.}$$

Definition 2.2.3. *Frequently only the ratios of marginal likelihoods of models M_i and M_j are interesting, the so-called Bayes factor [149]:*

$$B_{ij} = \frac{p(D|M_i)}{p(D|M_j)} = \frac{p(M_j)}{p(M_i)}\frac{p(M_i|D)}{p(M_j|D)}. \tag{2.13}$$

The Bayes factor shows the change of the ratio of prior beliefs to the ratio of the posteriors, which is interpreted as substantial, strong and decisive evidence below 10, between 10 and 100, and above 100 (for applications, see Section 8.2).

2.2.5 Model transformation and reparameterization

We close the general discussion of the Bayesian framework with the issue of model transformation, because of its relevance for Chapter 10. If $\omega = t(\theta)$ is a one-to-one differentiable function with inverse $\theta = t^{-1}(\omega)$ then the transformed density exists and is given by

$$p_\omega(\omega) = p_\theta(t^{-1}(\omega))|\det(\mathcal{J}_{t^{-1}(\omega)})|,$$

where $det(\mathcal{J}_{t^{-1}(\omega)})$ is the determinant of the Jacobian of the inverse transformation $\theta = t^{-1}(\omega)$.

An important consequence is that in general a prior supposed to be neutral (e.g., uniform) will loose its property (e.g., will not be uniform) after a transformation. This led to the concept of the invariance principle and the corresponding Jeffreys' prior and in multidimensional case to its extension of the

reference prior approach [34]. In the thesis we use the term *noninformative* prior as a reference prior, which does not incorporate relevant domain knowledge (such context is denoted with ξ^-). Another consequence of transformation is that the maximum a posteriori value, or in general the modes of the posterior θ^{MAP} are not invariant to parameter transformations, contrary to the invariance of the values maximizing the likelihood function.

2.3 Inference with Monte Carlo methods

The results of Bayesian inference are the predictive and parametric posterior distribution and usually various general statistical quantities and domain-specific quantities defined by the posterior can be reported, such as the optimal observables and parameters with minimal losses, model posterior, posterior probabilities of arbitrary semantic propositions or credible regions, Bayes factors. All the previous examples, except the last two, actually have the same form of an expectation with the posterior. For notational simplicity in this section we will assume that the "target" probability space is defined by a vector-valued random variable $x \in \mathcal{R}^k$ with "target" density $\pi(x)$ and the target function to be integrated w.r.t. $\pi(x)$ is $f(x)$, that is

$$\bar{f} = \mathrm{E}_{\pi(X)}[f(X)]. \tag{2.14}$$

The computation of this integral or summation has a similarly central role in the Bayesian framework as of optimization in the frequentist statistical framework. Furthermore, as we have to resort to probabilistic algorithms such as simulated annealing to approximate global optimization, because of the lack of general deterministic global optimization methods, similarly, the expectation above can be approximated with probabilistic algorithms in general. In the thesis the following two kinds of expectation have to be computed

$$p(x|D) = \int p(x|\theta)p(\theta|D)\,\mathrm{d}\theta, \tag{2.15}$$

$$p(\alpha(M)|D) = \sum_{k=1}^{K} p(M_k|D)\alpha(M_k). \tag{2.16}$$

In the first predictive case we will fix the model structure in the case of logistic regressions and multilayer perceptrons, so there is no summation over model structures (see Chapter 9). In the second case we compute the probability of a structural property of Bayesian networks defined by the sentence α, in which case the parametric integration will have a closed form, see Chapter 3.

2.3.1 Markov Chain Monte Carlo methods

In the case of an unnormalized posterior $\pi(X)$, importance sampling provides a baseline tool to approximate the expectation in Eq. 2.14 and in a resampling

setup it provides a method for the generation of samples from the target distribution. However, a central issue in importance sampling is the iterative checking and refinement of the closeness of the importance distribution to the posterior. From this point of view, it is interesting that there are distribution free, automated methods to perform jointly an iterative approximation to the target distribution, which asymptotically provide samples from the target distribution. The general idea is to construct a stochastic process with a limiting distribution $\pi(X)$ that can be efficiently simulated. The discrete time, homogeneous processes with Markov property are an ideal candidate for this purpose because of their analytic tractability and easy simulations. We start with summarizing the essential concepts and results for discrete time, homogeneous Markov chains with discrete and finite state space, which are mostly used in the thesis, then we discuss a universal construction scheme and its practical application.

2.3.1.1 Markov chains

Let $\mathcal{X} = \{X_0, X_1, \ldots\}$ is a sequence of random variables. The values of X_t are frequently interpreted as states from a state space, the index parameter frequently has a temporal or in biological sequence analysis a location interpretation. In many problems, the assumption of bounded effect is a reasonable assumption, which is formalized by the Markov assumption.

Definition 2.3.1. *A sequence of random variables $\mathcal{X} = \{X_0, X_1, \ldots\}$ is called a (first-order) Markov chain, if $p(X_t|X_{t-1}, \ldots, X_0) = p(X_t|X_{t-1})$. The Markov chain is (time-)homogeneous, if the transition kernel $p(X_t|X_{t-1})$ does not depend on t.*

In this section, unless otherwise stated the values of X_t are discrete and finite, denoted by nonnegative integers $S = \{0, 1, \ldots, K\}$. We use the notation $p^{(t)}$ for the distribution of X_t and $p(X_t = i) = p_i^{(t)}$. We always assume homogeneity, which allows a shorthand notation p_{ij} for the *transition probabilities* as $p_{ij} = p_{ij}^{(t)} = p(X_{t+1} = j|X_t = i)$, which are forming the (one-step) *transition probability matrix* $P = [p_{ij}]$ (a stochastic matrix). Clearly, the "n-step" transition probability matrix $P^{(n)}$ containing $p_{ij}^{(n)} = p(X_{t+n} = j|X_t = i)$ is the nth power of P and

$$p^{(n)^T} = p^{(0)^T} P^{(n)}, \text{ where } P^{(n)} = P^n. \tag{2.17}$$

A special distribution is the so-called *invariant distribution* $\tilde{p}$.

Definition 2.3.2. *The distribution $\tilde{p}$ is called an invariant distribution of a homogeneous Markov chain $\mathcal{X}$ with transition probability matrix P, if $\tilde{p}^T = \tilde{p}^T P$.*

Consequently, if $p^{(0)} = \tilde{p}$, then $p^{(t)} = \tilde{p}$ for $\forall\ t$. The invariant distribution $\tilde{p}$ is frequently called a *stationary distribution*, because for a first-order Markov chain $\mathcal{X}$ the identical marginals imply that $p^{(t)} = \tilde{p}$ ($\mathcal{X}$ is strongly *stationary*, if the distributions of time-shifted finite marginals are identical).

This indicates that if we could construct P such that the target distribution $\pi(X)$ is a corresponding stationary distribution and we could sample from $\pi(X)$

at least a correct prior distribution to start the chain, then we could sample from the target distribution by an efficient simulation of the chain. In lack of this, we try to construct such a P that $p^{(t)}$ converges to $\pi(X)$. To formalize this idea, we need the following concept [102, 120].

Definition 2.3.3. *A Markov chain $\mathcal{X}$ is stable, if $\lim_{t\to\infty} p(X_t) = p^{(\infty)}$ exists, it is a distribution (called limiting distribution or equilibrium distribution), and independent of the initial distribution $p(X_0)$.*

Now we need the concept of irreducibility and aperiodicity to state a central result about the limiting and invariant distributions.

Definition 2.3.4. *The discrete and finite state space Markov chain $\mathcal{X}$ is called*

1. *Irreducible, if there exists $n_{ij} > 0$ for all i, j such that $p_{ij}^{(n_{ij})} > 0$,*
2. *Aperiodic, if for some i (and with irreducibility, for all), there exists $n_i > 0$ that for all $n \geq n_i$ $p_{ii}^{(n)} > 0$.*

Theorem 2.3.1 ([102]). *If a discrete and finite state space Markov chain $\mathcal{X}$ is irreducible and aperiodic, then the chain is stable and the limiting distribution is the unique invariant distribution (i.e., $p^{(\infty)}$ is a unique, nonnegative solution of ${p^{(\infty)}}^T = {p^{(\infty)}}^T P$ and $\sum_i p_i^{(\infty)} = 1$).*

To simplify notation, for a stable chain we denote this unique limiting and invariant distribution $(p^{(\infty)}, \tilde{p})$ with $\pi(X)$, because in our case it will be the target distribution. Frequently in the literature, a stable chain $\mathcal{X}$ is called ergodic.

The convergence to the stationary distribution $\pi(X)$ allows various ergodic theorems, for example an analog of the law of large numbers [102, 120].

Theorem 2.3.2. *If a discrete and finite state space Markov chain $\mathcal{X}$ is stable and $\bar{f} = \mathrm{E}_{\pi(X)}[f(X)] < \infty$, then $P(\lim_{N\to\infty} \hat{f}_N = \bar{f}) = 1$, where $\hat{f}_N = 1/N \sum_{t=1}^{N} f(X_t)$.*

To state an analog "central limit theorem", we need the following concept [102].

Definition 2.3.5. *The discrete and finite state space Markov chain $\mathcal{X}$ is geometrically ergodic (convergent), if there exists $0 \leq \lambda < 1$ and function $V(.) > 1$ such that*

$$\sum_j |p_{ij}^{(t)} - \pi_j| \leq V(i)\lambda^t \ \forall\, i. \tag{2.18}$$

The smallest such λ is called a rate of convergence, expressing the convergence speed to the limiting distribution (i.e., geometric convergence implies stability).

An analog "central limit theorem" for Markov chains is as follows [235, 102].

Theorem 2.3.3 ([102]). *If a discrete and finite state space Markov chain $\mathcal{X}$ is geometrically ergodic (and thus stable) and started with its invariant distribution $\pi(X)$, then for a real valued function f with $\bar{f} = \mathrm{E}_\pi[f(X)]$ and $\sigma^2 = \mathrm{var}_\pi(f(X))$, $\mathrm{E}_\pi[f(X)^{2+\epsilon}] \leq \infty$ with some $\epsilon > 0$*

$$\tau^2 = \sigma^2 + 2\sum_{k=1}^{\infty} \mathrm{E}_\pi[(f(X_0) - \bar{f})(f(X_k) - \bar{f})] \tag{2.19}$$

exists and nonnegative, and for $\hat{f}_N = 1/N \sum_{t=1}^{N} f(X_t)$

$$\sqrt{N}\frac{\hat{f}_N - \bar{f}}{\tau} \to N(0,1) \text{ in distribution as } N \to \infty. \tag{2.20}$$

This theorem provides the theoretical basis for the construction of asymptotic confidence intervals for the estimates $\hat{f}_N$ based on the dependent samples from a Markov chain Monte Carlo simulation by estimating τ, the so-called Monte Carlo variance.

Finally we define a property that provides an efficient method to check the invariance of a distribution and to construct Markov chains.

Definition 2.3.6. *The discrete and finite state space Markov chain $\mathcal{X}$ with transition probability matrix P and invariant distribution $\tilde{p}$ is called reversible, if it satisfies the detailed balance condition*

$$\forall\, i, j\ \tilde{p}_i P_{ij} = \tilde{p}_j P_{ji}. \tag{2.21}$$

By summation it gives $\tilde{p}^T P_{\cdot j} = \tilde{p}_j$, which is the defining equation of an invariant distribution. Consequently, if for a given P $\underline{q}$ satisfies detailed balance, then $\underline{q}$ is an invariant distribution and vice versa, if for a given target distribution $\underline{\pi}$ we can construct a P such that it satisfies detailed balance with $\underline{\pi}$, then $\underline{\pi}$ is its invariant distribution. Furthermore, if the constructed P is such that the corresponding reversible Markov chain is irreducible and aperiodic as well, then $\underline{\pi}$ is its unique, invariant, limiting distribution, so we can generate (dependent) samples by sequential simulation and use it to approximate expectations and to provide confidence measures.

2.3.1.2 MCMC with the Metropolis-Hastings scheme

The Metropolis-Hastings algorithm provides a scheme to generate samples from a given unnormalized distribution by implicitly defining and simulating a reversible Markov chain. Besides the target distribution, the scheme can incorporate proposal distributions in the defined transition probabilities, offering the possibility of specialization for a given domain, though the irreducibility and aperiodicity of the chain has to be guaranteed.

Let $\pi(X)$ denote the unnormalized, strictly positive target distribution over $S = \{0, 1, \ldots, K\}$ ($\pi_i = \pi(X = i) \geq 0$). Let Q be a transition probability matrix ($Q\underline{1} = \underline{1}$), the so-called proposal distribution (for transitions), such that

($q_{ij} \geq 0$) *iff* ($q_{ji} > 0$). Define a Markov chain $\mathcal{X}$ with probability transition matrix P such that

$$p_{ij} = q_{ij} \min\left(1, \frac{\pi_j q_{ji}}{\pi_i q_{ij}}\right); \forall i \neq j \tag{2.22}$$

and define $p_{ii} = 1 - \sum_{j \neq i} p_{ij}$. Note that for the construction only the ratios of the target distribution are needed, which fits to the practical case of unnormalized posterior in Bayesian analysis.

Now $\pi(X)$ is the stationary distribution of the defined Markov chain, which can be proved by showing that the detailed balance condition is satisfied.

Furthermore, clearly, if Q is irreducible, so will be P and the same is true for aperiodicity. Consequently, if we provide a proposal distribution Q that (its corresponding Markov chain) is irreducible and aperiodic, then for a given target distribution $\pi(X)$ the construction above defines a stable and reversible Markov chain with (invariant) limiting distribution $\pi(X)$.

2.3.1.3 Convergence and confidence issues

The Metropolis-Hastings scheme offers complete freedom to design the proposal distribution specific to the domain, because it is ensured that the distribution and the averages will converge asymptotically. However for a Metropolis-Hastings algorithm with a specified proposal Q and target $\pi(X)$ distributions there are no analytic results with general, practical applicability for the rate of convergence to the target distribution, for forgetting the starting values or for the Monte Carlo variance of the average Eq. 2.19. Consequently, the length of the necessary simulation is usually determined by observing and analyzing simulations, practically based on the actual sampling.

These two problems of the convergence to the limiting distribution and the convergence of the ergodic averages shows the dual usage of MCMC methods: generation of samples from the target distribution for its exploration and computing ergodic averages for approximating expectations. Because our primary goal is the reliable approximation of expectations of the target quantities and not per se the convergence of the induced distribution of the target quantity, an optimal method would provide an estimate with a confidence interval without answering the question of convergence to the limiting distribution (i.e., we need only the convergence of the average of the target quantity).

The visual analysis of the sequence of a scalar target quantity $Y = f(X)$ (called trace plot) is usually based on the inspection of some form of stability of the estimated mean, variance, smoothness of the curve, either in a single simulation $\{Y_i; i = 1, \ldots, N\}$, or more frequently in multiple simulations M with wide range of starting values $\{Y_{i,j}; i = 1, \ldots, N\, j = 1 \ldots, M\}$. However, with many target quantities or with complex models more formal approaches are required.

The methods used in the thesis can be grouped as determining the length of *burn-in* (or *convergence diagnostic* tools), when the limiting distribution is sufficiently approached for a reliable estimation of the target quantity, and as

determining *stopping time*, when the Monte Carlo variance is sufficiently small (see Section 8.5.3). Each of the methods is based on the sampled sequence of the target quantity from a stable Markov chain with a Metropolis-Hastings algorithm. For convergence diagnostics the single-chain test of Geweke and the multiple-chain R score of Gelman-Rubin were used [102, 108, 45, 213]. The Monte Carlo variances were estimated using partitionings of single-chains [102, 45].

2.3.2 The hybrid Markov Chain Monte Carlo method

The problem of designing an efficient proposal distribution, which ensures large movements in the state space while maintaining high acceptance rate can be approached by the use of mixture of proposals (possibly compiling multiple proposals) [102]. An example is the hybrid Markov Chain Monte Carlo method, which is applicable if the gradient is efficiently computable for $\log(\pi(x)),\ x \in \mathcal{R}^k$. It utilizes the gradient information to replace the random steps in random walk Metropolis with large deterministic movements and embeds the parameter space in a larger space to ensure high acceptance rate and efficient full exploration (see Section 10.2.3.2 and Section 10.6 for its applications).

2.4 Model evaluation and selection

In practice, model evaluation is of central importance first to enhance the predictive performance theoretically (i.e., by extending and refining the model), second to enhance the predictive performance computationally (i.e., by ensuring more efficient simulation) and third to support scientific understanding of the model. Model evaluation can be particularly important, for example if the prior elicitation, transformation and incorporation is a complex process with multiple choices and the goal is the evaluation of the effect of prior (i.e., the sensitivity analysis). As the prior is part of the joint model, model evaluation and selection naturally includes the evaluation (and selection) of priors, so standard techniques for model evaluations can be used for the prior evaluation as well (see Section 8.2 for its application).

2.4.1 The prequential framework

In the predictive sequential (prequential) framework, the quantification of the performance of a forecasting system (i.e. model evaluation) is based purely on the predictive sequential (online) performance of the forecasting system. It consists of a forecasting system observing a sequence $D = X_1, X_2, \ldots$ of uncertain quantities in turn, which provides a forecast F_{n+1} for the next quantity given the previous observations $D_n = \{x_1, \ldots, x_n\}$ at each step. The forecasts are evaluated by a score function $S(F_{n+1}, x_{n+1})$ and the total score S is defined by the cumulative sum. We discuss the application of this framework for Bayesian networks in Section 3.4 and report results in Section 8.2.

This framework is applicable for various forecasting systems (for general treatments see [68, 70]). If the forecasting system is probabilistic, then the joint or the conditionals $p_n(X_n|X_1,\ldots,X_{n-1})$ can be defined. In a Bayesian forecasting system this is achieved by defining a prior and sampling distribution and selecting the conditionals to define the appropriate posterior predictive distributions.

The following example introduces a Bayesian forecasting system, when the uncertain quantities have r discrete values denoted with integers $1,\ldots,r$.

Example 2.4.1. *Assume that the observed sequence $D_n = \{X_i; i = 1, 2, \ldots, n\}$ contains i.i.d. multinomial samples with r discrete values. The prior $p(\theta)$ is a Dirichlet prior with hyperparameters $\underline{\alpha} = (\alpha_1,\ldots,\alpha_r)$ and $\alpha_+ = \sum_k \alpha_k$:*

$$\mathrm{Dir}(\theta|\underline{\alpha}) = c\prod_k \theta_k^{\alpha_k-1}, \text{ where } c = \frac{\Gamma(\alpha_+)}{\prod_k \Gamma(\alpha_k)}. \tag{2.23}$$

This prior is conjugate for multinomial sampling, so the posterior predictive distributions of the defined Bayesian forecasting system are the updated Dirichlet with hyperparameters $\underline{\alpha}_i$ at step i and the posterior prediction for value x_i (i.e., the marginal posterior probability $\mathrm{E}[\theta_{x_i}]$) is

$$\begin{aligned}
p(x_i|x_1,\ldots,x_{i-1}) &= \int p(x_i|\underline{\theta})\mathrm{Dir}(\underline{\theta}|\underline{\alpha_i})\,\mathrm{d}\underline{\theta} \\
&= c\int \theta_{x_i}\prod_k \theta_k^{\alpha_{i,k}-1}\,\mathrm{d}\underline{\theta}, \text{ where } c = \frac{\Gamma(\alpha_{i,+})}{\prod_k \Gamma(\alpha_{i,k})} \\
&= c\int\prod_k \theta^{\alpha_{i+1,k}-1}\mathrm{d}\theta, \text{where } \alpha_{i+1,k}{=}\alpha_{i,k}, \text{except } \alpha_{i+1,x_i}{=}\alpha_{i,x_i}{+}1 \\
&= \frac{\Gamma(\alpha_{i,+})}{\Gamma(\alpha_{i+1,+})}\frac{\prod_k \Gamma(\alpha_{i+1,k})}{\prod_k \Gamma(\alpha_{i,k})} \\
&= \frac{\alpha_{i,x_i}}{\alpha_{i,+}},
\end{aligned} \tag{2.24}$$

so the marginal probability of the data set D_n with prior $\underline{\theta} \sim \mathrm{Dir}(\underline{\alpha}_1)$ and n_k occurrences of values $k = 1,\ldots,r$ is

$$p(x_1,\ldots,x_n) = \prod_{i=1}^{n} p(x_i|x_1,\ldots,x_{i-1}) \tag{2.25}$$

$$= \frac{\prod_{k=1}^{r}(\alpha_{1,k}\ldots(\alpha_{1,k}+n_k))}{\alpha_{1,+}\ldots(\alpha_{1,+}+n)} \tag{2.26}$$

$$= \frac{\Gamma(\alpha_{1,+})}{\Gamma(\alpha_{1,+}+n)}\frac{\prod_{k=1}^{r}\Gamma(\alpha_{1,k}+n_k)}{\prod_{k=1}^{r}\Gamma(\alpha_{1,k})}. \tag{2.27}$$

Now we turn to the question of score function assuming that the uncertain quantities have r discrete values $s_1,\ldots,s_r$ and the forecasts $\underline{q_n}$ of the

probabilistic system are based on the posterior predictive distributions $\underline{q_n} = p_n(X_n|x_1, \ldots, x_{n-1})$.

If we interpret the forecasts in a decision theoretic framework as reporting of the posteriors, then the score function is a loss function $S(\underline{q}, s_k)$ and $\underline{q_n}$ should correspond to the minimal loss forecast (see Eq. 2.9)

$$\arg\min_{\underline{q}} \sum_{k=1}^{r} S(\underline{q}, s_k) p_n(X_n = s_k|x_1, \ldots, x_{n-1}). \tag{2.28}$$

It can be shown that the requirements of honesty ("reporting true beliefs"), smoothness ("proportional penalty for errors") and decomposability (penalty depends on pairs of {forecasts-outcomes}) characterize a logarithmic score function, $S(\underline{q}, s_k) = A \log(q_k) + B_k$ where $A < 0$ and B_k are arbitrary constants [34].

Note that the expected loss of reporting $\underline{q} \neq \underline{p}$ under the logarithmic score function corresponds to the cross-entropy $H(p\|q) = \mathrm{KL}(p\|q) + H(p)$ (see Eq. 2.12 and Eq. 2.11).

Returning to the scoring of a probabilistic forecasting system, the adoption of a logarithmic score

$$S_n(p_n(X_n|x_1, \ldots, x_{n-1}), x_n) = -\log(p_n(X_n = x_n|x_1, \ldots, x_{n-1})) \tag{2.29}$$

has other useful consequences w.r.t. batch evaluation.

First, the score over a given data set D_n is the logarithm of the marginal data likelihood (see Eq. 2.8) and independent of the ordering (i.e., the score equivalently can serve as a batch score for analyzing the joint data set and the model):

$$\begin{aligned} S &= \sum_{i=1}^{n} S_i(p_i(X_i|x_1, \ldots, x_{i-1}), x_i) && (2.30) \\ &= -\log \prod_{i=1}^{n} p_i(x_i|x_1, \ldots, x_{i-1}) && (2.31) \\ &= -\log p(x_1, \ldots, x_n). && (2.32) \end{aligned}$$

Second, in the relative approach to model evaluation, the score of the forecasting system (M) is compared to the score of a reference system M^{ref}. The relative logarithmic score

$$\exp(S - S^{ref}) = \frac{p(x_1, \ldots, x_n|M)}{p(x_1, \ldots, x_n|M^{\mathrm{ref}})} \tag{2.33}$$

is the Bayes factor (see Eq. 2.13).

2.4.2 Maximum a posteriori analysis

Finding model classes M_k with large posterior probability and modes of the posterior $p(\underline{\theta}_k|D, M_k)$ is of essential importance in practice. Assume that our

goal is to find a *maximum a posteriori* (MAP) or *maximum likelihood* (ML) parameterization $\underline{\theta} \in \mathcal{R}^k$

$$\underline{\theta}^{\mathrm{ML}} = \arg\max_{\underline{\theta}} L(\underline{\theta}; D), \text{ where } \mathrm{L}(\underline{\theta}; \mathrm{D}) = \log(\mathrm{p}(\mathrm{D}|\underline{\theta})), \tag{2.34}$$

$$\underline{\theta}^{\mathrm{MAP}} = \arg\max_{\underline{\theta}} \log(p(\underline{\theta})|D), \tag{2.35}$$

when $L(\underline{\theta}; D)$ is efficiently computable, furthermore its gradient vector $L'(\underline{\theta}; D)$ and the Hessian matrix $H = L''(\underline{\theta}, D)$ are available as well. In lack of analytic solutions, a standard choice is to use deterministic optimization techniques such as *gradient descent*, which starts from $\underline{\theta}_0$ selected at random or based on prior knowledge and iteratively updates it as

$$\underline{\theta}_{t+1} = \underline{\theta}_t - \epsilon L'(\underline{\theta}_t; D). \tag{2.36}$$

An attempt to solve the question of optimal step size ϵ is the extension of the Eq. 2.36 with a so-called momentum term, which is the geometric average of the earlier updates with parameter μ. A more automated method is the *line search*, which performs an optimization along a given direction $\underline{d}_t$

$$\underline{\theta}_{t+1} = \underline{\theta}_t - \lambda \underline{d}_t, \text{ where } \lambda = \arg\min_{\lambda} L(\underline{\theta} + \lambda \underline{d}_t), \tag{2.37}$$

and selects appropriate (non-interfering) directions $\underline{d}_t$ such that consecutive steps will not deteriorate the results of previous optimization steps. This is achieved (up to a second-order approximation of $L(\underline{\theta}; D)$) by selecting conjugate directions

$$\underline{d}_{t+1} = -L'(\underline{\theta}_{t+1}; D) + \beta_t \underline{d}_t, \text{ that } \underline{\mathrm{d}}_{\mathrm{t+1}}^{\mathrm{T}} \underline{\underline{\mathrm{H}}} \underline{\mathrm{d}}_{\mathrm{t}} = 0. \tag{2.38}$$

For derivation and formulas for β_t including only gradients and not the Hessian (in the so-called conjugate gradient algorithms) see [36]. Furthermore, the line search can be replaced by using a second-order approximation based on the approximation of the Hessian as suggested in the so-called *scaled conjugate algorithm* [190]. These deterministic optimization algorithms provide only a local optimum, but they can be incorporated in a stochastic framework called simulated annealing that is theoretically capable for global optimization (it can be interpreted as random walk Metropolis algorithm with gradually decreased acceptance rate). The conjugate gradient algorithm and the scaled conjugate gradient algorithm is applied in the thesis for finding maximum a posteriori parameters of classifiers (see Section 10.2.3.3 for its application).

Chapter 3

Bayesian networks primer

We summarize the Bayesian network model class, its probabilistic and causal interpretations and its Bayesian application. Then we overview the main issues of knowledge engineering, model evaluation and finally the learning of Bayesian networks.

The Bayesian framework overviewed in Chapter 2 leaves open the question of the model class, it is equally applicable with domain models discussed in this chapter or with conditional models discussed in Chapter 9. In this chapter we investigate a domain model class called *Bayesian network*s, conditional models are discussed in Chapter 9. Bayesian networks form a subclass of graphical models that is using directed acyclic graphs (DAGs) instead of more general graphs to represent a probability distribution and optionally the causal structure of the domain. In an intuitive causal interpretation, the nodes represent the uncertain quantities, the edges denote direct causal influences, defining the model structure. A local probabilistic model is attached to each node to quantify the stochastic effect of its parents (causes). The descriptors of the local models give the model parameters.

The widespread popularity of this representation is probably the consequence of its applicability in multiple disciplines. The multifaceted nature of Bayesian networks follows from the fact that this representation addresses jointly three autonomous levels of the domain: the causal model, the probabilistic dependency-independency structure, and the distribution over the uncertain quantities. Additionally, the Bayesian network, as a complete probabilistic domain model, can be applied as an input-output model, for example as a classifier, so it can be investigated in the conditional framework as well (see Chapter 9 and 10).

First we summarize the probabilistic interpretation of Bayesian networks, which is based on a DAG representation of an independence model of a distribution and on a decomposed representation of a distribution by DAGs annotated with local probabilistic models. Then we introduce the causal interpretation of Bayesian networks. Next we discuss the Bayesian approach to the parameters and to the structure. Then we discuss the knowledge acquisition methods and model (prior) evaluation methodologies. Finally we discuss fundamental results

for model identification.

3.1 Representational issues

3.1.1 Three aspects: belief, relevance and causation

Suppose that our goal is to model uncertain events, furthermore we assume that the number of events and the corresponding outcomes (observables) are finite. According to the discussion in Chapter 2, it corresponds to modeling a subjective joint distribution over the event space with elementary events defined by the Cartesian product of the possible outcomes. We denote the joint set of random events with $\underline{V}$, $p(\underline{V})$ denotes the joint (mass) probability distribution representing the personal belief over events. If it is necessary to differentiate, capitals with underline such as $\underline{X}$, $\underline{Y}$, $\underline{Z}$ denotes subsets and capitals such as X, Y, Z single random events, lowercase letters denotes values (outcomes) such as $X = x$. To simplify terminology we call each discrete random event a random variable (i.e., as if their outcomes would be always in $\mathcal{R}$).

3.1.1.1 The model of observational independencies

We introduce now the notation for the independencies of random events.

Definition 3.1.1. *Let $p(\underline{V})$ be a joint distribution over $\underline{V}$ and $\underline{X}, \underline{Y}, \underline{Z} \subseteq \underline{V}$ are disjoint subsets. Then denote the conditional independence of $\underline{X}$ and $\underline{Y}$ given $\underline{Z}$ with $I_p(\underline{X}|\underline{Z}|\underline{Y})$, that is*

$$I_p(\underline{X}|\underline{Z}|\underline{Y}) \text{ iff } (\forall \underline{x}, \underline{y}, \underline{z}\ p(\underline{y}|\underline{z}, \underline{x}) = p(\underline{y}|\underline{z}) \text{ whenever } p(\underline{z}, \underline{x}) > 0). \tag{3.1}$$

Note that conditional independence is required for all the relevant values of $\underline{Z}$. A weakened form of independence is the contextual independence, if conditional independence is valid only for a certain value $\underline{c}$ of another disjoint set $\underline{C}$. Then denote the contextual independence of $\underline{X}$ and $\underline{Y}$ given $\underline{Z}$ and context c with $I_p(\underline{X}|\underline{Z}, \underline{c}|\underline{Y})$, that is

$$I_p(\underline{X}|\underline{Z}, \underline{c}|\underline{Y}) \text{ iff } (\forall \underline{x}, \underline{y}, \underline{z}\ p(\underline{y}|\underline{z}, \underline{c}, \underline{x}) = p(\underline{y}|\underline{z}, \underline{c}) \text{ whenever } p(\underline{z}, \underline{c}, \underline{x}) > 0). \tag{3.2}$$

Another notation for $I_p(X|Z|Y)$ is $(X \perp\!\!\!\perp Y|Z)_p$. If it is nonambiguous, the subscript from $I_p(.)$ is omitted as well as the empty condition part. The negated independence proposition (i.e., dependency) is denoted with $(X \not\perp\!\!\!\perp Y|Z)_p$. It is a *direct dependency*, if for any disjoint $X, Y, Z \subseteq V$ $(X \not\perp\!\!\!\perp Y|Z)$ holds. A set of independence statements is called *independence model* (note that this is always a finite set in our case). We use the terms (probabilistic) independence and (information) irrelevance interchangeably.

Whereas the independencies or the complete independence model is an ideal candidate to represent qualitatively the target distribution, the autonomous, local mechanisms (rules) composing modularly the domain are the basis of both common sense and scientific understanding and explanation. The autonomous relations are asymmetric w.r.t. time and interventions suggesting a causal interpretation.

3.1.1.2 The model of causal (in)dependencies

For the discussion of causality, we need a concept and notation for intervention.

Definition 3.1.2. *Let $do(x)$ denote the intervention of setting variable(s) X to value x and $p(Y|do(x))$ the corresponding interventional distribution [201].*

Note that despite the symmetry of the probabilistic dependence relation, the causal dependence relation is asymmetric. For example in a hypothetical world with two variables X, Y and a single causal relation $X \rightarrow Y$ inducing $p(X, Y)$, the intervention on X and the observation of X are identical operations, but the intervention on Y will not influence the cause X (i.e., $p(Y|do(x)) = p(Y|x)$, but $p(X|do(y))$ is equal to $p(X)$ and not to $p(X|y)$). Now we introduce a notation for the *causal irrelevance* (independency) [202, 101].

Definition 3.1.3. *Let $p(.|do(.))$ denote the appropriate interventional distributions over $\underline{V}$ and $\underline{X}, \underline{Y}, \underline{Z} \subseteq \underline{V}$ are disjoint subsets. Then denote the causal independence of $\underline{X}$ and $\underline{Y}$ given $\underline{Z}$ with $CI_p(\underline{X}; \underline{Y}|\underline{Z})$, that is*

$$CI_p(\underline{X}; \underline{Y}|\underline{Z}) \text{ iff } (\forall \underline{x}, \underline{y}, \underline{z}\, p(\underline{y}|do(\underline{z}), do(\underline{x})) = p(\underline{y}|do(\underline{z}))) \tag{3.3}$$

A set of causal (in)dependence statements is called *causal model*.

3.1.2 Probabilistic Bayesian networks

Before investigating the role of directed acyclic graphs (DAGs) in representing causal relations, we have to clarify their purely probabilistic role in representing a joint distribution numerically and its (in)dependence model.

3.1.2.1 Markov conditions

Assume that each vertice (node) in DAG G corresponds to a random variable. We need the following concepts (cited from [200, 169, 60, 202]).

Definition 3.1.4. *A distribution $p(X_1, \ldots, X_n)$ is Markov relative to DAG G or factorizes w.r.t G, if*

$$p(X_1, \ldots, X_n) = \prod_{i=1}^{n} p(X_i|\,\mathrm{Pa}(X_i)), \tag{3.4}$$

where $\mathrm{Pa}(X_i)$ denotes the parents of X_i in G.

Definition 3.1.5. *A distribution $p(X_1, \ldots, X_n)$ obeys the ordered Markov condition w.r.t. DAG G, if*

$$\forall\, i = 1, \ldots, n : (X_{\prec(i)} \perp\!\!\!\perp \{\{X_{\prec(1)}, \ldots\, X_{\prec(i-1)}\} \setminus \mathrm{Pa}(X_{\prec(i)})\}|\,\mathrm{Pa}(X_{\prec(i)}))_p, \tag{3.5}$$

where $\prec$ is some ancestral ordering w.r.t. G (i.e., compatible with arrows in G) and $\{X_{\prec(1)}, \ldots\, X_{\prec(i-1)}\} \setminus \mathrm{Pa}(X_{\prec(i)})$ denotes all the predecessors of $X_{\prec(i)}$ except its parents.

Definition 3.1.6. *A distribution $p(X_1, \ldots, X_n)$ obeys the local (or parental) Markov condition w.r.t. DAG G, if*

$$\forall\, i = 1, \ldots, n : (X_i \perp\!\!\!\perp \text{Nondescendants}(X_i) | \, \text{Pa}(X_i))_p, \tag{3.6}$$

where Nondescendants(X_i) *denotes the nondescendants of X_i in G (i.e., without directed path from X_i).*

Definition 3.1.7. *A distribution $p(X_1, \ldots, X_n)$ obeys the global Markov condition w.r.t. DAG G, if*

$$\forall\, X, Y, Z \subseteq V : \; (X \perp\!\!\!\perp Y | Z)_G \Rightarrow (X \perp\!\!\!\perp Y | Z)_p, \tag{3.7}$$

where $(X \perp\!\!\!\perp Y|Z)_G$ denotes that X and Y are d-separated by Z, that is if every path p between a node in X and a node in Y is blocked by Z as follows

1. *Either path p contains a node n in Z with non-converging arrows (i.e., $\rightarrow n \rightarrow$ or $\leftarrow n \rightarrow$),*
2. *Or path p contains a node n not in Z with converging arrows (i.e., $\rightarrow n \leftarrow$) and none of the descendants of n is in Z.*

Now we can state a central result connecting the DAG representation of the joint distribution and the DAG representation of the independence model [169].

Theorem 3.1.1 ([169]). *Let $p(V)$ be a probability distribution and G a DAG, then the conditions in Def. 3.1.4, 3.1.5, 3.1.6, and 3.1.7 are equivalent:*

(F) p is Markov relative G or p factorizes w.r.t G,

(O) p obeys the ordered Markov condition w.r.t. G,

(L) p obeys the local Markov condition w.r.t. G,

(G) p obeys the global Markov condition w.r.t. G.

Because of their equivalence, we can refer to these as the (directed) Markov condition for the pair (p, G). To show the necessity and sufficiency of these conditions, we refer to a result that a sound and complete, computationally efficient algorithm exists to read off exactly (!) the independencies that are valid in all distributions that are Markov relative to a given DAG G [200].

Theorem 3.1.2 ([200]).

$$\forall\, X, Y, Z \subseteq V : \; (X \perp\!\!\!\perp Y | Z)_G \Leftrightarrow ((X \perp\!\!\!\perp Y | Z)_p \text{ in all } p \text{ Markov relative to } G.$$

Two further properties are implied by any of the (FOLG) conditions: the pairwise Markov condition [169] and the boundary Markov conditions [200].

Definition 3.1.8. *A distribution $p(X_1, \ldots, X_n)$ obeys the pairwise Markov condition w.r.t. DAG G, if for any pair of variables X_i, X_j nonadjacent in G and $X_j \in$ Nondescendants(X_i), $(X_i \perp\!\!\!\perp X_j | \text{Nondescendants}(X_i) \setminus \{X_j\})_p$ holds [169].*

To state the other implication, we need the following concepts.

Definition 3.1.9. *A set of variables $MB_p(X_i)$ is called a Markov blanket of X_i w.r.t. the distribution $p(X_1, \ldots, X_n)$, if $(X_i \perp\!\!\!\perp V \setminus \mathrm{MB}(X_i) | \mathrm{MB}(X_i))_p$ (see Fig. 3.1). A minimal Markov blanket is called Markov boundary [200].*

Definition 3.1.10. *A distribution $p(X_1, \ldots, X_n)$ obeys the boundary Markov condition w.r.t. DAG G, if the boundary $\mathrm{bd}(X_i, G)$ is a Markov blanket of X_i, where $\mathrm{bd}(X_i, G)$ denotes the set of parents, children and the children's other parents for X_i (i.e., parents with common child with X_i, see Fig. 1.1 and Fig. 3.1):*

$$\mathrm{bd}(X_i, G) = \{\mathrm{Pa}(X_i, G) \cup \mathrm{Ch}(X_i, G) \cup \mathrm{Pa}(\mathrm{Ch}(X_i, G), G)\}. \tag{3.8}$$

The boundary $\mathrm{bd}(X_i, G)$ coincides with the standard graph-theoretic notion of boundary (i.e., set of neighbours) of X_i in the moral graph of G, which is the graph where edges are added between parents with a common child and the orientation is dropped [60].

Theorem 3.1.3 ([200]). *The (FOLG) Markov condition for (p, G) implies that the set $\mathrm{bd}(X_i, G)$ is a Markov blanket ($\mathrm{MB}_p(X_i)$) for X_i.*

Note that the set $\mathrm{bd}(X_i, G)$ is not necessarily Markov boundary as it may not be minimal (because of the non-optimality of G to p). In the Bayesian context this problem is negligible as Th. 7.1.2 and the discussion in Section 3.1.2.3 show, so we will also refer to $\mathrm{bd}(X_i, G)$ as the Markov blanket for X_i in G using the notation $\mathrm{MB}(X_i, G)$ by the implicit assumption that p is Markov compatible with G and stable. The induced (symmetric) pairwise relation $\mathrm{MBM}(X_i, X_j, G)$ w.r.t. G between X_i and X_j

$$\mathrm{MBM}(X_i, X_j, G) \Leftrightarrow X_j \in \mathrm{bd}(X_i, G) \tag{3.9}$$

is called *Markov blanket membership* [96]. In short, the set $\{\mathrm{MBM}(X_i, G)\}$ includes the variables with non-blockable pairwise (observational) dependencies 3.1 to X_i including the unconditionally related variables (parents and children) and the purely conditionally related ones (the rest).

Finally, we introduce here the definition of the Markov Blanket (sub)Graph (MBG) (for a discussion of the MBG feature, see Section 7.2).

Definition 3.1.11. *A subgraph of G is called the Markov Blanket (sub)Graph or Mechanism Boundary (sub)Graph $\mathrm{MBG}(X_i, G)$ of variable X_i if it includes the nodes in the Markov blanket defined by $\mathrm{bd}(X_i, G)$ and the incoming edges into X_i and into its children $\mathrm{Ch}(X_i, G)$ (see Fig. 1.1 and Fig. 3.1).*

Fig. 3.1 shows an example for a Markov Blanket set and the Markov Blanket graph in a Markov chain.

3.1.2.2 Definitions of Bayesian networks

The equivalence of the conditions *FOLG* in Th. 3.1.1 allows versatile definitions of Bayesian networks. A neutral definition is as follows.



Figure 3.1: A Bayesian network structure G defining a Markov chain $p(X_1, X_2, Y, X_4, X_5)$. Underscore denotes the members of a Markov Blanket set of variable Y $\mathrm{MB}_p(Y)$, which is the unique Markov Boundary $\mathrm{MB}(Y, G)$ as well (defined by the boundary $\mathrm{bd}(X_i, G)$). Solid lines denote the edges of the Markov Blanket Graph $\mathrm{MBG}(Y, G)$.

Definition 3.1.12. *A directed acyclic graph (DAG) G is a Bayesian network of distribution $p(V)$, if the variables are represented with nodes in G and (G, p) satisfies any of the conditions F, O, L, G such that G is minimal (i.e., no edge(s) can be omitted without violating a condition F, O, L, G).*

If the distribution P is strictly positive, then the Bayesian network compatible with a given ordering $\prec$ is unique (i.e., composed of the unique minimal parental sets that makes the variable independent of the variables before w.r.t $\prec$) [200]. Note that depending on the ordering different Bayesian networks can be gained, representing more or fewer independencies of P.

In engineering practice Bayesian networks are frequently informally defined as a DAG annotated with local probabilistic models for each node.

Definition 3.1.13. *A Bayesian network model M of a domain with variables V consists of a structure G and parameters $\underline{\theta}$. The structure G is a directed acyclic graph (DAG) such that each node represents a variable and local probabilistic models $p(X_i|\,\mathrm{pa}(X_i))$ are attached to each node w.r.t. the structure G, that is they describe the stochastic dependency of variable X_i on its parents $\mathrm{pa}(X_i)$. As the conditionals are frequently from a certain parametric family, the conditional for X_i is parameterized by $\underline{\theta}_i$, and $\underline{\theta}$ denotes all the parameters of the model.*

When the conditionals are combined together as in Eq. 3.4, they define an overall joint distribution p. It trivially satisfies Markov relativity to G and the structure satisfies the conditions O, L, G. The lack of minimality requirement causes only potential redundancy (parameters) and fewer implied independencies. In most cases, we use the term Bayesian network to refer to both structure and parameters.

3.1.2.3 Stability

A limitation of DAGs to represent a given (in)dependency model is that (1) probabilistic dependencies are not necessarily transitive and (2) lower order (e.g., pairwise) probabilistic independencies does not imply higher order (e.g., multivariate) independencies. These are illustrated with the following examples.

Example 3.1.1. *Consider $p(X, Y, Z)$ with binary X, Z and ternary Y in a Markov chain ($X \to Y \to Z$). The intransitivity condition — $(X \not\perp\!\!\!\perp Y)$, $(Y \not\perp\!\!\!\perp Z)$, and $(X \perp\!\!\!\perp Z)$ — can be rewritten as an equation system with the probabilities.*

Its solvability demonstrates that the "naturally" expected transitivity of dependency can be destroyed by properly selected values. For the other case, consider $p(X,Y,Z)$ with binary variables, where $p(x) = p(y) = 0.5$ and $p(Z|X,Y)$ is defined by the logical function $Z = XOR(X,Y)$. In this case $(X \perp\!\!\!\perp Z)$ and $(Y \perp\!\!\!\perp Z)$, but $(\{X,Y\} \not\perp\!\!\!\perp Z)$, which demonstrates that pairwise independence does not imply total independence.

However, such numerically encoded independencies correspond to solutions of systems of equations describing these constraints, which are not stable for numerical perturbations. This leads to the following definition.

Definition 3.1.14. *The distribution p is stable* (or faithfull), if there exists a DAG called perfect map exactly representing its (in)dependencies (i.e., $(X \perp\!\!\!\perp Y|Z)_G \Leftrightarrow (X \perp\!\!\!\perp Y|Z)_p$, $\forall\ X,Y,Z \subseteq V$). The distribution p is stable w.r.t. a DAG G, if G exactly represents its (in)dependencies.*

Whereas in many domains the possibility of an unstable distributions is a real cause for concern, particularly containing deterministic relations, the following result shows that it is reasonable to expect that in a natural, "noisy" domain almost all the distributions are stable in a strict sense, which is also relevant for the applied Bayesian framework. If a "smooth" distribution is defined over the distributions Markov relative to G (such as in Section 3.1.5.1 in the Bayesian framework), it can be shown that the measure of unstable distributions is zero (as being a solution of a system of equations) [186]. It allows to sharpen Th. 3.1.2 that the DAG-based relation $(X \perp\!\!\!\perp Y|Z)_G$ offers a computationally efficient algorithm to read off exactly the independencies that are valid in a distribution Markov relative to G in case of "almost all" such distributions.

3.1.2.4 Equivalence classes of Bayesian networks

The assumption of stability and strict positivity does not exclude the possibility of having multiple perfect maps encoding the same independencies in p.

Example 3.1.2. *Consider a Markov chain $\mathcal{X} = \{X_1, \ldots, X_n\}$ with a stable distribution. Its independence model includes i=1, $\ldots$, n:$(X_i \perp\!\!\!\perp \{X_1, \ldots X_{i-2}\}|X_{i-1})$, and also the implied $(X_i \perp\!\!\!\perp \{X_1, \ldots X_{i-2}, X_{i+2}, \ldots, X_n\}|\{X_{i-1}, X_{i+1}\})$. This independence model can be exactly represented by n equivalent linear Bayesian networks without introducing convergent arrows, including the two special cases of the "forward" and the "backward" network (see Fig. 3.2).*

The induced independence models allow the definition of an equivalence relation between DAGs [200, 248, 186].

Definition 3.1.15. *Two DAGs G_1, G_2 are observationally equivalent, if they imply the same set of independence relations (i.e., $(X \perp\!\!\!\perp Y|Z)_{G_1} \Leftrightarrow (X \perp\!\!\!\perp Y|Z)_{G_2}$).*

*For a different interpretation of this term in probability theory, see [212].

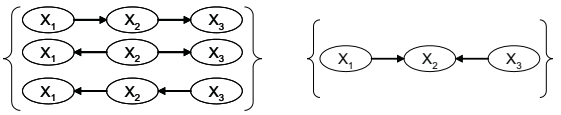


Figure 3.2: The equivalence classes of Bayesian network structures over three variables with direct dependences between X_1, X_2 and X_2, X_3, but not between X_1, X_3.

The implied equivalence classes may contain $n!$ number of DAGs (e.g., all the full networks representing no independencies) or just 1 (e.g., the empty DAG representing total independence of the variables). The characterization of the DAGs within the same equivalence class relies on two observations. First, the undirected skeleton of the observationally equivalent DAGs are the same, because an edge in a DAG denotes a direct dependency, which has to appear in any Markov compatible DAG [200]. Second, the direct dependencies between X, Y and Y, Z without direct dependence between X, Z and without independence such that $(X \perp\!\!\!\perp Z | \{Y, S\})$ has to be expressed with a unique converging orientation $X \rightarrow Y \leftarrow Z$ creating a so-called *v-structure* according to the global semantics. The theorem characterizing the DAGs within the same observational (and distributional) equivalence class is as follows.

Theorem 3.1.4 ([200, 49]). *Two DAGs G_1, G_2 are observationally equivalent, iff they have the same skeleton (i.e., the same edges without directions) and the same set of v-structures (i.e., two converging arrows without an arrow between their tails) [200]. If in the Bayesian networks $(G_1, \underline{\theta}_1)$ and $(G_2, \underline{\theta}_2)$ the variables are discrete and the local conditional probabilistic models are multinomial distributions, then the observational equivalence of G_1, G_2 implies equal dimensionality and bijective relation between the parameterizations $\underline{\theta}_1$ and $\underline{\theta}_2$ called distributional equivalence [49].*

The limitation of DAGs to represent uniquely a given (in)dependency model poses a problem for the interpretation of the direction of the edges. It also poses the question of representing the identically oriented edges in observationally equivalent DAGs. As the definition of the observational equivalence class suggests the common v-structures identify the starting common edges and further identical orientations are the consequences of the constraint that no new v-structures can be created. This leads to the following definition (for an efficient, sound, and complete algorithm, see [186]).

Definition 3.1.16. *The essential graph representing DAGs in a given observational equivalence class is a partially oriented DAG (PDAG) that represents the edges that are identically oriented among all DAGs from the equivalence class (called compelled edges) in such a way that exactly the compelled edges are directed in the common skeleton, the others are undirected representing inconclusiveness.*

3.1.3 Causal Bayesian networks

Now we continue with the causal interpretation of Bayesian networks, because of its relevance for prior acquisition and incorporation (i.e., knowledge acquisition from experts, for the discovery from scientific publications and for prior incorporation in Chapters 6, 8).

3.1.3.1 On the possibility of causal interpretation

The classical problem of "from (observational) correlation to causation", that is the question of determining causal status of a passively observed dependency between X and Y can be decomposed using the concepts introduced earlier to the question about the DAG-based representation of independencies (i.e., probabilistic Bayesian network), the existence of exact representation (i.e., stability) and the existence of unambiguous representation (i.e., essential graph). First, we have to consider whether all direct dependencies among the constructed domain variables are causal. This assumption is highly questionable and is discussed in detail below. Second, we have to consider stability that would guarantee that a corresponding Bayesian network exactly represents the independencies. Third, we have to adopt the "Boolean" Ockham principle, namely that only the minimal, consistent models are relevant (see Section 7.4, for the "soft" Ockham principle in the Bayesian approach to causal discovery). The essential graph resulting from the joint analysis of the observational conditional independencies (i.e., "correlations) indicates causal relations under these conditions. In short, under the condition of stability the essential graph represents the direct causal dependencies and the orientations that are dictated by (in)dependencies in the domain through the minimal models (DAGs) compatible with them. Furthermore, the direction of the edges corresponds to the intuitive expectation as the intransitive dependency triplets are represented as v-structures.

Correspondingly we can define a causal model as a Bayesian network according to Definition 3.1.13 with the causal interpretation that edges denote direct influences.

Definition 3.1.17. *A DAG is called a causal structure over a set of variables V, if each node represents a variable and edges direct influences. A causal model is a causal structure extended with local probabilistic models $p(X_i|\,\mathrm{pa}(X_i))$ for each node w.r.t. the structure G describing the causal stochastic dependency of variable X_i on its parents $\mathrm{pa}(X_i)$. As the conditionals are frequently from a certain parametric family, the conditional for X_i is parameterized by $\underline{\theta}_i$, and $\underline{\theta}$ denotes all the parameters, so a causal model consists of a structure G and parameters $\underline{\theta}$.*

With further assumption of stability, the essential graph shows exactly the independency relations and exhaustively the identifiable causal relations, which suggests that whereas the question of causation is underconstrained for a pair of variables (restricted to "no dependency-no causation), the joint analysis of the system of independencies allows partial identification.

3.1.3.2 The Causal Markov Condition

The following condition ensures the validity and sufficiency of a causal structure.

Definition 3.1.18. *A causal structure G and distribution p satisfies the Causal Markov Condition, if p obeys the local Markov condition w.r.t. G.*

The Causal Markov condition relies on Reichenbach's "common cause principle" that dependency between events X and Y occurs either because X causes Y, or Y causes X or there is a common cause of X and Y (it is possibly an aggregate of multiple events) [202, 116]. Consequently, the precondition of the Causal Markov condition for (p, G) is that the set of variables V is *causally sufficient* for P, that is all the common causes for the pairs $X, Y \in V$ are inside V. Note that hidden variables are allowed fitting to the usually high level of abstraction of the model, only variables that influence two or more variables in V are necessary for causal sufficiency. Interestingly, in the presence of potential hidden common causes (*confounders*), that is if the Causal Markov Condition is violated, certain causal dependencies can still be identified [202].

The causal Markov condition links the causal relations to dependencies and states sufficiency to model the observed probabilistic dependencies. On the other hand, the condition of stability of P w.r.t. a causal structure G states the necessity of G.

These two assumptions guarantee that observational (in)dependence (3.1) is exactly represented by the DAG-based relation (Def. 3.1.7) in a Markov compatible graph G and that causal (in)dependence (Def. 3.3) is exactly represented by standard separation in the causal structure G [101]. Furthermore, the Causal Markov condition allows the computation of interventional distributions corresponding to the *do*() operation (3.1.2) according to the "*Manipulation theorem*" ([229]) or "graph surgery" ([202]). It is performed simply by deleting the incoming edges for the intervened variables in the *do*() operator and omitting these factors from the factorization in Eq. 3.4.

3.1.3.3 The interventionist and mechanistic views

In general, the causal structure G satisfying the Causal Markov Condition for a domain with (observational) distribution P can encode all the interventional distributions in a single causal model, which is formalized in the interventionist definition of "causal Bayesian networks" [202].

This definition of causal Bayesian networks explicitly shows that the concept of causation is based on the concept of intervention, more exactly on the systematic ability to intervene. This boils down to the assumption of autonomous, local "mechanisms" composing the domain, which can be triggered by interventions independently and can be understood independently. A formalization of this "mechanism-based interpretation" of DAG representations is offered by the so-called "*functional Bayesian networks*" using a formalism of mechanisms as deterministic functions with disturbances (cf. with structural equation) [78, 202]. Whereas the functional Bayesian network formalism allows the probabilistic

modeling of counterfactuals, in the thesis we adopt a more modest causal interpretation termed "mechanism-based interpretation" meaning that under the Causal Markov condition the local probabilistic dependency models correspond to the autonomous, local mechanisms in the causal model.

3.1.3.4 Pairwise causal relations

The causal interpretation of Bayesian networks allows the definition of the following logical pairwise relations in a causal structure (recall that in stable causal models the dependency relations always represent exactly the probabilistic dependency relations):

1. *Causal path (P, $CaP(X_i, X_j|G)$):* There is a directed path from node X_i to node X_j in DAG G (also denoted by $X_i \prec_G X_j$).
2. *Causal edge (E, $CaE(X_i, X_j|G)$):* There is an edge from node X_i to node X_j in DAG G (also denoted by $(X_i \rightarrow_G X_j)$).
3. *Compelled edge (CompE, $CompE(X_i, X_j|G)$):* There is a compelled edge from node X_i to node X_j in the essential graph for DAG G.
4. *(Pure) Confounded (Conf, $Conf(X_i, X_j|G)$):* The two nodes X_i and X_j have a common ancestor in DAG G. The confounded relation is said to be pure, if there is no edge or path between the nodes.
5. *Independent (I, $Ind(X_i, X_j|G)$):* None of the previous.

Note that these pairwise relations can be also used in an acausal context using the differences w.r.t. the independence relation.

3.1.4 On the relativity of the interpretations

The causal interpretation has been challenged from many points of view. The Causal Markov assumption can be questioned as the presence of unobserved (hidden) variables as potential confounders seriously constrains the causal interpretation and automated causal discovery (for the Bayesian analysis of potentially infinite number of confounders, see [116]). Another violation called *selection bias* can occur if the observations depend on the joint combination of otherwise independent events, which induces non-causal dependencies between them. The next difficulty is related to the mixture of causal models, if conditionally both X causes Y and vice versa. A similar problem is the presence of feedback and indirectly temporality. Finally, the causal nature of the relations can be questioned because of global physical and semantic constraints between the variables [257]. It can occur if there is a global constraint on the joint set of the variables, outside the scope of the modeled domain or if the definitions of the variables are overlapping (i.e., there are logical dependencies).

In both the causal and probabilistic interpretations, the assumption of stability can be also questioned, for example because of deterministic dependencies,

resulting in the lack of guarantee for the uniqueness and exactness of the representation.

Finally, obviously the (in)dependencies are relative to the set of variables and specifically, also to the values of the variables (consider the conversion of a nth order Markov chain into a first-order by augmenting the state space), so both the probabilistic and causal interpretation has to be conditional on the set of variables and values [116].

These considerations are free of any data size issue and they are free of the question of the subjectivity of the prior in the Bayesian analysis of causation. The data set and the subjective prior information are further essential factors in the relativity of the causal and probabilistic inferences.

3.1.5 Bayesian networks in the Bayesian framework

In the Bayesian framework the prior probabilities over the Bayesian network model is represented by a joint distribution $p(G, \underline{\theta})$) over the DAG structures G and corresponding parameters $\underline{\theta}$. Because of the generality of the Bayesian network representation this distribution itself can be represented by a Bayesian network as we shall see below. However the specification of the joint or the conditionals $p(G)$ and $p(\underline{\theta}|G)$ requires practical simplifications and careful theoretical considerations, because of the huge size of the space and because of the observational equivalence of the structures. As in the thesis in general, in this section we also assume that the variables $V = \{X_1, \ldots, X_n\}$ are discrete with r_i number of values. We start with the parameter prior and then discuss the structure prior.

3.1.5.1 Parameter priors for Bayesian network models

The specification of parameter prior $p(\underline{\theta}|G)$ for Bayesian networks poses the following questions: the parametric form of the prior, the relation of the decomposition of the prior to the decomposition of P, the consistent confidence of the decomposed priors for the parts of a single structure, the consistency of the priors for observationally equivalent structures (recall that observational equivalence implies distributional equivalence in the discrete, multinomial case, see Th. 3.1.4). There is a remarkable result to clarify these problems. First, if the parameter prior decomposes w.r.t. the structure and the parameter priors are equivalent for observationally equivalent structures, then the parameter prior is a Dirichlet distribution. Furthermore, if the parts of the decomposed parameter prior are invariant w.r.t. the structure, then for any structure G $p(\underline{\theta}|G)$ can be derived from a point-specification θ_0 of a complete model and from the number of a priori seen complete cases. To state this formally, we need the following concepts. The concept of parameter independence ([228, 60]) is as follows:

Definition 3.1.19. *For a Bayesian network structure G, the global parameter*

independence assumption means that

$$p(\underline{\theta}|G) = \prod_{i=1}^{n} p(\underline{\theta}_i|G), \tag{3.10}$$

where $\underline{\theta}_i$ *denotes the parameters corresponding to the conditional* $p(X_i|\operatorname{Pa}(X_i))$ *in* G*. The local parameter independence assumption means that*

$$p(\underline{\theta}_i|G) = \prod_{j=1}^{q_i} p(\underline{\theta}_{ij}|G), \tag{3.11}$$

where q_i *denotes the number of parental configurations (*$\operatorname{pa}(X_i)$*) for* X_i *in* G *and* $\underline{\theta}_{ij}$ *denotes the parameters corresponding to the conditional* $p(X_i|\operatorname{pa}(X_i)_j)$ *in some fixed ordering of the* $\operatorname{pa}(X_i)$ *configurations. The parameter independence assumption means global and local parameter independence.*

The concept of likelihood equivalence extends observational equivalence of the structure coherently to the parameters ([131, 104]).

Definition 3.1.20. *The likelihood equivalence assumption means that for two observationally equivalent Bayesian network structures* G_1, G_2*,*

$$p(\underline{\theta}_V|G_1) = p(\underline{\theta}_V|G_2), \tag{3.12}$$

where $\underline{\theta}_V$ *denotes a non-redundant set of the multinomial parameters for the joint distribution over* V*. (The multinomiality of local models ensures distributional equivalence and that the Jacobian for parameter transformation exists.)*

Now the following theorem can be stated [104, 131].

Theorem 3.1.5 ([104, 131]). *The assumption of positive densities, likelihood equivalence and parameter independence for complete structures* G_c *implies that* $p(\underline{\theta}_V)$ *is a Dirichlet distribution with hyperparameters* $N_{x_1,\ldots,x_n}$*.*

The $p(\underline{\theta}_i|G_i) = J_{G_i} p(\underline{\theta}_V)$, where J_{G_i} is the Jacobian of the transformation from $\underline{\theta}_V$ to $\underline{\theta}_{G_i}$. It is remarkable that a structure level acausal constraint (i.e., likelihood equivalence of structures with multinomial local dependency models) implies a strong parameter-level constraint (i.e., Dirichlet parameter priors). To state the following theorem it is convenient to rewrite the hyperparameters as $N' = \sum_{x_1,\ldots,x_n} N_{x_1,\ldots,x_n}$ called *prior or virtual sample size* and $p(x_1,\ldots,x_n|\xi^+) = N_{x_1,\ldots,x_n}/N'$. Furthermore, we need the following concept:

Definition 3.1.21. *The parameter modularity assumption means that if* $\operatorname{pa}(X_i)$ *are identical in two Bayesian network structures* G_1, G_2*, then*

$$p(\underline{\theta}_{ij}|G_1) = p(\underline{\theta}_{ij}|G_2), \tag{3.13}$$

where $\underline{\theta}_{ij}$ *denotes the parameters corresponding to the conditional* $p(X_i|\operatorname{pa}(X_i)_j)$ *in some fixed ordering of the* $\operatorname{pa}(X_i)$ *configurations.*

The assumption of parameter modularity allows to induce parameter distributions for incomplete models from the parameter prior of a complete model.

Theorem 3.1.6 ([104, 131]). *If N' is the global prior sample size and $p(\underline{\theta}_V)$ is a Dirichlet distribution with hyperparameters $N_{x_1,\ldots,x_n} = N'p(x_1,\ldots,x_n)$ and the parameter modularity assumption holds and for all complete networks G_c, $p(G_c) > 0$, then for any network structure G the parameter independence and the likelihood equivalence holds and the decomposed distribution of the parameters is the product of Dirichlet distributions*

$$p(\underline{\theta}|G,\xi^+) = \prod_{i=1}^{n}\prod_{j=1}^{q_i}\prod_{k=1}^{r_i}\theta_{ijk}^{N'p(X_i=k,\mathrm{pa}(X_i,G|\xi^+)=\mathrm{pa}_{ij})-1}, \tag{3.14}$$

where r_i denotes the number of values of X_i, q_i denotes the number of parental configurations $\mathrm{pa}(X_i,G)$ and pa_{ij} denotes the values of the parents for the jth parental configuration in some fixed ordering of the $\mathrm{pa}(X_i)$ configurations.

Th. 3.1.6 offers a practical method to specify (likelihood equivalent) parameter priors for all the structures: by specifying point parameters for a complete or for a maximally detailed model $p(\underline{V}|G_c,\xi^+)$ and expressing confidence by specifying a prior sample size N' representing the complete cases underlying the point estimates (see Section 8.2.1 and Section 10.6 for its application). Then for any other model G we can compute hyperparameters according to the theorem.

However, Th. 3.1.6 also indicates that incomplete prior observations inducing different confidence for various parts of the network cannot be incorporated without violating these assumptions. For example, specifying a parameter prior as product of Dirichlets according to a structure with hyperparameters incompatible w.r.t. the theorem cannot be transformed to a product of Dirichlets for another observationally equivalent structure (i.e., the parameter prior will be different for observationally equivalent structures). In this case, the prior knowledge can be represented by a collection of incomplete cases called *prior database* instead of a *prior data set* with complete cases [116].

In case of a fixed structure G, the usage of Dirichlets with parameter independence can be attractive on its own right to specify a parameter distribution $p(\underline{\theta}|G,\xi^+)$ as follows

$$p(\underline{\theta}|G,\xi^+) = \prod_{i=1}^{n}\prod_{j=1}^{q_i}\mathrm{Dir}(\underline{\theta}_{ij}|N_{ij}) \propto \prod_{i=1}^{n}\prod_{j=1}^{q_i}\prod_{k=1}^{r_i}\theta_{ijk}^{N_{ijk}-1}. \tag{3.15}$$

3.1.5.2 Structure priors for Bayesian network models

The Bayesian approach to the parameters of Bayesian network models (reported from the end of the eighties [226, 227, 60]) provided answers for many longstanding objections against the elicitation and usage of complex probabilistic models ([43]). The Bayesian approach to the structure of Bayesian networks was similarly proposed from the beginning of the nineties, but was hindered

by the high computational demand. An ordering-specific, analytic approach was reported in [40], general analytic results and methodology were reported in [57], and the application of MCMC methods to perform Bayesian inference over structural features in [178]). With the increase in computational resources it became possible to investigate structural properties of Bayesian networks. Consequently, recently there is much emphasis on the automated or manual construction of the structure prior $p(G)$ for incorporation and for evaluation against a reference as well (see Section 8.1, 8.4, 8.6 and 10.7). Note the structure prior $p(G)$ complements the earlier investigated parameter prior $p(\underline{\theta}|G)$.

3.1.5.2.1 Using a prior data set Whereas the parameter prior and the structure prior can be specified independently, the structure prior can be induced from the *prior data set* $D_{N'}^{+}$ using Eq. 3.34 [179].

3.1.5.2.2 Using reference structure and substructures Other suggestions for the structure prior include the use of *deviation priors* (penalizing the deviations from a prior "reference" structure) and the *feature priors* (penalizing the presence and absence of various independent or dependent structural features).

The deviation prior [131] is defined by a "reference" network structure G_0 and a probability κ penalizing each missing or extra edge e_{ij} independently:

$$p(G) \propto \kappa^{\delta}, \text{where } \delta = \sum_{1 \le i < j \le n} 1(1(e_{ij} \in G) \neq 1(e_{ij} \in G_0)).$$

The *feature priors* are defined proportionally by the product of priors for the individual features (as they were totally independent). By denoting the value of feature F_i in G with $F_i(G) = f_i$, $i = 1, \ldots K$, we have

$$p(G) = c \prod_{i=1}^{K} p(F_i(G)), \tag{3.16}$$

where the c normalizing constant deals with the probability of inconsistent feature combinations $f_1, \ldots, f_K$. The possible structural features include the undirected edges or compelled edges (as direct relations or direct causal relations under the causal Markov Assumption), pairwise or partial ancestral ordering (related to causal ordering), relevance relations (Markov blankets) and even arbitrary subgraphs. However, these features are dependent in general, because of the global DAG constraint, so either the feature set should be selected carefully, or preprocessing can be applied to increase its approximation or the strength of the attached prior should reflect its approximative nature.

3.1.5.2.3 Modular priors It is particularly useful in the Bayesian analysis, if the features are "modular" in the following sense [40, 57, 131, 97]

Definition 3.1.22. *The structure modularity holds, if each feature function $F_i(G)$ depends only on the parents of X_i for $i = 1, \dots n$, defining the modular prior*

$$p(G) \propto \prod_{i=1}^{n} p(\mathrm{pa}(X_i, G)). \tag{3.17}$$

Because the DAG constraint creates dependencies, the modular features are not independent (i.e., $(F_i(G) \not\perp\!\!\!\perp F_j(G)|DAG(G))$, see Section 7.1.6), but it provides an efficient approach to define a decomposable ratio for the priors of valid structures (for certain automated corrections of the distortion because of the DAG constraint, see [42]).

A generalization of the modular prior is the *ordering-modular prior*, when modularity holds only conditionally on the orderings.

3.1.5.2.4 Edge priors With further assumption about the a priori independence of membership of edges in parental sets, we get the *directed pairwise prior* that defines the probability of each parental set as a product of individual arc probabilities. In general, the prior is defined only proportionally as follows by denoting the parents of X_i with $\mathrm{pa}(X_i) = \{\mathrm{pa}(X_i)_1, \dots, \mathrm{pa}(X_i)_{L_i}\}$:

$$p(\mathrm{pa}(X_i)) \quad \propto \quad \prod_{k=1}^{L_i} p(\mathrm{pa}(X_i)_k \in \mathrm{Pa}(X_i)) \prod_{Y \notin \mathrm{pa}(X_i)} (1 - p(Y \in \mathrm{Pa}(X_i))).$$

Originally, modular priors and directed pairwise priors were suggested conditional on a fixed ordering $\prec_0$ of the variables [40],

$$p(\mathrm{pa}(X_i)) = \prod_{\substack{X_j \prec_0 X_i \\ X_j \in \mathrm{pa}(X_i)}} p(X_j \in \mathrm{Pa}(X_i)| \prec_0) \prod_{\substack{X_j \prec_0 X_i \\ X_j \notin \mathrm{pa}(X_i)}} (1 - p(X_j \in \mathrm{Pa}(X_i)| \prec_0)), \tag{3.18}$$

in which case these features remain independent in the joint distribution over DAGs compatible with the ordering $\prec_0$. In fact, the assumption of "edge independence" first appeared implicitly in the noisy-OR canonical local dependency model, because its parameterization can be interpreted as encoding the probability of the edges [200].

To reach independent pairwise features for DAGs without constraining the ordering, we have to further simplify the features to avoid global constraints due to their interactions. Note that with independence, the marginals are not distorted and the prior is normalized, which allows the introduction of hyperparameters for modifying the prior to satisfy higher-order constraints as follows. By defining the prior over the skeleton in a pairwise manner (i.e., by retaining only the directness and omitting directionality), we get the *undirected pairwise prior* $p_{ij} \triangleq p(X_j \in \mathrm{Pa}(X_i) \vee X_i \in \mathrm{Pa}(X_j))$ represents the beliefs in direct influence between X_i and X_j [16]. The edge probabilities define the following prior

probability for a structure G:

$$P(G|\xi) \propto \prod_{i=1}^{n} \prod_{j=1}^{i-1} p_{ij}^{1(e_{ij} \in G)} (1 - p_{ij})^{1(e_{ij} \notin G)}. \tag{3.19}$$

The expectation of the number of edges L is given by $\sum_{0<i<j<n} p_{ij}$. Assuming that there is an a priori estimate for the number of direct influences in the overall model or related to a single variable, the prior p_{ij} can be scaled by an exponent ν to approximate this edge density in the prior Bayesian network (see [16]). By denoting the value that scales the expectation of the number of parental edges to L_0 with $\nu(L_0)$ we define the following scaling (it is always possible if we apply a lower limit $\epsilon < p_{ij}$ for the edge probabilities):

$$q_{ij} \triangleq p_{ij}^{\nu(L_0)}, \quad \text{with } \nu(L_0) \text{ so that } \sum_{0<i<j<n} q_{ij} = L_0. \tag{3.20}$$

Note that the scaling of p_{ij} provides an option to control the penalization (i.e., to express the prior beliefs in the prior structure). These priors except the undirected pairwise prior assign potentially different values for observationally equivalent structures (i.e., violates the structural *prior equivalence* property [131]). Because they are closely related to the causal, mechanism-based interpretation of Bayesian networks, they offer the possibility of representing a priori beliefs about the individual mechanisms in the domain and we call them *causal (structure) priors* vs *acausal (structure) priors*.

3.2 Inference methods

The Bayesian network model makes possible various types of inferences thanks to the possibility of

1. the multiple interpretation, such as causal vs. probabilistic,
2. the multilevel interpretation, such as at the level of domain values, independence relations or causal relations,
3. the adoption of the Bayesian framework at the parameter or the structure level,
4. embedding the Bayesian network model into a larger knowledge base to formulate more complex propositions (see Chapter 5).

Next we catalogue these inferences, summarize results and techniques used in the thesis.

3.2.1 Inference over values with observations

The goal in the following cases is to compute the value of marginal or conditional probabilities over domain values $P(\underline{y}|\underline{x})$ and possibly related quantities.

3.2.1.1 Fixed parameter and fixed structure

In the simplest case the structure and the parameters of a Bayesian network model are fixed. The computation of $p(\underline{y}|\underline{x})$ is NP-complete in general in the number of variables [55]. However in practice, an exact inference method has demonstrated its applicability, the *clique-tree* or *join-tree* algorithm [226]. We used this exact algorithm following the recommendations for implementation from [143]. The algorithm is exponential in the largest clique size of an intermediate Markov network and our experience similarly shows that the networks arisen in knowledge engineering and learning can be efficiently managed with this algorithm. A general result shows that the Monte Carlo approximation is hard as well: if $NP \nsubseteq RP$, then there is no random algorithm with polynomial time-complexity, whose estimate $\hat{p}$ is $|p(y|\underline{x}) - \hat{p}| < \epsilon$ accurate with δ confidence for all $\epsilon, \delta < 1/2$ [64].

3.2.1.2 Bayesian parameter and fixed structure

In case of a Bayesian approach to parameters with a fixed structure G, a parameter distribution $p(\underline{\theta}|G)$ is specified. The conditional probability over the domain values $p(\underline{y}|\underline{x}, \underline{\Theta})$ is a random variable and its mean, variance, credible regions are the target.

If the parameter distribution $p(\underline{\theta}|G)$ is specified according to the conditions of Th. 3.1.6, then it guarantees that $p(\underline{Y}|\underline{x}, \underline{\Theta})$ has a Dirichlet distribution with hyperparameters $Np_0(\underline{Y}, \underline{x})$, so the mean and credible regions can be efficiently computed.

If the parameter distribution $p(\underline{\theta}|G)$ is specified by using Dirichlet distributions and assuming parameter independence, but with arbitrary hyperparameters according to Eq. 3.15, then the marginal distribution $\bar{p}(X_1, \ldots, X_n)$ over the domain values is given by

$$\begin{aligned} \bar{p}(x_1, \ldots, x_n) &= \int p(x_1, \ldots, x_n, \underline{\theta}_1, \ldots, \underline{\theta}_n) \prod_{i=1}^{n} p(\underline{\theta}_i)\, \mathrm{d}\underline{\theta} && (3.21) \\ &= \prod_{i=1}^{n} \int p(x_i|\,\mathrm{pa}(x_i), \underline{\theta}_i) p(\underline{\theta}_i)\, \mathrm{d}\underline{\theta}_i && (3.22) \\ &= \prod_{i=1}^{n} \bar{p}(x_i|\,\mathrm{pa}(x_i)), && (3.23) \end{aligned}$$

where the $\bar{p}(x_i|\,\mathrm{pa}(x_i))$ are the local mean probabilities [228, 227, 60]. The expectations of the parameters at each node for each parental configuration (i.e., the integration of the Dirichlets) have a closed form solution (see Eq. 2.24)

$$\bar{p}(X_i = k|\,\mathrm{pa}(X_i) = \mathrm{pa}_{ij}) = \mathrm{E}_{\underline{\Theta}_i}[p(X_i = k|pa_{ij}, \underline{\Theta}_i)] = \mathrm{E}_{\underline{\Theta}_{ij}}[\Theta_{ijk}] = N_{ijk}/N_{ij}.$$

The closed solution for $\bar{p}(X_1, \ldots, X_n)$ ensures that any Bayesian inference over the domain values can be equivalently performed using this mean-valued

point parameters, instead of Bayesian averaging over the parameter space [228, 57], that is

$$\mathrm{E}_{\underline{\Theta}}[p(\underline{y}|\underline{x},\underline{\Theta})] = \bar{p}(\underline{y}|\underline{x}). \tag{3.24}$$

3.2.1.3 Bayesian parameter and structure

In the general case there is a distribution over the structures $p(G)$ and over the corresponding parameters $p(\underline{\theta}|G)$. The conditional probability over the domain values $p(\underline{y}|\underline{x},\underline{\Theta},G)$ is a random variable itself and its mean, variance, credible regions are the target. The computation of these quantities, for example of the mean involves both a summation over the space of DAGs and the integration over the parameters.

$$\bar{p}(\underline{y}|\underline{x}) = \mathrm{E}_{p(G)}[\mathrm{E}_{p(\underline{\theta}|G)}[p(\underline{y}|\underline{x},\underline{\theta},G)]]. \tag{3.25}$$

3.2.2 Inference over domain values with interventions

In the thesis the analyzed data set is observational. The interventional "do" semantics was necessary only for the causal interpretation, which is used in developing models for the analysis of domain literature with Bayesian networks. For the conversion of causally defined quantities $P(y|do(x),z)$ into "do"-free observational quantities $P(y|w)$ (question of identifiability) or to more appropriate causal quantities $P(y|do(x'),z')$ see [201, 101, 202].

3.2.3 Inference over model parameters

After the inference over the domain values we summarize now a basic result about the inductive Bayesian inference over the parameters. Let us assume the observation of a complete case x, parameter independence, and Dirichlet priors $\underline{\theta}_{ij} \sim \mathrm{Dir}(\alpha_{ij1},\ldots,\alpha_{ijr_i})$ for $i = 1,\ldots,n$ and $j = 1,\ldots,q_i$ (where r_i is the number of values of variable X_i, q_i are the number of parental configurations $\mathrm{pa}(X_i,G)_j = \mathrm{pa}_{ij}$ for variable X_i w.r.t. the Bayesian network G). Then the a posteriori distribution for an "observed" parameter family $\underline{\theta}_{ij_0}$ where j_0 is the index of $pa_i(x)$ is given by

$$\begin{aligned} p(\underline{\theta}|x) &= \frac{\prod_{i=1}^{n} p(x_i|pa_i(x),\underline{\theta}_{ij_0})p(\theta_{ij_0})}{p(x)} \prod_{i=1}^{n}\prod_{j\neq j_0} p(\theta_{ij}) && (3.26)\\ &\propto \prod_{i=1}^{n} \theta_{ij_0x_i}\mathrm{Dir}(\theta_{ij_0}|\underline{\alpha}_{ij_0}) && (3.27)\\ &\propto \prod_{i=1}^{n} \mathrm{Dir}(\theta_{ij_0}|\alpha_{ij_01},\ldots,\alpha_{ij_0x_i}+1,\ldots,\alpha_{ij_0r_i}), && (3.28) \end{aligned}$$

which shows that the parameter posterior preserves the parameter independence property and that local standard Bayesian updating is performed on the hy-

perparameters of the "observed" Dirichlets (the hyperparameters for the other parameter families $\underline{\theta}_{i_0 j}$ with $j \neq j_0$ are unchanged).

3.2.4 Inference over model structures

The posterior of the Bayesian network (structure) is the product of the model likelihood and the structure prior.

$$p(G|D_N) \propto p(G) \int p(D_N|\underline{\theta}, G) p(\underline{\theta}|G) \, \mathrm{d}\underline{\theta} = p(G) p(D_N|G). \tag{3.29}$$

To reach a closed form for the likelihood term we continue with the assumption of the previous paragraph: N complete observations, i.i.d. multinomial sampling, Bayesian network model with parameter independence and Dirichlet parameter priors following [57, 227, 131]. Under these assumptions the observation of a complete case results in a local standard Bayesian updating of the hyperparameters of the "observed" Dirichlets retaining the parameter independence (see Eq. 3.26). The maintained parameter independence allows a standard parental decomposition w.r.t. the Bayesian network G for each observation (see Eq. 3.21), which allows the following rearrangement:

$$\begin{aligned} p(x^{(1)}, \ldots, x^{(N)}|G) &= \prod_{l=1}^{N} \prod_{i=1}^{n} p(x_i^{(l)}|pa_i^{(l)}) && (3.30) \\ &= \prod_{i=1}^{n} \prod_{l=1}^{N} p(x_i^{(l)}|pa_i^{(l)}) && (3.31) \\ &= \prod_{i=1}^{n} \prod_{j=1}^{q_i} \prod_{l=1}^{N} p(x_i^{(l)}|pa_{ij})^{1(\mathrm{pa}_{ij}=\mathrm{pa}_i^{(l)})}, && (3.32) \end{aligned}$$

where $pa_i^{(l)}$ denotes the value(s) of parental set of X_i in case l. The marginal probability of the data for a single Dirichlet prior and multinomial sampling was derived in Eq. 2.24 and Eq. 2.24, 2.25. Now if r_i denotes the cardinality of the discrete values of variable X_i, α_{ijk} the initial Dirichlet hyperparameters, and n_{ijk} the number of occurrences for the variable X_i, its parental configuration pa_{ij} and its value r_k, then for each variable X_i and parental configurations j independently

$$\begin{aligned} \prod_{l=1}^{N} p(x_i^{(l)}|pa_{ij}, G)^{1(\mathrm{pa}_{ij}=\mathrm{pa}_i^{(l)})} &= \frac{\prod_{k=1}^{r_i} (\alpha_{ijk} \ldots (\alpha_{ijk} + n_k))}{\alpha_{ij+} \ldots (\alpha_{ij+} + n)} && (3.33) \\ &= \frac{\Gamma(\alpha_{ij+})}{\Gamma(\alpha_{ij+} + n_{ij+})} \prod_{k=1}^{r_i} \frac{\Gamma(\alpha_{ijk} + n_{ijk})}{\Gamma(\alpha_{ijk})}, \end{aligned}$$

Putting everything together, if the prior satisfies the structure modularity, then the posterior of the Bayesian network structure has the following product

form

$$p(G|D_N) \propto \prod_{i=1}^{n} p(\mathrm{Pa}(X_i, G))S(X_i, \mathrm{Pa}(X_i, G), D_N) \text{ where} \tag{3.34}$$

$$S(X_i, \mathrm{Pa}(X_i, G), D_N) = \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij+})}{\Gamma(\alpha_{ij+} + n_{ij+})} \prod_{k=1}^{r_i} \frac{\Gamma(\alpha_{ijk} + n_{ijk})}{\Gamma(\alpha_{ijk})}.$$

3.3 Knowledge engineering

As discussed in Section 3.1 and enumerated in List 3.2, the Bayesian network can serve as a multilevel (structural or parametric), multiple-point-of-view (probabilistic or causal) representation of the domain. Besides being a model ("*surrogate*"), it fulfills other important roles of a knowledge representation (following the proposed roles from [67]): *ontological* (what kind of objects and relations exists in the domain), *inferential* (what kind of inference is possible in the domain), *computational* (what kinds of embedding of the model and real-world applications are possible), *communicational* (what kind of understanding and communication is supported by the model between domain experts, knowledge engineers, and users).

Because of the versatility of the Bayesian network representation as a knowledge representation, knowledge engineering methodologies are necessary for proper and efficient real-world applications. Particularly, if a Bayesian network model serves as a probabilistic expert system or as the engine of a decision support system, its construction should be subject to engineering standards, which include specifications with quantitative quality measures for the process and the product and complexity measures related to budgetary, personal and time limits, etc. However, these issues are still largely unexplored and the knowledge engineering of Bayesian networks is still in its early stage (described for example in [1]). The main reasons are the versatility of the representation mentioned above, the continuing extensions of the representation and the newly evolved knowledge engineering context of the "e-science" era.

The "classical" knowledge engineering of Bayesian networks in complex domains was criticized as aiming at a "one-shot" and "monolithic" Bayesian network. Its extension led to new representational methods, especially to modularized representations [207, 182, 77, 196, 168]. The object-oriented and frame-based approaches were partly responses to problems of modularization, validation, verification, maintenance and reuse [167, 155, 156]. Other approaches extended the Bayesian network representation itself. The multi-net representation was partly a response to a problem related to the elicitation and representation of contextual independencies [105]. The qualitative Bayesian networks and other semantic extension of the represented relations were partly a response to the problem of the elicitation and refinement of parameters [254, 170, 211], similarly to the investigation of special local dependency models [130, 94].

3.4 Prequential analysis by Bayesian networks

The Bayes factor in Eq. 2.13 is typically used in a non-sequential setup. In Section 2.4.1 we summarized the prequential framework, which evaluates the model from a forecasting point of view by scoring its sequential predictions based on the actual observations [227, 60]. Because of its sequentiality, it also offers a sample-by-sample evaluation of the compatibility of the data and the model (see Section 8.2). For us, the case of a (discrete and finite) probabilistic forecasting system (PFS) is relevant predicting a distribution $p(X_i|x_1,\ldots,x_{i-1})$ for the observation at step i. For the application of the prequential evaluation for Bayesian networks and parts of the model we have to interpret them as PFSs and compare them using the logarithmic score (see Eq. 2.33).

The PFS shall be defined as a Bayesian forecasting system (see Section 2.4.1) using a fixed Bayesian network structure with Dirichlet parameter priors under the condition of parameter independence.

The *global monitor* tracks the overall performance of the Bayesian network model $M = (G, \underline{\theta})$ over a data set D_N:

$$\begin{aligned} S(M; D_N) &= \sum_{l=1}^{N} -\log p(x^{(l)}|x^{(1)},\ldots,x^{(l-1)}, M) && (3.35)\\ &= -\log p(x^{(1)},\ldots,x^{(N)}|M). && (3.36)\end{aligned}$$

The equation shows the ordering-insensitivity and batch-sequential equivalence of the log-score for PFSs. By noting that this is the model likelihood derived in Eq. 3.30, 3.33, the score is given by

$$S(M; D_N) = -\log \prod_{i=1}^{n}\prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij+})}{\Gamma(\alpha_{ij+}+n_{ij+})} \frac{\prod_{k=1}^{r_i}\Gamma(\alpha_{ijk}+n_{ijk})}{\prod_{k=1}^{r_i}\Gamma(\alpha_{ijk})}. \quad (3.37)$$

In line with the decomposition w.r.t. the structure (see Eq. 3.34) various monitors were suggested for the parts of the Bayesian network model.

The (unconditional) *node monitor* tracks the performance of the Bayesian network model M w.r.t. a given variable X_i:

$$S(X_i; D_N) = -\log \prod_{l=1}^{N} p(x_i^{(l)}|x^{(1)},\ldots,x^{(l-1)}, M). \quad (3.38)$$

Two variants of the node monitor are the conditional node monitors, because the target variable is predicted conditioned on all the other variables or only on the parental set in the actual case. This monitor was called a "conditional node monitor" [227]), but in the case of complete data assumption this is equivalent with scoring the predictive performance of the Markov blanket subgraph $\mathrm{MBG}(X_i, G)$. So we will adopt the term *Markov blanket subgraph monitor*.

$$S(\mathrm{MBG}(X_i, G); D_N) = -\log \prod_{l=1}^{N} p(x_i^{(l)}|x^{(l)} \setminus \{X_i\}, x^{(1)},\ldots,x^{(l-1)}). \quad (3.39)$$

Conditioning only on the parental set in a causal approach, we get the *mechanism monitor* that tracks the performance of the parental set $\mathrm{Pa}(X_i, G)$ in forecasting a variable:

$$S(\mathrm{Pa}(X_i, G); D_N) = -\log \prod_{l=1}^{N} p(x_i^{(l)} | \, \mathrm{pa}(X_i) = \mathrm{pa}_i^{(l)}, x^{(1)}, \ldots, x^{(l-1)}). \quad (3.40)$$

The specialization of the mechanism monitor is the *configuration monitor* that tracks the performance of a parental set in case of a specific parental configuration pa_{ij}:

$$S(\mathrm{pa}_{ij}; D_N) \quad = \quad -\log \prod_{l=1}^{N} p(x_i^{(l)} | pa_{ij}, x^{(1)}, \ldots, x^{(l-1)})^{1(\mathrm{pa}_i^{(l)} = \mathrm{pa}_{ij})}. \quad (3.41)$$

By these definitions we can rewrite the model score as the sum of the mechanism monitors or the total sum of all of the configuration monitors in M.

$$S(M; D_N) = \sum_{i=1}^{n} S(\mathrm{Pa}(X_i, G); D_N) = \sum_{i=1}^{n} \sum_{j=1}^{q_i} S(\mathrm{pa}_{ij}; D_N). \quad (3.42)$$

The application of the model monitor, mechanism monitor and parent-child monitor in the ovarian cancer domain are reported in Section 8.2.

3.5 Learning Bayesian networks

By now we summarized a framework for general, normative, inductive inferences using probabilistic domain models: the Bayesian decision-theoretic framework with Bayesian networks. Frequently, it is restricted to optimization, particularly over structures, which is termed the "standard" Bayesian network (structure) learning, not necessarily within the Bayesian decision theoretic framework. This mode of operation is particularly relevant if a large amount of data is available w.r.t. the complexity of the model. So, in this section we finish our overview with the summary of the score-based learning of Bayesian networks, including Bayesian and non-Bayesian inductive scores and search algorithms.

Another large family of methods for finding complete models best fitting the observations are the constraint-based algorithms. These construct a network by performing independence tests with certain prespecified significance level. Assuming no hidden variables, a stable distribution and correct hypothesis tests, the Inductive Causation (IC) algorithm correctly identifies a Bayesian network that exactly represents the independencies (see [202, 116, 229]). It means that the score-based and the constraint-based learning algorithms behave identically for stable distributions in the limit w.r.t. the sample size. However, there is no generally recommendable prespecified significance level and final significance level for the identified model. Furthermore, because of the frequentist approach, there is no principled way to incorporate uncertain prior information. On the

other hand, efficient constraint-based algorithms exist that work in the presence of hidden variables, which is currently not tractable with Bayesian methods.

Our assumption of complete, observational and discrete data modeled with a fixed set of discrete variables is a serious restriction, but it provides a sufficient conceptual framework to develop the main topics in the thesis such as the (automated) construction of priors, the computation of posteriors of complex structural features and their role in classification. We direct the reader to the following sources regarding the treatment of mixture of discrete and continuous variables [169, 60, 131]; the mixture of observational and interventional data [116]; the issue of incomplete data[108, 90]; the issue of special local probabilistic dependency models [94] and the issue of temporal data and variables [218].

3.5.1 Score functions and their properties

The score-based learning of Bayesian networks best fitting to the data D_N consists of the definition of a score function $S(G, D_N) : \{G \times D_N\} \rightarrow \mathcal{R}$ and a search method in the space of DAGs. In a Bayesian decision theoretic framework the score function is specified as the expected loss $\mathrm{E}_{P(\hat{G}|D)}[L(G, \hat{G})]$ of selecting (i.e., reporting) the structure $\hat{G}$. Whereas the advantages of knowledge rich utility functions are apparent, standard score functions lack domain knowledge. For example, in case of 0-1 utility function the model with maximum expected utility corresponds to the structure with *maximum a posteriori* probability or in case of uniform prior to finding the *maximum likelihood* structure.:

$$G^{\mathrm{MAP}} = \arg\max_{\hat{G}} \mathrm{E}_{p(G|D)}[L(G, \hat{G})] = \arg\max_{\hat{G}} p(\hat{G}|D), \text{ if } L(G, \hat{G}) = 1(G = \hat{G}). \tag{3.43}$$

In Eq. 3.34 we derived a closed form for the posterior of a structure G,

$$p(G, D_N) = p(G) \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij+})}{\Gamma(\alpha_{ij+} + n_{ij+})} \prod_{k=1}^{r_i} \frac{\Gamma(\alpha_{ijk+n_{ijk}})}{\Gamma(\alpha_{ijk})}, \tag{3.44}$$

termed *Bayesian Dirichlet* metric [131]. If the initial hyperparameters $\underline{\alpha}$ satisfy the conditions of Th. 3.1.6 (ensuring indistinguishability within an equivalence class), then it is denoted as BD_e. If the initial hyperparameters $\underline{\alpha}$ are constant 1 then it is denoted by $\mathrm{BD}_{\mathrm{CH}}$ [57]. If the initial hyperparameters are the converse of the number of parameters corresponding to the local, overall multinomial models of the variables then it is denoted by $\mathrm{BD}_{\mathrm{eu}}$ [40, 131]. The corresponding score functions are defined as $\mathrm{BD}(G; D_N) = \log(p(G, D_N))$.

Another family of non-Bayesian score functions can be derived within the likelihood framework. The maximum likelihood score is defined as follows

$$\mathrm{ML}(G; D_N) = \max_{\underline{\theta}} p(D_N|G, \underline{\theta}). \tag{3.45}$$

We used only the following MDL/BIC-score defined as follows

$$\mathrm{BIC}(G; D_N) = \log(\mathrm{ML}(G; D_N)) - \frac{1}{2}\dim(G)\log(N), \tag{3.46}$$

where $dim(G)$ denotes the number of free parameters. For overviews of other score functions and for the derivation of the BIC-score, see [163, 39, 50, 99, 132]. We discuss now the properties scoring metrics w.r.t. observational equivalence and sample size.

Definition 3.5.1. *A score function $S(G; D_N)$ is called score equivalent, if for each pair of observationally equivalent Bayesian network structure G_1, G_2 the scores are equal $S(G_1; D_N) = S(G_2; D_N)$ for all D_N [131].*

Theorem 3.5.1 ([131]). *The $BD_e(G; D_N)$ scoring metric is likelihood equivalent, that is if G_1, G_2 are observational equivalent, then $p(D_N|G_1) = p(D_N|G_2)$. Furthermore, if the structure prior is acausal (i.e., equal for such G_1, G_2), then the BD_e scoring metric is score equivalent [131].*

Consequently, the score can be used directly in an acausal approach if the hypotheses are the observational equivalence classes. In a causal approach to Bayesian network structure learning with the BD metrics the structure prior can incorporate information differentiating observationally equivalent structures, which means an asymptotically vanishing term w.r.t. the likelihood term. The differentiation within an equivalence class by a non-likelihood equivalent BD score (i.e., by a non-likelihood equivalent parameter prior such as the $\mathrm{BD_{CH}}$) is similarly vanishing.

The score equivalence of the BIC score is the direct consequence of the result that the number of free parameters (i.e., the term $dim(G)$) are equal in observationally equivalent Bayesian networks (here again as throughout the thesis, we assume discrete variables and multinomial local dependency models) [39, 50, 49].

Theorem 3.5.2 ([49]). *The $BIC(G; D_N)$ scoring metric is score equivalent.*

Results about asymptotic consistency and rate of convergence results for maximum likelihood scores are derived in [39, 99]. For the sample complexity of parameter learning, see [65].

3.5.2 Search algorithms for finding high-scoring BNs

As discussed in the beginning of this section, the recently used loss functions or more generally the score functions $S(G, D_N)$ are usually efficiently computable in $\mathcal{O}(nN)$. It is partly the consequence of the decomposability of the score, which allows even further computational speed-ups as discussed later on. However, the global DAG constraint does not allow the decomposition, so we have to perform a combinatorial optimization in the space of DAGs over n nodes (variables). The cardinality of the space of DAGs is given by a recursion [57]. By neglecting the DAG-constraint, this can be bounded by the number of the combinations of the edges between different nodes ($2^{n(n-1)}$). By limiting the maximum number

of parents to k it is still super-exponential (consider that the number of parental sets for a given ordering of the variables is in the order of n^{kn}, so $2^{\mathcal{O}(kn \log n)}$ [96]).

The computational complexity of learning BNs in the constraint-based and in the score-based framework is bounded by the following two theorems (assuming $P \neq NP$). The first states the NP-hardness of finding a Bayesian network for the observations (as minimal representation of the observed independencies, see Def. 3.1.12) [39]. The second theorem states the NP-hardness of finding a best scoring Bayesian network (i.e., the NP-hardness of optimization over DAGs) [50].In the special case of $k = 1$ (that is for trees and polytrees) standard maximum weight spanning tree (MWST) construction algorithms can be applied, which has polynomial time complexity, see [200, 50]. The NP-hard nature of the problem remains if the learning takes place over the smaller space of equivalence classes [50, 152].

Consequently, a frequently used suboptimal approach is to use iterative improvement algorithms with local search. These start from a good or at least a neutral candidate satisfying the prior knowledge and the DAG constraint. In each step i a structure with an improved score is selected from the prespecified neighborhood $Nb(G_i)$ of G_i, otherwise the algorithm is stopped. Usually this neighborhood is defined as structures with 1 edge difference. However, the result of the iterative improvement algorithms with local search is probably a local optimum, so frequently the algorithms are restarted with a random initial candidate. This problem can be avoided by replacing the greedy element of the algorithm with a stochastic scheme allowing selections of structures with worse score, as in the simulated annealing algorithm. A greedy algorithm called $K2$ can be applied if the score is decomposed and the ordering of the variables are well-restricted, because for each ordering the parental sets can be optimized independently with a greedy algorithm [57]. Studies of the performance of various iterative improvement algorithms using local search and simulated annealing are reported in [50, 39], which indicate a robustly good performance with relatively low computational complexity for the $K2$ algorithm without tuning to the domain, data set, etc. Our experiments in the ovarian cancer domain with various iterative improvement algorithms with local search and simulated annealing algorithm similarly strengthened this result. In the thesis the reported results are usually computed with a $K2$ variant algorithm using the implementational tricks of the sample tree to compute the score for a parental set in $\mathcal{O}(N)$ as proposed in [57] and storing the parental scores as also proposed in [40].

Chapter 4

Prior knowledge and data about ovarian cancer

We overview domain variables, known risk factors, preoperative classification models, and we describe the statistical data sets from the IDO and the IOTA project. We document the results of knowledge engineering. On the one hand, this chapter summarizes the elicited expert knowledge about domain variables, pairwise relations and complete domain models, partly with complete parameterization. On the other hand, it summarizes the automatically collected original and the derived, secondary electronic resources, such as the so-called "literature data" sets.

4.1 The biomedical background, the IDO, and the IOTA projects

We shall illustrate our theoretical developments on a real-world medical problem related to *ovarian cancer*. Ovarian malignancies represent the greatest challenge among gynecologic cancers. Early detection is of primary importance for the survival of the patient, since currently more than two-thirds of the patients are diagnosed only at an advanced stage and therefore have poor prognosis. A reliable test to discriminate between benign and malignant tumors before surgery (i.e., a preoperative diagnosis) would be of considerable help to clinicians. It would help them recognize patients for whom treatment with minimally invasive surgery or conservative management suffices versus those for whom referral to a gynecologic oncologist is needed for more aggressive treatment. There are two different types of information for the development of such predictive models: the biological and medical information about the disease and the growing amount of patient data.

The doctoral research started within the framework of the IDO project at the K.U.Leuven, which was aimed at developing "Predictive computer models for

medical classification problems using patient data and expert knowledge". The main work took place in the context of the *International Ovarian Tumor Analysis Consortium* (*IOTA*), which is a multicenter study on ovarian tumors [240]. Its main goal is the preoperative prediction of malignancy of ovarian masses by fusing expert knowledge and statistical data. This study also includes the multicenter collection of patient data and the corresponding data collection protocols. For an overview of the process of the web-based medical data collection and quality checking, see [5].

4.1.1 The domain and domain concepts

The abundant background knowledge is diverse: for example, the MEDLINE collection of abstracts from biomedical journal papers contains thousands of items about ovarian cancer. The most common ovarian malignancies are the epithelial cancers, which arise from the cover of the ovary. Various theories hypothesize that the malignant transformation is related to the number of ovulations, to the level of gonadotropins, carcinogens, and also to genetic defects. Factors known to affect the risk of malignancy are parity (number of pregnancies), infertility treatment, duration of lactation, oral contraceptives, foreign bodies (carcinogens), family history of breast and ovarian cancer, genetic defects, age, age at menopause, and hysterectomy. Besides clinical data, additional measurements and observations used in standard clinical diagnosis are the following: bilaterality of the tumor, pelvic pain, morphological descriptors of the mass (such as smoothness and solidness), descriptors of its echogenicity and vascularization, level of several serum tumor markers, such as CA125, amount of fluid in the abdominal cavity and the day of the cycle. While the effect of some of these variables can be quantified reliably such as the effect of the family history and genetic defects (e.g., familial $BRCA_1$ and $BRCA_2$ mutations), other effects are only qualitatively known and highly subjective (e.g., the use of the vascularization indices).

In the experiments, we used thirty-five *domain variables*, which had been previously evaluated as the most relevant domain variables, such as pathology (benign vs. malignant), parity, drug treatment for infertility, use of oral contraceptives, family history of breast and ovarian cancer, age, bilaterality of the tumor, pain, descriptors of the morphology, echogenicity, and vascularization of the mass, or the level of CA125 tumor marker (see Table A.1 for their definitions). For the IOTA nomenclature and taxonomy, and the measurement procedures, see [240]. Twenty of the variables are nominal or a nominal interpretation has been provided by the IOTA protocol. For the rest of the variables, a medical expert provided commonly used thresholds for their discretization, which are shown in Table A.2.

4.1.2 Previous predictive models

The first predictive models were based on single variables (such as CA125, resistance index) or risk indices (Lerner's scoring system, risk of malignancy index

(RMI)). Standard statistical studies indicated that a multimodal approach — the combination of several variables — is necessary for the discrimination between benign and malignant tumors [63, 41, 137]. Therefore several studies have applied logistic regression [239], multilayer perceptrons [249, 250, 238], support vector machines [173, 172], and Bayesian networks [15, 14].

4.2 The data sets

In addition to the prior background information, two continuously growing data sets were used in our work, the IDO and IOTA data sets.

4.2.1 The IDO data set

The IDO data set has been collected prospectively from 300 consecutive patients who were referred to a single institution (University Hospitals Leuven, Belgium) from August 1994 till June 1997. The data collection protocol excludes other causes with similar symptoms, such as infection or ectopic pregnancy and ensures that the patients with persistent extrauterine pelvic mass undergo surgery. This eliminates the possibility of false negatives and the quality of this single center study provides reliable pathology values as gold standard (for a detailed description, see [237, 238, 240]).

4.2.2 The IOTA data sets

The *IOTA* data sets have been collected in the framework of the IOTA project, consisting of 68 parameters for over 1,150 tumors [240]. In our experiments, we included the cases satisfying the IOTA protocol, excluded cases without measurement of the serum CA125 level and use of oral contraceptives, which were not mandatory variables for the data collection but used in our prior extraction.

Because of the ongoing data collection, two data sets have been formed from this source. The *IOTA-1.1* data set contains thirty-one variables and the completely observed cases with respect to the selected variables (604 cases) denoted by D^{IO_1}. The variables and the corresponding univariate statistics is shown in Table A.2. Figure A.1 shows the biplot of the IOTA-1.1 data set and the variables. The biplot shows variables and cases in the plane spanned by the first two principal components. In particular, a small angle between variables such as (Age, Meno, PostMenoY) indicates high correlation between those variables. The observations of malignant tumors (indicated by $\diamondsuit$) tends to be correlated with high values for certain morphologic variables, such as Papillation or WallRegularity, but with relatively low values for variables such as PillUse and Shadows. The *IOTA-1.2* data set D^{IO_2} includes four additional variables: the "Familial history of breast cancer" (FamHistBrCa) and the "Familial history of ovarian cancer" (FamHistOvCa) (which are the original variables from which the variable "Familial history" variable (FamHist) is derived), and the "Postmenopausal age" (PMenoAge) and the "Reproductive years" (ReprYears)

derived from the variables "Age", "Postmenopausal years" (PMenoY) assuming 12 years for the age at menarche. It contains 782 complete cases, including the samples of the IOTA-1.1, but with a few errors detected and corrected. The biplot and the sorted eigenvalues of the covariance matrix of the IOTA-1.2 data set containing 782 complete cases are shown in Fig. A.2 and Fig. A.3. Note that both the IOTA-1.1 and IOTA-1.2 data sets differ from the data set of the official first release of the IOTA consortium [239, 137].

4.3 Knowledge engineering BNs

We developed a series of Bayesian networks and various corresponding formal and informal resources as the prior background information. In the first phase between 1999-2000 we followed a "classical" knowledge engineering methodology to construct Bayesian belief networks with parameter priors. It was done with the help of a domain expert Dirk Timmerman and by pooling statistics manually from heterogenous sources reported in the literature. This phase took place within the framework of the IDO project using its data. It was reported in [28, 27] and it is summarized in Section 4.3.1. This model was applied in our work for use as an auxiliary domain model with classifiers, particularly in the experiments on transforming a Bayesian network parameter prior into a parameter prior for a conditional model by projection and virtual sample.

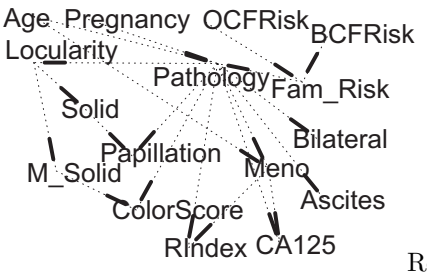
In the second phase between 2000 and 2001 at the start of the IOTA project, first we concentrated on the electronic domain literature and domain ontologies, and we constructed various textual resources as a foundation for the ABN-KB. This is reported in Section 4.3.4. Mainly influenced by the evaluation of the pairwise, associative statistical text-mining methods, with the help of Dirk Timmerman we constructed a knowledge base about the relevance (dependency) relations in the domain, particularly focusing on the direct, pairwise dependency relations and identifying the acausal, semantic relations.

In the third phase in 2002, we performed on the one hand an experiment to elicite the point parameters for a small Bayesian network and on the other hand a domain expert constructed three embedded causal domain structures. The pairwise relevance information and the reference structures from the domain experts are reported in Section 4.3.3, the estimated point parameters are in Section 4.3.2.

4.3.1 An early Bayesian network for ovarian cancer

For the early Bayesian networks, we used thirteen variables where the continuous and integer variables were discretized according to the medical literature and expert knowledge. We built a "heterogeneous" belief network containing biological models of the underlying mechanism quantifiable by the literature, parts quantified by a medical expert and parts quantified from previous studies (for a more detailed description of the model construction process, see [28]). The prior belief network is shown in Fig. 4.1 and we derived the hyperparameters

for the Dirichlet parameter priors manually from heterogenous sources shown in the table in Fig. 4.1.



Variable	Source
Genetic risk	[33, 81, 165, 256, 52]
Bilaterality	[244]
CA125	[87]
Color score	[245]
Locularity	[118]
Papillation	[118]
Pathology	[33, 81, 165, 256, 52, 124]
Resistance index	[233]

Figure 4.1: A early BN model for the ovarian cancer problem (left). Relevant publications providing information for certain variables in addition to the expert's opinion (right).

This work started in the "classical" knowledge engineering framework described in Section 3.3, but it has been gradually shifted towards the "Bayesian" knowledge engineering framework described in Section 5.1. Because of the extensive and complex usage of the prior knowledge, we used a strict documentation method to track the route of the prior information from studies into the model. With the Bayesian approach to data analysis, it has led to the concept of a Bayesian computational environment with rich informal and formal knowledge elements, the ABN-KBs and ABNs (see Section 5.2), to the concept of model profiling and ABN-based information retrieval (see Chapter 5), and to the concept of ABN-based text-mining (see Chapter 6).

4.3.2 Parameter priors for a small-scale model

In the second experiment of parameter elicitation *parameter prior* to restrict the number of free parameters (to 400) we used only highly relevant variables and relations in a small-scale model (see Fig. A.4).

4.3.3 Elicitation of structural priors

We elicited three kinds of structural information: relevance in the domain, reference structures and decomposed beliefs for structural properties. To identify the relevant domain variables, we asked the experts to construct minimal sets of variables and score their relevance for the prediction of the target variable *Pathology*. Because of our assumption of complete observations, this task was equivalent to score the Markov blankets MB(*Pathology*). These are identifiable from the domain structure; and indeed, for the expert it was difficult to score abstractly the Markov blankets without considering the structural aspects behind them. So we used this information only informally for designing experiments and not as an element of a structural prior. Second, the expert specified three embedded structures, which were used as reference in evaluation and in

deviation priors. Third, the expert specified his belief in direct pairwise, possibly directed (causal) relations, which can be used in edge priors. We summarize these results below with their comparisons.

4.3.3.1 Prior structures from a model-based approach

The three Bayesian network structures G^H, G^M, and G^R specified by the medical expert are shown in Fig. 4.2 (they are embedded). We proposed the causal-mechanistic interpretation for determining the parental sets, even though the logical relations, the abstraction level and the set of variables corresponding to the IOTA study sometimes hindered it. In this interpretation, the belief in domain models is the belief in the joint collection of mechanisms for each domain variable. This joint belief in our case can be approximately decomposed (i.e., the embedded prior domain models can be conceived of an embedded prior for each variable for its mechanisms or practically for its parental sets). The corresponding set of edges (i.e. parents) for variable X_i with "high", "moderate" and "reasonable" relevance are denoted by S_i^H, S_i^M, and S_i^R, which define the embedded G^H, G^M, and G^R reference structures respectively.

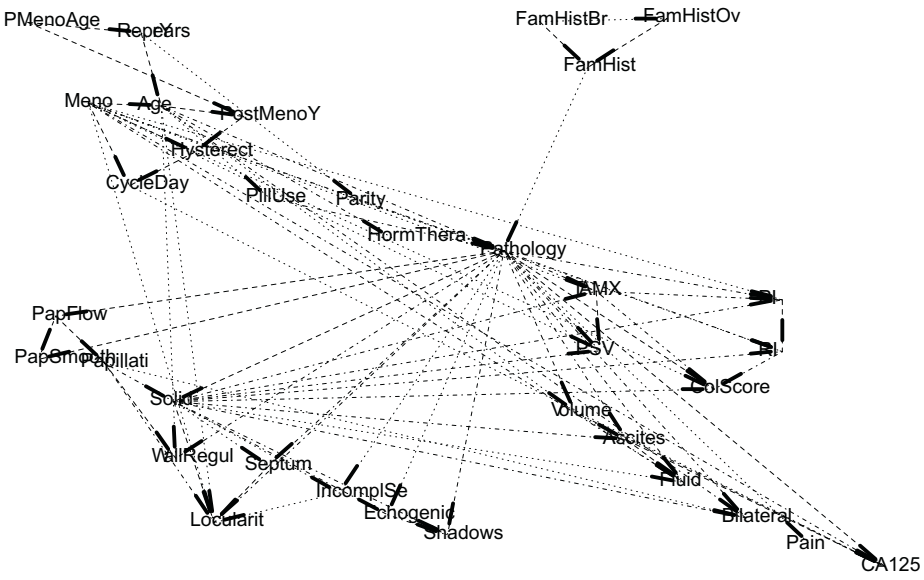


Figure 4.2: Three Bayesian network structures G^H, G^M, and G^R specified by the medical expert for the thirty-five IOTA variables used in the second stage. The most parsimonious BN G^H is denoted with dashed lines (containing a highly relevant collection of parental sets), the medium BN G^M with dashed-dotted lines and the relevant BN G^R with dotted lines (they are embedded).

4.3.3.2 Priors from a pairwise relevance approach

The expert rated the pairwise relations inclusively as 'highly', 'moderately', 'reasonably" and 'weakly" relevant direct dependencies. Furthermore, the expert provided rankings for the pairwise relations, which was manually transformed into a prior $R_{\text{Expert}}(X;Y)$ for undirected edges. We applied the scaling in Eq. 3.20 to the prior score to satisfy the condition that the average pairwise direct relations per variable is 3, furthermore we set a lower limit ϵ to avoid the a priori exclusion of edges. The expert also indicated the tentative causal ordering of the pair or the logical relation, specifically as $\{one, many\} \times \{one, many\}$ relations.

For example, $R_{\text{Expert}}(\text{Pathology};Y)$ represents an assessment of the relevance of each domain variable Y with respect to the *Pathology* variable—that is, to discriminate between benign and malignant tumors. Later we use the notation that S^h denotes the set of "most relevant" relations, S^m denotes the set of both "most relevant" and "moderately relevant" relations, and S^r denotes the set of all relations and $S_P^{h,m,r}$ denote the respective subsets of the relations corresponding to the central variable *Pathology*.

4.3.3.3 The causal ordering of variables

The elicited multiparental and pairwise structural relations define embedded partial orderings of the variables. The reference total ordering $\prec^c$ was derived (Table A.1) by the resolution of the partly logical pairwise relations.

4.3.4 Electronic resources for knowledge engineering

To derive prior knowledge about the domain concepts and their relations from electronic resources, we experimented with the UMLS meta-ontology, which includes multiple taxonomies, standardized vocabularies such as the ULT93, MSH2002-06 and SNMI98 collections [197]. We mapped the IOTA entities (groups, variables, values) to UMLS concepts. Then we derived structural priors by inducing quantified relations from UMLS relations and from their various combinations. The results of these experiments were unsuccessful, because of the heterogenous and very noisy ontologies within the UMLS, so we report results w.r.t. electronic free-text.

4.3.4.1 Text kernels

To use the electronic resources we constructed a *text kernel* for each domain variable, which includes the name of the variable, synonyms, a free-text description (the kernel) and references to documents. To ensure consistent, objective annotations, we used the IOTA protocols without modification as primary sources. A corresponding Ph.D. thesis [237] provided an extension for the IOTA descriptions. Together, these compose the text kernels, on average a hundred-word description for each of the domain variables. Additionally, we let these kernels contain references to the Merck Manual [141], the Online Medical Dictionary

[140], the CancerNet Dictionary [139] and the MEDLINE collection of abstracts of the US National Library of Medicine [142], which are used optionally in deriving a vector representation with user-specified weights (see Section 5.3).

4.3.4.2 Document collections

We asked medical experts to select the *most relevant* journals for the domain (2 journals), the *highly relevant* (3 journals), the *moderately relevant* (33 journals) and the *relevant* journals (93 journals). Based on these, four embedded collections of MEDLINE abstracts were constructed containing 5,367, 71,845, 231,582, and 378,082 abstracts denoted by ME_{HMR-} selected from the MEDLINE corpus. We also constructed a more restricted series of embedded collections based on the results of Pubmed [142] to the query "ovarian mass/tumor/cancer" in March of 2003 denoted by PM_{HMR-}, which contain $2,256$, $3,301$, $9,372$, $12,038$, and $35,562$ abstracts. The Medline corpus contains papers mostly between 1985 and 2000; the PM corpus from 1980 till 2002. The annual number of papers in the MEDLINE (ME) and in the PubMed (PM) corpora with rates high (3), medium (2), reasonable (1) and all (0) between 1980 and 2003 are reported in Fig. 4.3.

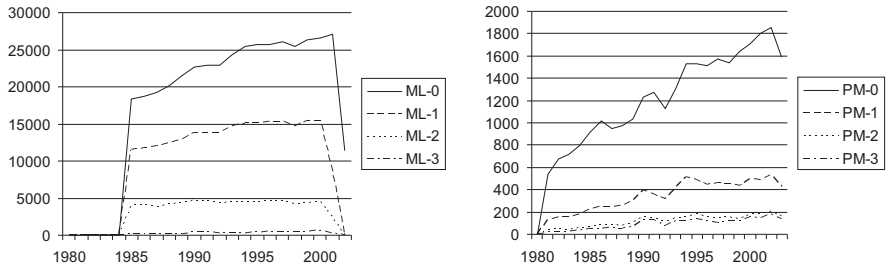


Figure 4.3: The annual number of papers in the MEDLINE and in the PubMed corpora with rates high (3), medium (2), reasonable (1) and all (0) between 1980 and 2003.

4.3.4.3 Domain vocabularies

We constructed manually a small domain vocabulary containing less than one thousand words with domain-specific phrases and synonyms. It follows the IOTA terminology [240] and the guidelines for controlled indices as well [51]. Furthermore, we constructed a domain vocabulary containing more than one million words by incorporating statistically relevant words and manually curated general medical vocabularies, such as MeSH*.

*http://www.nlm.nih.gov/mesh/meshhome.html

Chapter 5

Fusing BNs and logical knowledge bases

We first summarize the new context of knowledge engineering of Bayesian networks. Second we describe a method for fusion of logical knowledge bases and complex distributions, knowledge bases including informal (i.e., free-text) information and distributions over values or structures of Bayesian networks. Third we describe a statistical keyword profiling for such hybrid annotated Bayesian network knowledge bases, which are formalized as elements of the knowledge base. Finally, we presents an ABN-based information retrieval language for an integration of knowledge engineering and data analysis.

The availability of electronic resources as prior knowledge with the advent of Bayesianism shall lead us to the concept of Bayesian knowledge engineering and knowledge intensive statistical data analysis (see Section 5.1). Because of the relevance of the informal prior information usually available as free-text, we developed the concept of an integrated knowledge base of formal and free-text (informal) prior knowledge organized around a causal Bayesian network of the domain. The concept of annotated Bayesian network knowledge base (ABN-KB) was introduced as the collection of free-text, though frequently structured description of objects related to the Bayesian network representation, such as sets of discrete values, variables and subgraphs. The purpose was multiple, though each can be related to Bayesian knowledge engineering. First, its goal is to support the collection of prior information from electronic textual resources including the collection of relevant papers and the discovery of relevant information from scientific domain literature (i.e., integrating information retrieval and knowledge engineering). Second, to support the formulation of complex, knowledge rich probabilistic sentences related to the Bayesian network representation (i.e., knowledge engineering and data analysis). Third, to represent subjective information from the experts relevant for formulating the Bayesian prior knowledge model (that are not necessarily related to the descriptions), that is to define priors over consistent combinations of network fragments.

The integrated knowledge base of factual, free-text prior knowledge embedding also posteriors over Bayesian networks allows statistical keyword profiling of the annotations of various objects. First, these profiles allow the exploration of the prior knowledge itself by direct browsing, by clustering of the profiles or by the visualization of the similarity of the profiles. Second, if the knowledge base is expanded with collection of domain publications, the keyword profiles allow the integrated exploration of the knowledge-model and the domain literature. The exploration methods are based on a simple representation, on standard statistical keyword profiles of the free-text annotations of various objects and of publications. First, we report the functions that allow the definition of various profiles, then their use for the exploration of the model and finally their use for the joint exploration of the model and publications, such as finding papers that are relevant for certain aspects of the model (i.e., contextual information retrieval by providing a personal and domain-specific context). In short, we report an ABN-based integrated knowledge modeling and information retrieval system to support the Bayesian knowledge engineering and knowledge-rich statistical data analysis. Another use of the profiles for extracting prior knowledge from the publications are reported in Chapter 6 under the title of statistical text mining with Bayesian networks.

5.1 Bayesian knowledge engineering

Besides the modularity issue discussed in Section 3.3, another factor behind the problems of the applicability of the "classical" knowledge engineering methodology is that its context, particularly the following two background assumptions, have changed drastically in the last ten years:

1. *Immediacy:* The domain knowledge is provided by domain expert(s) and moderate amount of domain literature. That is the prior is conceivable and manually formalizable by the knowledge engineer and there is no significant involvement of automated extraction and reformulation from electronic domain literature or existing knowledge bases.

2. *Data independence:* The goal of the knowledge engineering process is to produce and use a self-sufficient knowledge representation without statistical data. The inductive refinement of the knowledge base and the support of statistical data analysis are optional.

The current context of knowledge engineering of Bayesian network can be characterized with three additional features besides the modularity issue:

1. *Electronic vs. printed and expert domain knowledge.* The availability of the semantic web with electronic domain literature and knowledge bases contrary to the earlier case relying exclusively on experts and on printed domain literature.

2. *Statistical data vs. test cases.* The importance of the support of data analysis and the availability of a significant amount of statistical data for automated theory refinement contrary to earlier anecdotic test cases.

3. *Bayesianist vs. frequentist.* The availability of Bayesian methods by increased computational power offers a principled method for prior incorporation, theory refinement, and using a significant number of models.

These factors have redefined the knowledge engineering process for Bayesian networks in the following respects.

1. *Automated, meta knowledge engineering.* The knowledge engineering process has to provide methods for exploring and collecting the electronic domain knowledge, and for extracting or possibly discovering relevant domain knowledge in the electronic domain knowledge.

2. *Construction of priors.* A goal of knowledge representation is to formalize prior(s). The "final" (usually implicit) knowledge model is provided by the posterior of the Bayesian update.

3. *Interpretation and evaluation of posteriors.* Another goal of knowledge engineering is to provide a context for formulating knowledge rich statements with posteriors for evaluation and interpretation.

In short, because of the electronic domain knowledge and Bayesian methods, new goals for knowledge engineering are to specify (1) a Bayesian prior knowledge model, (2) indirectly over the electronic resources, (i.e., *meta-prior specification*) and (3) compute the posterior of complex, semantic statements. We call the knowledge engineering in this context *Bayesian knowledge engineering.*

The Bayesian conception of knowledge engineering as prior formulation, update and posterior analysis was envisioned in a seminal paper by Buntine [40] under the title of "Bayesian theory refinement". Similarly, the use of already existing knowledge bases in constructing Bayesian network were investigated in [255, 209, 84]. This view also fits to the knowledge intensive and "open" trend of Bayesian statistical data analysis [219, 32]. The usage of the electronic textual resources for prior formulation and extraction were reported in our works [22, 23, 20, 16, 25].

5.2 Probabilistic knowledge bases by embedded Bayesian networks

The Bayesian network $G, \underline{\theta}$ specifying $p(\underline{V}) = p(X_1, \ldots, X_n)$ can be conceived of a probabilistic propositional knowledge base KB over the domain variables V by interpreting the propositions $X_i = x_i$ as the corresponding random variables in the Bayesian network. Because the Bayesian network assign a probability to each atomistic event $\underline{x} = x_1, \ldots, x_n$, this induces a probability for any

well-formed sentence α over the domain propositions according to the rules of probability theory (as the expectation of its truth):

$$p(\underline{x} : \alpha(\underline{x})|KB) = \mathrm{E}_{p(\underline{x}|KB)}[\alpha(\underline{x})] = \sum_{\alpha(\underline{x}) \; is \; true} p(\underline{x}|KB). \tag{5.1}$$

Similarly, the conceptualization of the posterior $p(G|D)$ over the set of structures ($\mathcal{G}$) as a probabilistic knowledge base was proposed from the beginning of the field [40, 57].

The application of Bayesian networks as (1) "monolithic", (2) "propositional" and (3) "isolated" probabilistic knowledge base is severely restricted. First, the set of propositions or in other words the set of domain variables is fixed and it cannot be changed dynamically according to the domain (e.g., by duplicating a subset with known probabilistic relations). Second, there are no objects and relations, functions in propositional logic and the language does not support the formation of general statements. Third, the probabilistic knowledge base is separated from the free-text or semi-structured background information, contrary to first-order logic, in which it can be incorporated in a standard manner.

The "monolithic" restriction of Bayesian networks were addressed in the works on representing network "fragments" [167], object-oriented Bayesian networks [155, 205], probabilistic frame-based systems and relational probability models [156, 92] to represent complex distributions over dynamically changing set of domain variables.

The "isolated"restriction were addressed partly by the above mentioned works and also by the works on textually annotated Bayesian networks investigating various usage of semantically incorporated free-text or semi-structured background information text [22, 23, 20, 16, 25].

The "propositional" restriction was addressed in the works on the probabilistic first-order logic.In first-order logic this approach requires a distribution over possible worlds with interpretations (i.e., over models M containing potentially varying number of objects and predicate and functional relations between them, see Section 7.4). Related work can be grouped as research on probabilistic logics and on the generalization of Bayesian networks towards first-order logic (FOL) (for a recent overview see e.g. [62]). One of the early works in the first group attempted to combine logic and probability [121], which defines the probabilistic knowledge base from elementary probabilistic building blocks. The BLOG (Bayesian Logic) language and Markov logic networks are also members of the first-order probabilistic logic family [76, 187]. The concept of Relational Bayesian networks [145] is another possible approach.

Following the proposed possible world interpretation from [121], we specialize this general approach for the fusion of factual (free-text) and uncertain knowledge based on data [25, 21]. We restrict the knowledge base to a voluminous factual part consisting of established ontologies and papers from the domain and to an uncertain part defined by an arbitrary distribution over Bayesian network structures with fixed set of domain variables. This hybrid approach defines the

distribution over the models $p(M)$ by the combination of a logical knowledge base and a probabilistic model assuming their independence. The logical knowledge base KB^l describes the factual knowledge in the domain and defines the set of models $\mathcal{M}(KB^l) = \{M : KB^l \text{ is true in M}\}$. The probabilistic knowledge base KB^p expresses the remaining uncertain knowledge by defining a distribution over these models $p(M|M \in \mathcal{M}(KB^l))$. That is, the uncertain knowledge does not narrow further the set of models but weighs them.The probability of a sentence α is defined as the expectation of its truth in valid worlds.

$$p(M : 1(\alpha, M)|KB^l, KB^p) = \sum_{M \in \mathcal{M}(KB^l)} 1(\alpha, M)p(M|KB^p). \tag{5.2}$$

where $1(\alpha, M)$ denotes the α's truth-value in the model M. If the models vary only in a well-defined aspect such as a given object, this regularity can be used to define the distribution over the models based on a distribution over this aspect.

In practice the textual annotations for Bayesian network objects, such as values, variables and substructures, can be structured, possibly containing formalized, even numeric information. The prior knowledge base in the ovarian cancer domain constructed and used in the thesis includes a four-graded rating (high/medium/low/none) for the pairwise dependency relations and for the causal mechanisms (i.e., for the parental sets). Furthermore, the pairwise relations are annotated with monotonicity information (+/-), logical and causal information, four-graded rating, a derived probability, besides the optional free-text annotation (see Chapter 4 for details). Because of the multiple uses of textual annotations for Bayesian networks, we use the term of *Annotated Bayesian Network* (*ABN*) to encompass the enhanced functionalities of such minimally enriched Bayesian networks.

Definition 5.2.1. *A Probabilistic Annotated Bayesian Network Knowledge Base (pABN-KB) K for a fixed set $\underline{V}$ of discrete random variables is a first-order logical knowledge base using standard graph, string and BN related predicates, relations and functions. Let G represent a target DAG structure including all the target random variables. The knowledge base includes free-text descriptions for the subgraphs and for their subsets. We assume that the models of the knowledge base differs only w.r.t. G (i.e. there is a bijection $G \leftrightarrow M$) and the distribution $p(G)$ is available. For a well-formed sentence α, its probability is defined as the expectation of its truth*

$$\mathrm{E}_{p(M|K)}[1(\alpha, M)] = \sum_G 1(\alpha, M(G))p(G|K), \tag{5.3}$$

where $M(G)$ denotes the model defined by G.

This hybrid approach defines a distribution over the set of models $\mathcal{M}$ by combining a logical knowledge base with a probabilistic model. The logical knowledge base describes the factual knowledge in the domain defining a set of models (legal worlds) and the probabilistic part $p(G)$ expresses the uncertain

knowledge over these worlds. If the annotations in the knowledge base are compatible with a single Bayesian network model, then we use the term *Annotated Bayesian Network* . For the approximation of the expectation in Eq. 5.3 see Section 7.5.

Note that the logical knowledge base usually excludes a priori certain structures G, so only an unnormalized distribution is available. However, this is not a serious restriction, since $p(G)$ usually is an unnormalized posterior. Another problem is that typically not the most probable sentences are the most interesting(e.g., it is possible that a sentence α is a tautology or entailed by the factual knowledge base K, so it has probability 1). This led us to the formalization of the *most probable sentence subset selection* problem [189].

From a syntactic point of view, the model-based semantics can be reformulated as follows. The KB^l is extended with a set of predicates $\mathcal{S}_{KB^p}$ representing the uncertain part. The extension happens w.r.t. the distribution KB^p, so the probability of a sentence can be defined as the probability of its provability with a sound and complete inference method $\vdash$, approximated with a constrained theorem prover $\vdash_i$:

$$p(\alpha|KB^l, KB^p) \triangleq p(KB^l \cup \mathcal{S}_{KB^p} \vdash \alpha) \approx p(KB^l \cup \mathcal{S}_{KB^p} \vdash_i \alpha).$$

An ABN-KB can be formally represented using the formalisms of the object-oriented Bayesian network approach or the probabilistic frame-based system approach. In our case the emphasis is not on the formalism, but on the functionality of such textually enriched Bayesian network (or Bayesian network fragments) in knowledge engineering, in model evaluation and refinement by defining complex ABN-propositions incorporating textual background knowledge and in learning of Bayesian networks from domain literature and clinical data (see Section 8.4.1, 8.4.2).

5.3 Keyword profiles of ABN-KB objects

According to our assumption, an ABN-KB contains free-text annotations for subsets of the variables (e.g., the singular variables themselves), for subsets of the discrete values and for both directed and undirected subgraphs (see Def. 5.2). In this chapter we assume again that the ABN-KB defines target models as single Bayesian networks, a tree-like class hierarchy over the domain variables and possibly a distribution over the target model to ensure the model-based probabilistic semantics for ABN sentences as defined in Eq. 5.3. We will focus on the ABN-KB itself (i.e., on the logical part and not on the probabilistic extension), which follows our earlier formalization reported in [22, 23].

The keyword profiles for the annotations of ABN-KB objects and later for the publications are based on an algebraic representation, called the *vector space model*, which encodes a document in a vector space where each component represents a corresponding word in the vocabulary described in Section 4.3.4.3. This approach thus neglects the grammatical structure of the text. We used

the Porter stemmer to canonize the words [88], processed the essential domain-specific phrases and synonyms appropriately and applied a standard stopword list to remove general words [29, 161, 185]. The weights for the vector model were computed using the *term frequency–inverse document frequency* (tf-idf) term weighting scheme [29, 161, 185], but the raw term frequency and the boolean presence/absence representation is used as well. The weighted frequency of term t_j in document d_i is

$$w_{ij}^{\text{tf-idf}} = f_{ij} \log(\frac{L}{n_i}), \tag{5.4}$$

where f_{ij} is the number of occurrences of t_j in d_i, L is the total number of documents and n_i is the number of documents containing term i (in our largest MEDLINE corpus). If the text kernel of a domain variable contains references, then we used the linear combination of the vector representations of the literal annotation and of the references with user-specified weights λ_i for the sources (e.g., 0.1 for the corpus with medium rate and 0.5 for the IOTA protocol).

A standard similarity metric for a pair of documents d_i, d_j is the cosine of the angle between their normalized tf-idf vector representation $\underline{W}_i, \underline{W}_j$:

$$\text{sim}(d_i, d_j) \quad = \quad \cos(\underline{W}_i, \underline{W}_j). \tag{5.5}$$

Continuing the first-order logic (FOL) formalization of an ABN-KB described in Section 5.2, we expand it by introducing the following functions (in addition to the standard set of string functions and relation, we assume standard axioms of set theory and arithmetics specified in FOL).

Example 5.3.1. *1. Annotation(s^E)/$\mathcal{A}(s^E)$: the concatenation of the descriptions for the objects in the set s^E*

2. Index(s^A,type)/$\mathcal{I}(s^A)$: the vector representation of a text object s^A (or the set of vector representations of objects in the set s^A). The available vector representations are the boolean, raw frequency and the TF-IDF weights.

3. IndexOperation(s_1^I,s_2^I,type): the arithmetic combination of vector representations as average or the term-by-term multiplication, which can be used to select an interesting subset of terms (nulling the rest).

4. Similarity(s_1^I,s_2^I,type)/$\Delta(s_1^I, s_2^I, options)$: the similarity of the vector representations, where the similarity can be the cosine similarity in Eq. 5.5 or for Boolean representation a ratio of overlapping terms [185].

These functions allow various keyword profiles for wide range of ABN-KB objects and collections. Conceptually we can think of them as the following sequence: the sets of ABN entities, their annotations and their vector representation (i.e., their keyword profiles):

1. *ABN-(entity) set*: set composed by set operations over domain variables and classes. The properties of the Bayesian network and the ontology can be used in the set definitions (such as parents or Markov blanket of a given variable, descendants or ascendants of a class).

2. *Annotation set*: set composed of the annotations of ABN-(entity) sets and possibly directly specified free-text.

3. *Index set*: the corresponding set of vector representations of the members of an annotation set using a Boolean or TF-IDF scheme. It may be a certain combination of the vector representations of the members, such as the average, minimum or maximum value of the index weights. The result can be restricted to a set of keywords.

4. *Word set*: a directly specified set of keywords or the keywords of an index set with average weights above a specified threshold and the union, intersection, and difference of word sets.

In Section 5.5.2 we will show a formal language corresponding to these constructs to use ABN-based keyword profiles in information retrieval.

5.4 Explorations by keyword-based profiles

The profiles allow the exploration of the prior knowledge base itself by direct browsing, by clustering of the profiles or by the visualization of the similarity of the profiles.For a more refined usage of the profiles for model exploration we implemented functions to visualize and explore interactively the cosine similarity of the profiles of the variables defined in Eq. 5.4, 5.5. It offers the direct visualization as a network and it applies a hierarchical clustering with Ward linkage over the pairwise cosines. (see Figure 5.1).

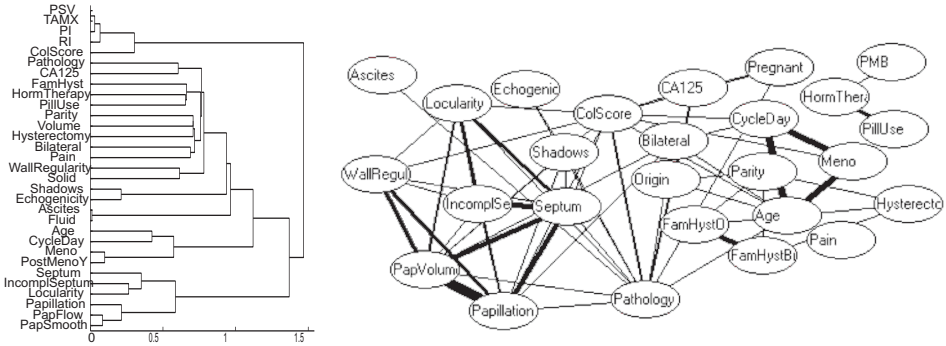


Figure 5.1: (Left) The hierarchical clustering of the variables based on the cosine of the *tf-idf* vector representation of the kernel documents of the variables. (Right) The similarity of the annotations of the variables using the pairwise cosines of the TF-IDF representations of the variables.

5.5 An ABN-based information retrieval language

Information retrieval (IR) deals with methods for indexing, searching, and recalling data, particularly text and other unstructured data forms [29, 161]. Two major trends leading to an increased efficiency of the information retrieval process are the utilization of user-specific information and domain-specific information. The purpose of both is to increase the convenience and the efficiency of expressing the information need of the user and to increase the quantitative performance of the retrieval.

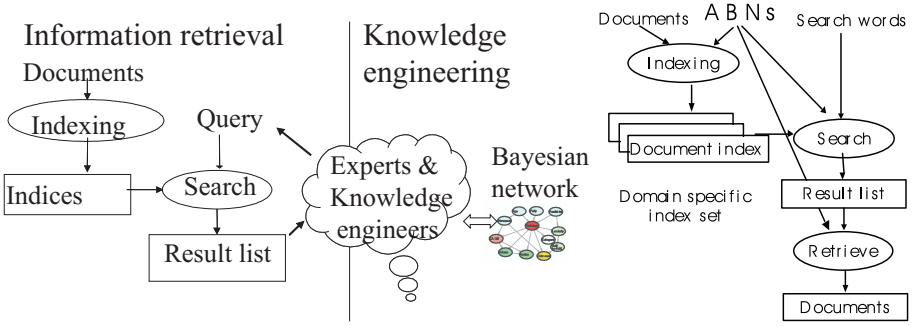


Figure 5.2: The current separated information retrieval (left) and the proposed annotated Bayesian network based information retrieval (right) to support the knowledge engineering and the learning of Bayesian networks. In the first case the knowledge engineer links the information resources and the built or learned BN model, whereas in the second case the annotated BN model can be incorporated directly in the information retrieval.

The ABN-KB can serve both as a probabilistic user profile and domain ontology, which allows another variant of the long suggested probabilistic expansion of the query [100, 61, 44, 38]. In fact, its use in the information retrieval process means its integration in the KE process . First consider the general case of having a distribution over the target DAGs and then the case of using only the logical part (the ABN-KB itself). In both cases, formally we assume that the publications are part of the ABN-KB as structured objects p_i whose textual parts are available by the $\mathcal{A}(p_i)/\mathcal{A}(p_i)$ function.

5.5.1 Informational relevance expressed by ABN sentences

If we have a probabilistic ABN-KB with a distribution over the structure of the target DAG G according to Def. 5.2, then the extension of the ABN-KB functions in Example. 5.3.1 and the publications allow integrated probabilistic queries incorporating informal knowledge and the formal model properties (again using the probabilistic semantics of the ABN sentences defined in Eq. 5.3). For example, the following sentence expresses the query that there is a publication p that is similar (i.e., relevant) to all the annotations of the variables in the

Markov blankets of a given variable Y (i.e., it covers the boundary of variable Y).

$$\forall X Publ(p) \wedge Var(Y) \wedge (X \in \mathrm{MB}(Y, G)) \Rightarrow (\Delta(\mathcal{I}(p), \mathcal{I}(\mathcal{A}(X))) < \delta)$$

5.5.2 An IR language for contextual relevance

Next, we consider the more practical case when we use only the logical part of the ABN-KB as a user-specific and domain-specific context. To increase the efficiency of information retrieval, particularly in knowledge modeling and model building, a special language was proposed for the detailed characterization of documents. It allows the definition of complex expressions based on the ABN (i.e., based on the domain model and the personal annotations), which are interpreted and evaluated as a relevance measure to select the matching documents by the information retrieval system. The ABN-based information retrieval language implements the functions listed in Example 5.3.1 by using the constructs listed in Section 5.3: ABN-(entity) sets, annotation sets, index sets, word sets. The implemented ABN-based information retrieval system was reported in [23], its language is summarized below (using the formalism of rewrite rules in grammars). The "Publication" non-terminal denotes the publications, $[]^!$ denotes the selection of exactly one alternative.

$$\begin{aligned}
\mathrm{S} &\rightarrow \mathrm{Relevance} \\
\mathrm{Relevance} &\rightarrow \Delta(\mathrm{Index_set}, \mathcal{I}(\mathrm{Publication})[\mathrm{cosine}|\mathrm{boolean}])|\mathrm{scalar}| \\
&\quad (\mathrm{Relevance}[+|*|\mathrm{min}|\mathrm{max}\ldots]\mathrm{Relevance}) \\
\mathrm{Index_set} &\rightarrow \mathcal{I}(\mathrm{Annotation_set})| \\
&\quad \mathrm{IndexOperation}(\mathrm{Index_set}, \mathrm{Index_set}, [\mathrm{min}|\mathrm{max}|\mathrm{average}|*|\mathrm{null}]) \\
\mathrm{Annotation_set} &\rightarrow \mathcal{A}(\mathrm{ABN_set})|\mathrm{string}|\mathrm{Annotation_set}[\cup|\cap|\setminus]^!\mathrm{Annotation_set} \\
\mathrm{ABN_set} &\rightarrow \mathrm{ABN_Variable}|\mathrm{ABN_Class}|(\mathrm{ABN_set}[\cup|\cap|\setminus]^!\mathrm{ABN_set})| \\
&\quad \mathrm{Parents}(\mathrm{ABN_set})|\mathrm{Children}(\mathrm{ABN_set})|\mathrm{Markov_Blanket}(\mathrm{ABN_set}) \\
\mathrm{ABN_Class} &\rightarrow \mathrm{c}_0|..|\mathrm{c}_\mathrm{L} \\
\mathrm{ABN_Variable} &\rightarrow \mathrm{v}_0|..|\mathrm{v}_\mathrm{n}
\end{aligned} \tag{5.6}$$

Quantitative evaluation of certain aspects of the ABN-based information retrieval language were reported in [23].

Chapter 6

Text mining with BNs

We discuss the application of Bayesian networks in statistical analysis of free-text publications, which offers a generative, publication model-based method. We discuss conditions for this method, its causal interpretation, and its complementarity to currently prevailing bottom-up, manually supported extraction methods.

Rapid accumulation of biological data and the corresponding knowledge poses a new challenge of making this voluminous, uncertain and frequently inconsistent knowledge accessible. Despite recent trends to broaden the scope of formal knowledge bases in biomedical domains, free-text electronic literature is still the central repository of the domain knowledge. This central role will probably be retained in the near future. The extraction of explicitly stated knowledge or the discovery of implicitly present latent knowledge requires various techniques ranging from purely linguistic approaches to machine learning methods. In this chapter we investigate a domain-model based approach to statistical inference about dependence and causal relations given the literature using minimal linguistic preprocessing. We use Bayesian Networks (BNs) as causal domain models to introduce generative models of publication (i.e., we examine the relation of domain models and generative models of the corresponding literature).

In a wider sense our work provides support to statistical inference about the structure of the domain model. In Chapter 8 we present a unified view of the literature and the data, but one of the attractive alternatives is a two-step approach, which consists of the reconstruction of the beliefs in mechanisms from the literature by model learning and their usage in a subsequent learning phase. Here, the Bayesian framework is an obvious choice. Earlier applications of text mining provided results for the domain experts or data analysts, whereas our aim is to go one step further and use the results directly in the statistical learning of the domain models. The first step consists of reconstructing collective beliefs from the literature as parameters of generative models. Actually it can be conceived as an a posteriori belief given the "literature data" (see Sections 6.1 and 8.1). In the second phase the Bayesian inference about the posteriors of structural properties of the domain model given the clinical or biological data is

the practical choice. Finally the link between these two steps can be formalized using the principled probabilistic semantics (i.e., our goal is to provide the a priori probabilities on the structural properties of the domain model derived from the literature, see Fig. 1.3).

The central assumption of our work is that causal relations (mechanisms, see Section 3.1.3.3) are important factors influencing most of biomedical publications. The explicitly known or implicitly reported mechanisms exert their effects as building blocks in generative models of the occurrences of domain entities in publications. Fig. 1.3 illustrates our assumptions about (1) the mechanism uncertainty in the domain, (2) the corresponding literature data, (3) the reconstructed generative probabilistic model, and (4) the application of reconstructed mechanism uncertainty as prior in statistical inferences about domain models.

The chapter is organized as follows. In Section 6.1 we define an algebraic representation of the literature. In Section 6.2 we review the types of uncertainties in biomedical domains from a causal, mechanism-oriented point of view. In Section 6.3 we summarize recent approaches to information extraction and literature mining based on natural language processing (NLP) and "local" analysis of occurrence patterns. In Section 6.4 we propose generative probabilistic models for the occurrences of biomedical concepts in scientific papers. Section 6.1 presents textual aspects of the OC domain. In Chapter 8 we present a unified view of the literature, the data and their models and report results on learning BNs given the literature.

6.1 The literature data

As the foundation of the followed statistical text-mining approach, we introduce a vector representation of the text. According to our assumption, for each domain variable X_i a name and its synonyms and a text kernel is available (see Section 4.3.4). We denote the occurrence of the name (and synonyms) of an ABN variable X_j in document d_i with a binary x_{ij}^O value. $\underline{\underline{D}}^{C^O}$ denotes the complete matrix for a given document corpus C. This matrix will be used in the name co-occurrence methods. Note that this co-occurrence representation cannot handle repetition and proximity or separation into distinct paragraphs, sentence, and so on; but in our experiments this scheme gave satisfactory performance (for the comparison of such options, see [74]).

We define another binary representation of MEDLINE abstracts based on the kernel documents using the TF-IDF vector representation and the cosine similarity in Eq. 5.4, 5.5. It consists of binary values defined as

$$x_{ij}^R = \begin{cases} 1 \text{ if } \tau < \text{sim}(k_j, d_i) \\ 0 \text{ else} \end{cases}, \tag{6.1}$$

which expresses the relevance of kernel document k_j to document d_i. We will use an experimentally selected fixed value for τ ($\tau = 0.1$). $\underline{\underline{D}}^{C^R}$ denotes the corresponding matrix for a given corpus C. For later references we introduce the following concept.

Definition 6.1.1. *The term literature data set (D^L, D^O, D^R) denotes the binary representation of the occurrence or relevance of predefined concepts in publications in a given corpus (the corresponding binary random variables of domain variables X_i are denoted respectively with X_i^L,X_i^O, or X_i^R).*

Source document collections are described in Section 4.3.4.

6.2 Concepts, associations, and causation

Frequently a biomedical domain can be characterized by a dominant type of uncertainty w.r.t the causal mechanisms. These types of uncertainty show certain sequentiality as described below. This sequence is related to the development of biomedical knowledge, even though a strictly sequential view is clearly an oversimplification.

(1) *Conceptual phase*: Uncertainty over the domain ontology (i.e., what are the relevant entities?).

(2) *Associative phase*: Uncertainty over the association of entities. Indirect, associative hypotheses, or frequently associated entities are reported in this phase. Though we accept the general view of causal relations behind associations, we assume that the exact causal functions and direct relations are unknown.

(3) *Causal relevance phase*: Uncertainty over the existence of causal relations (i.e., over mechanisms). Typically, direct causal relations are reported as processes and mechanisms.

(4) *Causal effect phase*: Uncertainty over the strength of the autonomous mechanisms embodying the causal relations.

We assume that the domain is already in the associative or causal phase (i.e., that the entities are more or less agreed upon, but that their causal relations are mostly in the discovery phase). This assumption holds true in many biomedical domains, particularly in those linking the biological and clinical levels. There the associative phase is a crucial but lengthy process of knowledge accumulation, where a wide range of research methods is used to report associated pairs or clusters of the domain entities (i.e., transitive closures of partially observed causal relations).

6.3 Literature mining

Literature mining methods can be classified into bottom-up — usually pairwise — and top-down (model based) methods. Bottom-up methods assume that the domain is at least partially in a causal phase and attempt to identify individual relations leaving the integration to the domain expert. The corresponding linguistic approaches assume that the individual relations are sufficiently known, formulated and reported for automated detection methods. On the contrary, top-down methods assumes only that the domain is in an associative phase. That is they assume that mainly causally associated entities are reported with

or without tentative relations and direct structural knowledge. Their linguistic formulation is highly variable, not conforming to simple grammatical characterization. Consequently top-down methods typically use agrammatical text representations and minimal linguistic support. To compensate, they concentrate on identifying consistent domain models by analyzing jointly the domain literature, which autonomously prune redundant, inconsistent, indirect relations by evaluating consistent domain models.

Until recently mainly bottom-up methods have been analyzed in the literature: *linguistic approaches* extract explicitly stated relations, possibly with qualitative ratings [208, 135]; *co-occurrence analysis* quantifies the pairwise relations of variables by their relative frequency [230, 147]; *kernel similarity analysis* uses the textual descriptions or the occurrence patterns of variables in publications to quantify their relation [221]; Swanson and Smalheiser [232] discover relationships through the heuristic pattern analysis of citations and co-occurrences; in [54] and [184] local constraints were applied to cope with possible hidden confounders, to support the discovery of causal relations; *joint statistical analysis* in [162] fits a generative model to the temporal pattern of corroborations, refutations and citations of individual relations to identify "true" statements.

The top-down method of the *joint statistical analysis* of de Campos [71] learns a restricted BN thesaurus from the occurrence patterns of words in the literature. Our approach is closest to this and those of Krauthammer et al. and Mani [162, 184].

The reconstruction of informative priors over domain mechanisms or models from research papers is further complicated by the *multiple aspects of uncertainty* about the existence, scope (conditions of validity), strength, causality (direction), robustness for perturbation and relevance of mechanism and the *incompleteness of reported relations*, because they are assumed to be well-known parts of *common sense* knowledge or of the already reported *paradigmatic* knowledge of the community.

6.4 BN models of publications

Considering biomedical abstracts, we adopt the central role of causal understanding and explanation in scientific research and publication [234]. According to this causal stance, we assume that the function of an occurrence of a domain concept (i.e., variable) is "explained" (explanandum) or "explanatory" (explanans), in addition, we allow the "described" status. This implicitly means that we assume that publications contain either description of the domain concepts without considering their relations or the occurrences of entities participating in known or latent causal relations.

Furthermore, we assume that mainly positive statements are reported and we treat negation and refutation as noise. We assume that exclusive hypotheses are reported for a given variable (i.e., we treat alternatives as one aggregated hypothesis) and that there is only one causal mechanism for each set of causes (i.e., we will equate a given set of causes and the mechanism based on it). Ad-

ditionally, we presume that most publications are causally ("forward") oriented (i.e., explanations mostly follow causal and not diagnostic line of reasoning). We attempt to model the transitive nature of causal explanation over mechanisms (e.g., that causal mechanisms with a common cause or with a common effect are surveyed in an article, or that successive causal mechanisms are tracked to demonstrate a causal chain). By contrast, we also have to model the lack of transitivity (i.e., the incompleteness of causal explanations, that is that certain variables are assumed as explanatory, others as potentially explained, except for survey articles that describe an overall domain model). Finally, we assume that the reports of the causal mechanisms and the univariate descriptions are independent of each other.

First, we experimented with a two-layer Bayesian network. The upper-layer variables represent the pragmatic functions (i.e., the intentions of the authors or the property of the given experimental technique), the lower-layer variables represent their observable occurrences. We assumed that lower-layer variables are influenced only by the upper-layer ones denoting the corresponding mechanisms, and not by any other external quantities (e.g., by the number of the reported entities in the paper). A further assumption is that the belief in a compound mechanism is the product of the beliefs in the pairwise dependencies. Consequently we use noisy-OR canonic distributions for the children in the lower layer [200]. This model cannot represent the dependencies between the reported associations, and its performance was not satisfactory.

To devise a more advanced model, we relax the assumption of the independence between the variables in the upper layer representing the pragmatic functions, and we adapt the models to the vector representation of publications (see Section 6.1). Consequently we analyze the possible pragmatic functions corresponding to the domain variables, which could be represented by hidden variables. We assume here that the explanatory roles of a variable are not differentiated (the "uniform transitivity" assumption), and that if a variable is explained (or described), then it can be explanatory for any other variable (the "full transitivity" assumption). We assume also the "full transparancy" of causal relevance (i.e., that the lack of occurrence of an entity in a paper means causal irrelevance w.r.t. the mechanisms and variables in the paper and not a neutral omission). These assumptions allow the merging of the explanatory, explained and described status with the observable reported status (i.e., we can represent the hidden and observed pairs jointly with a single binary variable). Fig. 6.1 shows these steps. Note that these assumptions remain tenable in case of report of experiments, where the pattern of relevances has a transitive, causal foundation.

Definition 6.4.1. *Let us assume that in a domain with variables $X_1, \ldots, X_n$ a research community accepts that a causal model G satisfies the Causal Markov Condition (see Def. 3.1.18). Let $X_1^L, \ldots, X_n^L$ denote the binary, random variables in the literature data representation of a corresponding document corpus C (see Def. 6.1.1). The "forward, transitive, transparent, causal" (FTTC) publication condition holds if G statisfies the Causal Markov Condition w.r.t.*

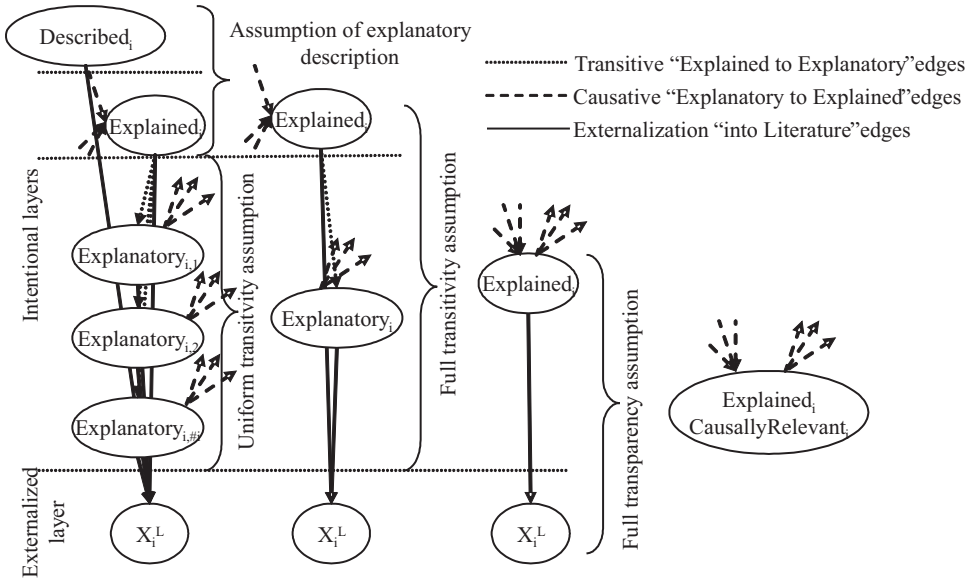


Figure 6.1: The derivation of the transitive publication model by the assumptions of "uniform transitivity", "full transitivity", and "full transparancy".

$p(X_1^L, \ldots, X_n^L)$ as well. We call the Bayesian network models of $p(X_1^L, \ldots, X_n^L)$ (binary, transitive, fully transparent) literature Bayesian networks $(G^L, \underline{\theta}^L)$.

A possible semantics for the parameters of a binary, transitive, fully transparent literature BN can be derived from a causal point of view that the presence of an entity X_i is influenced only by the presence of its potential explanatory entities (i.e., its parents). Consequently, $p(X_i = 1|Pa(X_i) = \mathrm{pa}_{x_i})$ can be interpreted as the belief that the parental variables with "present" status in pa_{x_i} can explain the entity X_i ($Pa(X_i)$ denotes the parents of X_i and $Pa(X_i) \rightarrow X_i$ denotes the parental substructure). In that way the parameters of a complete network can represent the priors for parental sets compatible with the implied ordering:

$$p(X_i = 1|Pa(X_i) = pa(X_i)) = P(Pa(X_i) = pa(X_i)) \tag{6.2}$$

where for notational simplicity $\mathrm{pa}(X_i)$ denotes both the parental set and a corresponding binary representation. This would imply that we can model only full survey papers, but the general, unconstrained multinomial dependency model used in the transitive BNs provides enough freedom to avoid this limitation.

The multinomial model allows entity specific modifications at each node, combined into the parameters of the conditional probability model, which are independent of other variables (i.e., unstructured noise). This permits the modeling of the description of the entities, the beginning of the transitive scheme of causal explanation, and the reverse effect of interrupting the transitive scheme (i.e., incorporating the probability of acausal description, and starting and ter-

minating a chain of causal explanation). Note that a "backward" model corresponding to an effect-to-cause or diagnostic interpretation and explanation method has a different structure with opposite edge directions.

In the Bayesian framework, there is a structural uncertainty also (i.e., uncertainty over the structure of the generative models themselves). So to compute the probability of a parental set $Pa(X_i) = pa(X_i)$ given a literature data set $D^L_{N'}$, we have to average over the structures using the posterior given the literature data:

$$\begin{aligned} & P(Pa(X_i) = pa(X_i)|D^L_{N'}) && (6.3)\\ &= \sum_{(pa(X_i)\to X_i)\sim G^L} p(X_i = 1|pa(X_i), G^L)P(G^L|D^L_{N'}) \\ &\approx \sum_{G^L} 1((pa(X_i) \to X_i) \sim G^L)P(G^L|D^L_{N'}) && (6.4)\end{aligned}$$

Consequently, the result of learning BNs from the literature can be multiple (e.g., using a maximum a posteriori structure and the corresponding parameters, or the posterior over the structures, see Eq. 6.3). In the first case, the parameters can be interpreted structurally and converted into a prior for a subsequent learning. In the latter case, we neglect the parametric information focusing on the structural constraints, and transform the posterior over the literature network structures into a prior over the structures of the real-world BNs (see Section 8.1).

6.5 Local scores for pairwise relationships

We use the following local (i.e., non-domain model based) score for pairwise relationships, which are the simplest approaches in statistical text mining (see Section 6.3 for the comparison of such statistical bottom-up methods).

Let $p(X^O_1, \ldots, X^O_n)$ denote the joint probability of occurrence of the names or synonyms of ABN random variables in a paper from a given corpus. For the kernel relevance, let $p(X^R_1, \ldots, X^R_L)$ denote the joint probability of the relevance of the kernels of the random variables in the ABN for a certain document. Based on the previous definitions, we can define several text scores to quantify the dependency of the pairs of random variables in the ABN. $R^{\mathrm{AND},C_i}_{\mathrm{COOC}}(X;Y)$ and $R^{\mathrm{AND},C_i}_{\mathrm{COREL}}(X;Y)$ denote a name co-occurrence and a kernel corelevance score. $R^{\mathrm{MI},C_i}_{\mathrm{COOC}}(X;Y)$ and $R^{\mathrm{MI},C_i}_{\mathrm{COREL}}(X;Y)$ denote the mutual information scores based on name occurrence and kernel relevance over the collection C_i.

$$R^{\mathrm{AND}}_{\mathrm{COOC}}(X;Y) \triangleq p(X^O = 1, Y^O = 1|(X^O = 1) \vee (Y^O = 1)) \quad (6.5)$$
$$R^{\mathrm{AND}}_{\mathrm{COREL}}(X;Y) \triangleq p(X^R = 1, Y^R = 1|(X^R = 1) \vee (Y^R = 1)) \quad (6.6)$$
$$R^{\mathrm{MI}}_{\mathrm{COOC}}(X;Y) \triangleq I(X^O;Y^O) \quad (6.7)$$
$$R^{\mathrm{MI}}_{\mathrm{COREL}}(X;Y) \triangleq I(X^R;Y^R) \quad (6.8)$$

Additionally, we introduce a relevance scoring for X and Y inspired by information retrieval. A standard similarity metric for the kernel descriptions of K_X and K_Y is the cosine of the angle between their corresponding normalized tf-idf vector representation as defined in Eq. 5.5. The definition is the following:

$$R_{\text{ASIM}}(X;Y) \triangleq \text{sim}(K_X, K_Y).$$

We refer to these text-based local relevance scores in general with $R^L_{\text{Text}}(X;Y)$.

6.6 Results

In the main application domain of the thesis, in ovarian cancer substantial prior knowledge and clinical data is available, which allows wide range of evaluations of the BN publication models. Such cross-comparison of expertise, clinical data based statistical inferences and literature data based statistical inferences are reported in Chapter 8 using the transitive, fully transparent BN model with a causal (forward) interpretation. Results w.r.t. complete models are reported in Section 8.4.2, (sequential) posteriors of simple pairwise structural are in Section 8.5.2.1 and posteriors of complex structural features such as MBG(Y) and MB(Y) relations are reported in Section 8.5.2.2. Anticipating these quantitative evaluations Fig. 6.2 shows a literature Bayesian network.

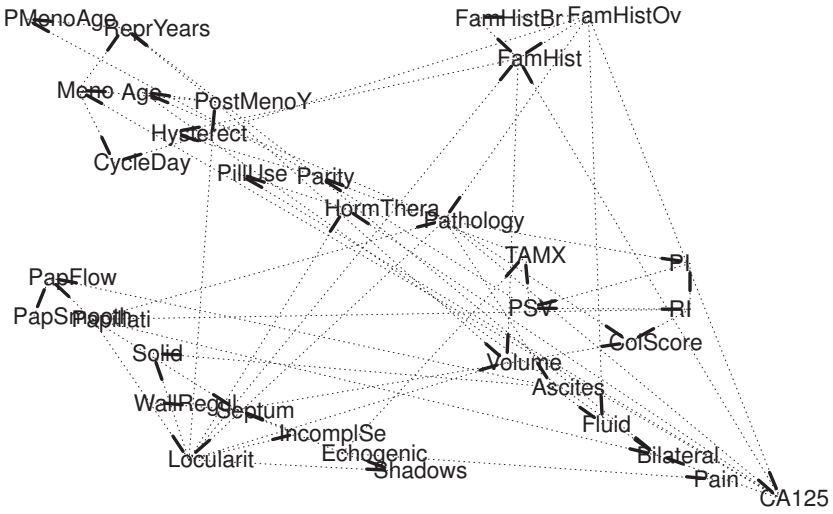


Figure 6.2: The maximum a posteriori Bayesian network using the $D^{PM^H_R}$ data set, the BD_{eu} parameter priors and noninformative structure priors and exhaustive search to 3 parents with K2 greedy continuation over 10^6 random ordering.

Chapter 7

Inference over BN features

First we categorize structural properties (i.e., features) of Bayesian networks, and introduce a feature called Markov blanket graph. Second we summarize the advantages of the Bayesian approach to BN features, and formalize the applicability and the statistical advantages of the ordering-based MCMC estimation method. Third we discuss the consequences of the exponential cardinality of feature values for decisions based on their MC estimates. Finally, the integration of estimation and search of high-scoring MBG feature values is analyzed.

The increasing complexity of the models, the incorporated prior knowledge and the queries leads to the issue of Bayesian inference over general properties of Bayesian networks (i.e., to estimation of the expectation of binary random variables). Although we discuss this problem from the point of inference over structural features, note that the expectation of functions over the space of DAGs w.r.t. a posterior appears in a wide range of problems, such as in the posterior of a feature (i.e., structural model property) F_c, in the posterior of an ABN sentence (see Def. 5.2), in the expected loss of the selection of a given model and in the full-scale Bayesian inference over domain values (see Eq. 3.25):

$$p(F_c = f_c|D_N) = \sum_G 1(F_c(G) = f_c)p(G|D_N) \tag{7.1}$$

$$p(\alpha(G)|\mathcal{K}, D_N) = \sum_{M(G)\in\mathcal{M}(\mathcal{K})} \alpha(M(G))p(G|D_N) \tag{7.2}$$

$$L_{\hat{G}|D_N} = \mathrm{E}_{p(G|D_N)}[L(G,\hat{G})] = \sum_G L(G,\hat{G})p(G|D_N), \tag{7.3}$$

$$p(\underline{y}|\underline{x}, D_N) = \mathrm{E}_{p(G|D_N)}[\mathrm{E}_{p(\underline{\Theta}|G,D_N)}[p(\underline{y}|\underline{x},\underline{\Theta},G)]]. \tag{7.4}$$

First, we overview Bayesian network features in Section 7.1 and introduce the Markov blanket subgraph feature in Section 7.2. In Section 7.3 and 7.4 we discuss the advantages of feature posteriors as confidence measures w.r.t. the bootstrap probabilities. In Section 7.5 we will concentrate on the approximation of Eq. 7.1, when the feature is a standard graph-theoretic property of DAG G

with values $F(G) = f_i$, $i = 1, \ldots K$. The growing importance of such model-based, feature-oriented statistical inferences is the result of (1) frequent high sample complexity for the identification of the complete model, (2) the lack of prior for the complete model, (3) the high computational complexity for the identification of the complete model, (4) the availability of computational resources and stochastic methods for estimation, and (5) the availability of complex semantic propositions with statistical semantics as the ABN sentences in Eq. 7.2.

The most important factor is the relatively small amount of data. A general expectation is that, in case of small amount of data, at least certain properties with high significance of a complex model can be inferred and perhaps with lower computational cost. So the goal is the automated learning of what is learnable with high confidence in the considered model space given the data and to support the interpretation of statistical inference by indicating confidence measures for such properties. Furthermore, the model properties with high significance can be used heuristically as "hard" constraints or "soft" bias to support the inference of the complete model, either by influencing it through priors in learning from heterogeneous sources or in the case of using the same data set by influencing the optimization process itself (see Chapter 8). Note the similarity of this approach to the frequentist constraint-based Bayesian network learning methods, which perform hypothesis tests on local model properties (on features) and integrate them into a consistent domain model. In a potential Bayesian analog the hypothesis tests are replaced by the model-based feature posteriors instead of the significance levels and p-values of hypothesis tests, enhancing their integration in subsequent phases of learning a complete domain model.

However, the Bayesian approach to feature learning has many additional aspects beside the estimation of the posterior. Such related issues are the effect of the cardinality of feature values on the selection of optimal value(s) and the integration of estimation and search processes in case of high numbers of features, which are discussed in Section 7.6 and 7.7. Additional issues related to classification in our case are the support of full scale Bayesian inference over domain values (i.e., the use of the estimated posterior distribution over the features as a probabilistic knowledge base) and the transformation or inducement of priors for a subsequent learning phase either using Bayesian networks or using other more specialized representations, for example logistic regression or multilayer perceptrons. These are discussed in Chapter 10.

Whereas these inferences are investigated mainly in fundamental research, they may soon appear in standard statistical data analysis software and in decision support systems as they can offer a more personalized and knowledge intensive environment for inductive inferences. For example, the combination of the electronic clinical and genomic patient data, the semantic web and evidence-based medicine can be driving force for such complex probabilistic queries over standardized knowledge bases and data-bases. A special case is the area of statistical analysis of biomedical literature, where we can treat the domain literature as a special data set and formulate queries against this voluminous

knowledge base (see Chapter 8). In general, it means that the knowledge intensive Bayesian approach over large, distributed knowledge and data-bases will get more and more emphasis within the area of knowledge and data analysis.

7.1 Bayesian network features

Before considering the induction of confidence measures over a Bayesian network feature F, first we overview standard Bayesian network features, together with proposed identification methods and the corresponding Bayesian tasks.

There is a large variety of features (i.e., model properties) to provide an overall or specialized characterization of the underlying model, such as the undirected edges or compelled edges (as direct relations or direct causal relations under CMA), pairwise or partial ancestral ordering (related to causal ordering), the parental sets, the pairwise relevance relations, the subset relevance relations (Markov blankets) or the partially parametric features such as the pairwise qualitative features. Despite this variety and the presence of the parental set features, which are the ultimate building blocks of Bayesian networks, the usefulness of these features are still seriously restricted by their unexplored dependency in all application areas, such as in data analysis, in probabilistic knowledge bases, in prior acquisition and in posterior-to-prior inducement for later phases of Bayesian learning. This seems to be unavoidable because even small sets of simple local features quickly become dependent, because of the DAG constraint, what biases this model-based approach with hardly estimatable effects.

A possible solution is the definition of complex features (subtheories) that are *sufficient* features for a given aspect of the domain theory and still more efficiently learnable than the complete domain model. So, it is an open issue to define complex features that on the one hand exactly model a semantically interesting fragment (subtheory) of the domain and on the other hand they are still considerable simpler than the complete domain model. Such a feature would exactly represent the interesting dependencies between the relevant simpler features and the statistical and computational complexity of the estimation of its distribution over the feature space would be lower and better interpretable.

In fact, we can define two approaches to Bayesian network features. The first approach relies on the assumption that the feature set is fixed, the features are significantly simpler than the complete domain model, though they provide an overall characterization as a fragmentary representation, and the number of features and feature values are tractable (not exponential, but linear or quadratic in the number of variables). Such features are the pairwise edge or relevance features (i.e., the compelled edges and Markov blanket relations). These simple features are easily interpretable or can be used to support a subsequent learning phase of a complete Bayesian network model. The main challenge in this approach is the computation of the corresponding expectations.

At the other extreme of feature learning we find the identification of arbitrary subgraphs with statistical significance, which is an idealistically autonomous ap-

proach to feature learning consisting of a mixture of search and the computation of the achieved significance. This is close to our approach to Bayesian network features investigated in the thesis, but we restrict the subgraphs to Markov blanket subgraphs to have a focused representation from a single, but complex point of view (i.e., from conditional modeling) and we use the Bayesian framework instead of the frequentist framework.

7.1.1 Edges: direct pairwise dependencies

The first family of frequentist algorithms for learning a Bayesian network feature targets the identification of "direct" (unconditional) causal pairwise relations ("direct" in the sense discussed in Section 3.1.3.2). If the hypotheses are the DAGs as causal models, then this feature corresponds to the edges. If the hypotheses are the observational equivalence classes as independence models, then such relations are exactly identified by the compelled edges assuming no hidden variables, the causal Markov condition and stability. The corresponding posteriors in the Bayesian context are the following

$$p(X_i \to_G X_j | D_N) = \sum_G 1(X_i \to_G X_j) p(G|D_N) \tag{7.5}$$

$$p(CompE(X_i, X_j|G)|D_N) = \sum_G CompE(X_i, X_j|G) p(G|D_N). \tag{7.6}$$

In the presence of possible hidden variables there are more advanced constraint-based algorithms for identifying relations with various causal interpretations, though not in the Bayesian framework (see [202, 116], [54, 224]).For the application of bootstrap and Bayesian method over edge features, see Section 7.3 and 7.5.2.3.

7.1.2 Ordering of the variables

Whereas the identification of the ordering of the variables rarely appears as a direct target, indirectly it is usually present in BN learning. In the acausal approach the identification of an acausal Bayesian network heavily influenced by the identification of a good ordering of the variables, because the learning of an acausal Bayesian network structure for a given ordering is computationally efficiently doable (both in the frequentist or Bayesian framework). In the causal approach when the hypotheses are the DAGs, the causal structures directly define causal orderings as ancestral orderings. Consequently a score for a Bayesian network G can be interpreted as an approximate scores for the underlying partial orderings. Recall that the ML structure score can be interpreted as the summed mutual information between the parents-child pairs and that the BD and the BIC scores are asymptotically equivalent (see Section 3.5.1). So, in a broad sense, any structure learning can be interpreted as an indirect learning of orderings, but certain algorithms explicitly use orderings as a central representation. For example, the use of genetic algorithms has been reported to

find the best ordering for the learning of Bayesian network structures [166]. The corresponding posterior over the complete orderings $\prec$ in the Bayesian context is the following

$$p(\prec |D_N) \quad = \quad \sum_G 1(G \in \mathcal{G}^{\prec}) p(G|D_N). \tag{7.7}$$

7.1.3 Relevant variables

The concept of relevance is a fundamental concept in the definitions of the Bayesian network representation (see Def. 3.1 and 3.3 for the observational and causal relevance), but it is also central to AI, to decision theory (e.g., the value of further information) and to statistics (for an overview, see [231]). An important special case is the relevance of explanatory variables to predict a target variable given a data set, hopefully with a domain-specific interpretation. The selection of the relevant variables in this context is called the *feature subset selection* (FSS) problem, which is part of the broader problem of input preprocessing, construction of variables (e.g., interaction terms) and dimensionality reduction. We will discuss only the relation of the FSS problem to BN feature learning. Note that even in the conditional approach in general the features are not independent, so the concept of relevance corresponds to the subsets and not to the individual features.

To explain the generality of the Bayesian approach to relevance using Bayesian network features, we summarize the most widespread conditional approaches to FSS in sequence (see Section 9.1 for the conditional Bayesian modeling). We start with the concept of relevance and with a non-Bayesian approach specific to the applied optimization algorithm, the data set, the model class, and the loss function. Then we generalize these specifics step by step, which leads to a standard conditional probabilistic concept of relevance in the end. Finally, we relate the Bayesian conditional approach to the general Bayesian approach, particularly to the Bayesian inference over Bayesian network features. In short, we show that the Bayesian inference over Bayesian network features offers an algorithm-free, model-free*, loss-free and non-conditional (i.e., domain model based) solution for the feature subset selection problem.

The *conditional approach* to FSS relies on the separate modeling of the dependence of a target variable Y on $\underline{X}'$ (i.e., without modeling the overall domain). It has been investigated using various conditional model classes M, such as linear regression, decision trees, logistic regression, multilayer perceptrons or support vector machines [138, 36, 125, 73]. It defines a score function $S^S(\underline{X}', D_N, M, L)$ for the subsets $\underline{X}' \subseteq \underline{X}$ and performs a search in the space of subsets of the features.

The wrapper approach to feature selection uses an optimization algorithm $\hat{f}_{\mathcal{C}}(\underline{X}') = \mathcal{C}(\underline{X}', D_N, M, L)$ [148, 153]. It defines the score function as

$$S^S(\underline{X}', D_N, M, L) = S^F(\hat{f}_{\mathcal{C}}(\underline{X}'), D_N, M, L).$$

*In the assumed case of discrete variables with multinomial conditionals.

The conditional model score $S^F(\hat{f}_C(\underline{X}'), D_N, M, L)$ may incorporate factors for the interpretability or complexity of the conditional models $f(\underline{X}') \in M^{\underline{X}'}$ and their estimated expected predictive loss (risk).

In an algorithm-free and asymptotic case the subset score $S^S(\underline{X}', M, L)$ can be defined as the best expected predictive loss in a conditional model class $M^{\underline{X}'}$ with features $\underline{X}'$

$$S^S(\underline{X}', M, L) = \arg \min_{f(\underline{X}') \in M^{\underline{X}'}} \int L(y, f(\underline{x}'))p(y|\underline{x}') \, \mathrm{d}yp(\underline{x}') \, \mathrm{d}\underline{x}'. \tag{7.8}$$

However, this asymptotic and algorithm-free optimality of a subset for a given model class is not appropriate to define the relevance of a feature, as it was demonstrated in [148, 153].

The model-free subset score $S^S(\underline{X}', L)$ can be defined as the best achievable risk with subset $\underline{X}'$ for a given loss L, called Bayes risk

$$R^*_L = \int L(y, g^*(\underline{x}'))p(y|\underline{x}') \, \mathrm{d}yp(\underline{x}') \, \mathrm{d}\underline{x}', \tag{7.9}$$

where g^* is the Bayes decision, which minimizes the expected loss of prediction for each x (see Section 2.2.4.2).

Because of the specific choice of the loss function $L(Y, \hat{Y})$, it is still possible that the minimal subset would miss certain features relevant for another loss. The following theorem for the case of binary output Y shows that the final loss-free generalization of the concept of relevance necessarily leads to the standard conditional probabilistic definition of relevance [73].

Theorem 7.1.1 ([73]). *A transformation $T(\underline{X}')$ is a mapping from the feature space $\mathcal{R}^n$ to $\mathcal{R}^{n'}$ and its Bayes risk with loss L is denoted with $R^*_{L,T}$. It is called admissible if for any loss function L, $R^*_{L,T} = R^*_L$, where R^*_L is the original Bayes risk. A transformation is admissible, if $T(\underline{X}')$ is a sufficient statistics (i.e., $p(Y|T(\underline{X}'), \underline{X}') = p(Y|T(\underline{X}'))$).*

The *relevance* of a feature can be defined in an algorithm-free, asymptotic, model-free and loss-free way as follows.

Definition 7.1.1. *A feature X_i is strongly relevant, if there exists some x_i, y and $s_i = x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n$ for which $p(x_i, s_i) > 0$ such that $p(y|x_i, s_i) \neq p(y|s_i)$. A feature X_i is weakly relevant, if it is not strongly relevant, and there exists a subset of features S'_i of S_i for which there exists some x_i, y and s'_i for which $p(x_i, s'_i) > 0$ such that $p(y|x_i, s'_i) \neq p(y|s'_i)$. A feature is relevant, if it is either weakly or strongly relevant; otherwise it is irrelevant [148, 153].*

The model-free, algorithm-free and loss-free conditional approach is called *filter approach* (for references, see [148, 153]). In the filter approach to feature selection we have to select a minimal subset $\underline{X}'$ that fully determines the conditional distribution of the target ($p(Y|\underline{X}') = p(Y|\underline{X}')$) without modeling the complete domain $p(Y, \underline{X}')$ or the explanatory variables $p(\underline{X}')$.

The Bayesian networks as representation of the independencies in the domain motivated a series of methods for identifying such a subset for the variable Y, particularly using the boundary of Y in DAG G in a distribution compatible with G (see Def. 3.1.9). However this set is not necessarily unique and not even minimal. The following theorem gives a sufficient condition for both [242].

Theorem 7.1.2 ([242]). *If distribution P is stable w.r.t. the DAG G, then the variables corresponding to the nodes in the boundary of Y, $\mathrm{bd}(Y,G)$ (the parents and children of Y and other parents of its children) forms a unique and minimal Markov blanket of Y, $\mathrm{MB}_P(Y)$ (the Markov boundary). Furthermore, $X_i \in \mathrm{MB}_P(Y)$, if X_i is strongly relevant.*

The Markov Blanket Approximating Algorithm assumes that the number of relevant variables is usually much larger for the target variable then for the explanatory variables, so it iteratively omits features for which there is a subset of features forming a Markov blanket without the target variable, consequently not influencing the conditional distribution of the target variable [157]. It uses pairwise correlation for finding a Markov blanket for the features and the KL distance to test the change of the conditional distribution. Recent extension of the algorithm and its applicaton to microarray data are reported in [259]. The Incremential Association Markov Blanket algorithm and its variants similarly use correlation measures and independence tests in forward-backward phases for identifying Markov Blankets, with asymptotic correctness and low computational complexity [242, 243]. Other filter methods directly use Bayesian networks for a preliminary feature selection, which provides usually a restricted set of variables for a computationally more intensive classifier learning in the next phase. The $K2MB$ method first identifies a parental set for the target variables from all the explanatory variables using the $K2$ greedy method (see Section 3.5.2), then it applies the $K2$ algorithm for random orderings of this subset [56]. The learning of a GBN classifier similarly first applies a Bayesian network learning method [46], then it selects the boundary of the target node $\mathrm{MB}(Y,g)$ as a Markov blanket from the resulting Bayesian network G and applies a general Bayesian network learning algorithm or the learning of Bayesian multinets representing also contextual independencies [47, 48, 105].

The *wrapper approach* to feature selection similarly can apply the Bayesian networks as classifiers, in this case jointly in the feature selection phase and the phase of classifier learning [210, 144]. These filter methods indicate that the feature subset selection problem can be approached in a conditional and a model-based way. In the first case, to avoid the statistical (sample) and computational complexity corresponding to complete domain models, the Markov blanket is inferred independently of any other aspect of the domain model (i.e., without evaluating the implications of the identified features for the domain model). Thus these Bayesian network methods have still conditional and frequentist foundation, beside being model free and loss free. So on the one hand, unavoidably the scores for the subsets in these model-free methods has a vague relation to the performance of a given loss function and algorithm over specific model class restricted to the subsets [148, 153]. But on the other hand, (1) the

scores do not utilize the potential of Bayesian networks as domain models (i.e., conditional scores), (2) they have hidden biases, and (3) they have no confidence measures with clear interpretation, partly because of the sequential application of statistical tests on a finite, frequently rather small amount of data. These can be answered in a domain model-based, Bayesian approach to the feature subset selection problem using Bayesian networks.

In the Bayesian conditional approach to feature selection θ encodes the presence of the explanatory variables, so $p(\theta|D)$ induce a (conditional) posterior distribution over the subsets (for an overview of using MCMC methods in a conditional model space over structures with varying input features, see [199], for applications [72, 215]). A hierarchical conditional approach is the Automatic Relevance Determination (ARD) method [194], in which certain parameters represent the weights W_i(relevance) of the inputs (features) X_i, so the parameter posterior for the inputs $p(W_1, \ldots, W_n|D_N)$ can be used for the evaluation of a feature subset.

In the Bayesian domain-based (non-conditional) approach a conditional model of the target variable cannot be separated from the overall domain model or at least the conditional model and the model over the potential explanatory variables are dependent. For example, it is generally so for Bayesian network structure priors, so, as we shall see, we have to average over the model space to derive posterior for the part of the model relevant conditionally (see Eq. 9.8).

As we saw in Th. 7.1.2, the boundary of the variable Y in the Bayesian network G identifies a minimal and unique Markov blanket $\mathrm{MB}(Y, G)$ for variable Y in any stable distribution w.r.t. the DAG G. Using Bayesian network with multinomial local dependency models as unconstrained domain models for discrete values and with Dirichlet parameter priors, the posterior probability of the Markov blanket expresses exactly the belief in the (observational) probabilistic relevance of the subset $\underline{X}'$:

$$p(\mathrm{MB}(Y) = \underline{X}'|D_N) = \sum_G 1(\mathrm{MB}(Y, G) = \underline{X}')p(G|D_N). \tag{7.10}$$

Recall that the structure posterior $p(G|D_N)$ represents the posterior belief in stable distributions w.r.t. G (the non-stables have measure zero see Section 3.1.2.3) and that DAGs in a equivalence class $G \in G^{\sim}$ represent the same set of independencies, so imply the same Markov blanket.

Though the concept of relevance corresponds to subsets, a corresponding pairwise measure can be introduced that defines individual "feature relevance"

$$p(\mathrm{MBM}(Y, X_i)|D_N) = \sum_G 1(X_i \in \mathrm{MB}(Y, G))p(G|D_N). \tag{7.11}$$

Because of model averaging it is still model-based (!), consequently biased towards "domain consistency", contrary to standard pairwise correlation and association measures. Note that only the Bayes risk based subset score is monotone, similarly to a mutual information based subset score, which makes the search in the space of subsets harder. For the application of bootstrap and Bayesian method over MBM features, see Section 7.3 and 7.5.2.3.

7.1.4 MBG subnetworks

The feature subset selection problem does not include explicitly the issue of dependencies between the features, though the interaction between the selected features is important for their interpretation. A generalization of the FSS problem includes the construction of a model containing the variables $\underline{X}'$ relevant to a target variable Y and their observational dependency and causal dependency relations w.r.t. Y.

As shown in Eq. 7.13, the classification performance of a Bayesian network in case of complete data is fully determined by the Markov blanket spanning subgraph $\mathrm{MBG}(Y, G)$ and its parameters (the local models for Y and its children). Another interpretation of the MBG feature is that it encompasses all the causal mechanisms directly related to a given variable Y. Because of the generality of the MBG feature discussed in Section 7.2, we call such model a Markov Blanket Graph or Mechanism Boundary Graph (a.k.a. classification subgraph, feature subgraph).

In the conditional approach, the importance of the MBG feature was already identified, because early methods used the score of a complete Bayesian network G to score the classification performance of the model and to score the Markov blanket of the target variable. As noted in [91] and discussed in Section 9.4.1, this is incorrect from the point of prediction of the target Y, particularly in the case of complete data, because this score includes (direct or indirect) complexity penalization w.r.t. the complete domain model that is not relevant for the MBG submodel relevant for classification. It is more appropriate to use special scores for the classification relevance of the MBG subnetwork and possibly even for scoring the feature subset. Such a classification oriented score is the conditional node monitor (or MBG monitor), its use was reported in [158, 159, 160, 4].

In conditional approaches using other models, the dependency models may contain such additional information about the conditional dependence structure. In Chapter 9 we discuss the logistic regression model, the tree augmented Bayesian network classifiers [91] and the augmented Bayesian classifier [150], which explicitly contain interactions and the MLP model, in which such information is rather implicit.

In the Bayesian framework using Bayesian networks, the corresponding score for the MBG feature is the posterior

$$p(\mathrm{MBG}(Y, G) = \mathrm{mbg}\,|D_N) = \sum_G 1(\mathrm{MBG}(Y, G) = \mathrm{mbg}) p(G|D_N). \qquad (7.12)$$

7.1.5 Learning of subnetworks

The most general structural feature is a general subgraph of a Bayesian network. The identification of subgraphs with statistical significance was reported in [203]. In the first phase, this method generates confidence measure for the pairwise Markov blanket memberships $\mathrm{MBM}(X_i, X_j)$ using the bootstrap. Next, using a heuristic threshold on the bootstrap probabilities for the pairs, it identifies components as starting seeds for a bottom-up expansion to generate multivariable

features from the pairwise features. The attractive assumption behind this approach is that pairwise features corresponding to the same or dependent causal mechanisms are dependent, so they can be identified jointly with higher significance. The evaluation indicated the advantage of this model-based (called "context specific" in their terminology) approach for detecting "correlation" compared to the investigation of direct associations of features with Pearson correlation. The continuation of this work similarly indicated the advantage of learning parts and modules using a special decomposed representation for the Bayesian network [220, 204]. This study also investigated the learning of global pairwise features, such as the existence of a directed path, causal effect between two variables and the learning of parametric features, such as the qualitative type of the local dependency models.

7.1.6 The properties and taxonomy of features

We introduce a terminology to analyze Bayesian network features, particularly the properties of a new BN feature we propose later. The concept of feature over DAGs (Bayesian networks) has a broad usage, it is used for random variables (i.e., a mapping from DAGs G to the real line), for their values, and even for mappings from DAGs G to a set of complete and mutually exclusive composite events. From another point of view, there are simple quantitative random graph properties such as mean in-degrees, out-degrees, clique sizes or lengths of directed paths, and there are complex indicators such as the ABN sentences or complex mappings to subgraphs such as the essential graphs. We use the term feature in a broad sense to denote any function over DAGs G or BNs (G, θ) (e.g., $F(G) : \mathcal{G} \to \mathcal{F}$). If the context allows, e.g. in case of binary features, we use the term feature to refer to the feature function, feature value, and also to the denoted graph property. Frequently a set of feature functions can be indexed by the variables $X_i \in V$ (i.e., $\{F_{X_i}(G)\}$) or pairs of the variables, etc., as it would be another argument of the feature function, so we can talk about univariate features $F(X_i, G)$ or pairwise features $F(X_i, X_j, G)$, instead of referring to the corresponding sets of features.

A feature F is a *local feature* $F(V', G)$, if its value depends only on the subgraph of G spanned by the argument variables $V' \subseteq V$ denoted with $G^{|V'}$ (i.e., $(G_1^{|V'} = G_2^{|V'}) \Rightarrow (F(G_1) = F(G_2))$, where $G^{|V'}$ contains nodes $V' \subseteq V$ and edges of G from V' to V'). A non-local called *global feature* indicates a potential relation to other features and increased computational complexity.

A feature F is a *modular*, if it depends only on the parental sets in DAG G (i.e., $(\mathrm{Pa}(G_1) = \mathrm{Pa}(G_2)) \Rightarrow (F(G_1) = F(G_2))$). A feature is *ordering-modular*, if for all except at most one feature value f and for each complete ordering $\prec$ there is a conjunctive normal form $C_1 \wedge \cdots \wedge C_n$ such that each clause $C_i(f, \prec), G$ for $i = 1, \ldots n$ depends only on $\mathrm{Pa}(X_i, G)$ for all $G^{\prec}$ (i.e., $C_i(f, \prec, \mathrm{pa}(X_i, G))$). Note that the compelled edge relation and the pairwise MBM relevance relation between X_i, X_j are not local, but the MBM relation (through its false value) is modular [97].

Another general type is the *observationally equivalent feature* F, if the mapped

subgraph $F(G)$ over the variables $V' \subseteq V$ depends on only the essential graph of G, $G^{\sim}$ (i.e., $(G_1^{\sim} = G_2^{\sim}) \Rightarrow (F(G_1) = F(G_2))$).

A feature F is called a *complex feature*, if the number of values of the feature is exponential in the number of domain variables.

A set of features $\{F_1, \ldots, F_L\}$ called *DAG-independent feature set*, if the values of the features can be selected arbitrarily without violating the DAG-constraint (i.e., for each L-tuples of feature values, there is one or more DAG G with these feature values: $\forall\{f_1, \ldots, f_L\}\exists G : (F_1(G) = f_1) \wedge \cdots \wedge (F_L(G) = f_L)$).

Finally, let S denote an elementary event (e.g., either G or $(G, \underline{\theta})$). A set of features $\{F_1, \ldots, F_L\}$ is called a *complete feature set*, if for each S the set of values $\{F_1(S), \ldots, F_L(S)\}$ identifies S (i.e., $(S_1 \neq S_2) \Rightarrow (\{F_1(S_1), \ldots, F_L(S_1)\} \neq \{F_1(S_2), \ldots, F_L(S_2)\})$). A set of features $\{F_1, \ldots, F_L\}$ is called complete w.r.t. a feature $F^*(S)$, if for each S the set of values $\{F_1(S), \ldots, F_L(S)\}$ identifies $F^*(S)$. In turn, a feature $F^*(S)$ is a *sufficient feature* for a set of features $\{F_1, \ldots, F_L\}$, if $\forall\ S, i :\ F_i(S) = F_i(F^*(S))$, consequently $p(F_1(S), \ldots, F_L(S))$ can be induced from the distribution of the complex feature $p(F^*(S))$. If additionally, the set of features are complete then the complex feature is called *exact feature* for the feature set (as it is a one-to-one/bijective relation).

7.2 The Markov Blanket (sub)Graph feature

In this section we propose a complex feature, Markov Blanket (sub)Graph feature $(\mathrm{MBG}(Y), \underline{\theta}_{\mathrm{MBG}})$, that includes all the direct causal and probabilistic relations corresponding to a given variable. This feature is at an intermediate level as its complexity is less than of the complete domain model. We show it is a necessary and sufficient feature w.r.t. classification of Y under the usual assumptions in the thesis, such as complete data, discrete values, multinomial local dependency models. The MBG feature can be equally derived from a causal point of view using the mechanism-interventionist interpretation as the minimal set of mechanisms directly relevant for Y, so we equally use the term Mechanism Boundary (sub)Graph feature. It means that the MBG feature represents such a fragment of the domain theory that its distribution is necessary and sufficient to induce the exact posteriors for any classification related feature, to support full scale Bayesian inference and to induce various priors for classifiers, such as logistic regression or multilayer perceptrons. In other words, the complex feature does not violate the dependency of (sub)features for these tasks by modeling them as independent (obviously the MBGs for different variables $(\mathrm{MBG}(X_i), \mathrm{MBG}(X_j), X_i \neq X_j)$ are dependent at the model level in general, so interpreting them as independent using $p(\mathrm{MBG}(X_i), \mathrm{MBG}(X_j)) = p(\mathrm{MBG}(X_i))p(\mathrm{MBG}(X_j))$ would be incorrect).

Definition 7.2.1 ([25, 21]). *The parametric Markov Blanket (sub)Graph feature or Mechanism Boundary Graph feature for a variable Y* $\mathrm{pMBG}(Y, G, \underline{\theta}_G)$ *maps Bayesian network models $(G, \underline{\theta}_G)$ to Markov Blanket Graphs of variable Y and to its parameters* $(\mathrm{MBG}(Y), \underline{\theta}_{\mathrm{MBG}(Y)})$. *The (non-parametric) Markov Blanket Graph feature for a given variable Y denotes the mapping of Bayesian net-*

work structures (G) *to the Markov Blanket Graphs of variable* Y *(see Def. 3.1.11, Fig. 1.1, and Fig. 3.1).*

Because of our general assumptions of global parameter independence and parameter modularity, we always assume that the parameter transformation is a simple selection, so the parameter distribution is unchanged (i.e., $\underline{\theta}_{\mathrm{MBG}(Y,G)} = \{\underline{\theta}_Y, \underline{\theta}_{\mathrm{ch}(Y,G)_1}, \ldots, \underline{\theta}_{\mathrm{ch}(Y,G)_K}\}$ is equal to the corresponding parameters in $(G, \underline{\theta})$, where $K = |\operatorname{ch}(Y, G)|$).

The characteristic property of the pMBG feature is that it completely defines the conditional distribution of Y given the other variables $V \setminus Y$ in a Bayesian network model $(G, \underline{\theta})$ by the local dependency models of Y and its children.

Proposition 7.2.1. *If* $p(\underline{V}|G, \underline{\theta})$ *is defined by a Bayesian network* $(G, \underline{\theta})$*, then the conditional distribution of the target variable* $Y \in \underline{V}$ $p(Y|\underline{V} \setminus Y, G, \underline{\theta})$ *is defined by its Markov Blanket (sub)Graph feature* $\mathrm{pMBG}(Y, G, \underline{\theta}_G)$*.*

Proof.

$$
\begin{aligned}
& p(Y|V \setminus Y, G, \underline{\theta}) \qquad (7.13)\\
&= p(Y|\operatorname{MB}(Y,G), G, \underline{\theta}) = p(Y|\operatorname{pa}(Y,G), \operatorname{ch}(Y,G), \operatorname{pa}(\operatorname{ch}(Y,G),G), \underline{\theta})\\
&\propto p(\operatorname{ch}(Y,G), Y|\operatorname{pa}(Y,G), \operatorname{pa}(\operatorname{ch}(Y),G), \underline{\theta})\\
&= p(Y|\operatorname{pa}(Y,G), \underline{\theta}) \prod_{j=1}^{|\operatorname{ch}(Y,G)|} p(\operatorname{ch}(Y,G)_j|\operatorname{pa}(\operatorname{ch}(Y,G)_j), \underline{\theta}),
\end{aligned}
$$

where $\operatorname{ch}(X_i, G)_j$ denotes the children of X_i in a compatible ordering with G. □

For notational simplicity we assume a binary target varible Y. Let us define a vector-valued feature called *conditional distributional feature* $\mathrm{CD}(Y, G, \underline{\theta})$ denoting the conditional distribution $p(Y|V \setminus Y, G, \underline{\theta})$.

Furthermore, we can state a Bayesian extension of Proposition 7.2.1.

Proposition 7.2.2. *In case of parameter independence, parameter modularity and Dirichlet parameter priors, the Markov Blanket structural and parametric marginals* $p(\mathrm{MBG}(Y,G) = \mathrm{mbg})$ *and* $p(Y|\operatorname{MBG}(Y,G) = \mathrm{mbg})$ *define the conditional distribution of* Y *given other variables* $V \setminus Y$ *in the Bayesian framework, where*

$$p(\mathrm{MBG}(Y,G) = \mathrm{mbg}) = \sum_G 1(\mathrm{MBG}(Y,G) = \mathrm{mbg}) p(G) \qquad (7.14)$$

and $p(Y|\operatorname{MBG}(Y,G) = \mathrm{mbg})$ *denotes the mean distribution* $\mathrm{E}_{\underline{\Theta}'}[p(Y|mbg, \underline{\Theta}')]$.

Proof.

$$
\begin{aligned}
& p(Y|V \setminus Y) \qquad (7.15)\\
&= \sum_G p(G) \int p(Y|G,\underline{\theta}) p(\theta|G)\, \mathrm{d}\underline{\theta} \\
&= \sum_G p(G) \int p(Y|\operatorname{MBG}(Y,G),\underline{\theta}_{\operatorname{MBG}(Y,G)}) p(\underline{\theta}_{\operatorname{MBG}(Y,G)}|G)\, \mathrm{d}\underline{\theta}_{\operatorname{MBG}(Y,G)} \\
&= \sum_G p(G) p(Y|\operatorname{MBG}(Y,G)) \\
&= \sum_{\operatorname{MBG}(Y,G)=\mathrm{mbg}} p(\mathrm{mbg}) p(Y|\mathrm{mbg}),
\end{aligned}
$$

□

Note that Proposition 7.2.2 also indicates that Bayesian model averaging for prediction can be performed in the MBG space, because the parametric marginal $p(Y|\operatorname{MBG}(Y,G)=\mathrm{mbg})$ is efficiently computable in case of Dirichlet parameter priors (see Eq. 3.21). However, in general there is no closed formula for the posterior $p(\operatorname{MBG}(Y,G)=\mathrm{mbg})$, but we can state the following theorem.

Theorem 7.2.1 ([25]). *If the parental set size is bounded by k and the scores $p(pa(X_i)|D_N)$ in Eq. 3.34 are available in $\mathcal{O}(1)$, then the ordering-conditional posterior $p(\operatorname{MBG}(Y,G)=\mathrm{mbg}\,|\prec)$ can be computed in polynomial time.*

Proof. If the parental set size is bounded by k, then

$$
\begin{aligned}
& p(\operatorname{MBG}(Y,G)=\mathrm{mbg}\,|D_N,\prec) \qquad (7.16)\\
&= p(\operatorname{pa}(Y,\mathrm{mbg})|D_N) \prod_{\substack{Y\prec X_i \\ Y\in\operatorname{pa}(X_i,\mathrm{mbg})}} p(\operatorname{pa}(X_i,\mathrm{mbg})|D_N) \prod_{\substack{Y\prec X_i \\ Y\notin\operatorname{pa}(X_i,\mathrm{mbg})}} p(Y\notin \operatorname{pa}(X_i,\mathrm{mbg})|D_N),
\end{aligned}
$$

where

$$
p(Y\notin \operatorname{pa}(X_i,\mathrm{mbg})|D_N) = \sum_{Y\notin\operatorname{pa}(X_i)} p(\operatorname{pa}(X_i)|D_N). \qquad (7.17)
$$

□

Clearly, for a given Markov Blanket structure and ordering Eq. 7.16 directly defines a conjunctive normal form, which gives the next property.

Corollary 7.2.1 ([25, 21]). *The Markov Blanket (sub)Graph feature $MBG(Y,G)$ is an ordering-modular feature.* □

The number of MBGs for a given variable $|\operatorname{MBG}(Y)|$ in case of n variables is still super-exponential (even if the number of parents is bounded above with k). Consider an ordering of the variables such that Y is the first and all the other variables are children of it, then the parental sets can be selected independently, so the number of alternatives is in the order of $(n-1)^{n^2}$ (or $(n-1)^{(k-1)(n-1)}$).

However, at the other extreme, if Y is last in the ordering, then the number of alternatives (i.e., parental sets) is in the order of 2^{n-1} or $(n-1)^{(k)}$). In case of MBG(Y, G), the types of the variable X_i can be (1) non-occurring in the MBG, (2) parent of Y ($X_i \in \text{pa}(Y, G)$), (3) children of Y ($X_i \in \text{ch}(Y, G)$) and (4) (pure) other parent in the MBG ($(X_i \notin \text{pa}(Y, G) \wedge (X_i \in \text{pa}(\text{ch}(Y, G)_j))))$). These types correspond to the categories irrelevant (1) and strongly relevant (2,3,4), as can be seen directly from the definitions of relevance (see, Def. 7.1.1). The number of DAG models $G(n)$ compatible with a given MBG and ordering $\prec$ can be computed as follows: the contribution of the variables $X_i \prec Y$ without any constraint and the contribution of the variables $Y \prec X_i$ that are not children of Y. Let us denote the number of such variables with N_B and N_A respectively, then assuming that the maximal number of parents is k, the number of compatible DAGs is $2^{\Theta((k-1)(N_B+N_A)\log(n))}$.

Proposition 7.2.1 and Proposition 7.2.2 offer two interpretations for the MBG feature. From a (conditional) probabilistic point of view the MBG(G) feature defines an equivalence relation over the DAGs w.r.t. the conditional distribution of Y given all the other variables under parameter modularity and global parameter independence. This is the consequence of Th. 3.1.3 and Th. 7.1.2, which allow the reduction of the space of DAGs to the space of MBGs from the point of view inferring a given variable. If the hypotheses are the observational classes (i.e. the parameter and structural priors are identical for observationally equivalent DAGs), then this conditionally induced equivalence relation is combined with the observational equivalence relation, which allows further reduction of the space of MBGs (for a partially oriented representation of the MBGs, see [3, 4]). We show certain properties of this combined equivalence, although in our exploratory context we assume causal priors, so we cannot simplify further the MBG space. In the non-Bayesian context let us define a pairwise relation over Bayesian networks as G_1 and G_2 are inferentially equivalent for variable Y, if they can encode the same set of conditional distributional features for Y (i.e., for each CD$(Y, G_1, \underline{\theta}_1)$, there exists a $\underline{\theta}_2$ such that CD$(Y, G_1, \underline{\theta}_1) =$ CD$(Y, G_2, \underline{\theta}_2)$). Clearly, observational equivalence and MBG equivalence of DAGs G_1, G_2 implies conditional distributional equivalence (IE), but MBG equivalence and conditional distributional equivalence does not imply observational equivalence. Interestingly, MBG equivalence is not implied by observational equivalence or by conditional distributional equivalence (i.e., the MBG feature is not a unique representant of an inferentially equivalent class of Bayesian networks and it can be different in observationally equivalent DAGs).

From a causal point of view, this feature uniquely represents the minimal set of mechanism including Y despite the non-uniqueness of the MBG feature w.r.t. the acausal conditional distributional equivalence. This offers the second interpretation of the MBG feature: the pMBG$(Y, G, \underline{\theta})$ feature includes exactly the mechanisms containing the variable Y, hence the name Mechanism Boundary (sub)Graph feature pMBG$(Y, G, \underline{\theta})$. The probability of an MBG is the sum of probabilities of the causal domain models that are compatible with this causal subtheory for the variable Y (Eq. 7.14), which shows that for example infer-

entially equivalent MBGs may have different probabilities in a causal context (e.g., in case of causal prior or interventionist data).

From Proposition 7.2.1 we can conclude that the MBG(Y, G) feature is necessary and sufficient to represent the mechanisms directly relevant for the variable Y and from the point of view of prediction of Y, it is a *sufficient* feature for the conditional distributional features of Y. In other words, under the conditions such as parameter modularity, global parameter independence and complete data assumption, this structural and parametric feature of the causal BN domain model is necessary and sufficient to support the manual exploration and automated construction of a causal, probabilistic, interpretable conditional dependency model. This "ultimate" property of the MBG feature suggests the concept of conditional feature and the generalization of the feature subset selection problem.

Definition 7.2.2. *A feature (function) F is called conditional feature for a given variable Y, if it depends only on* $(\text{MBG}(Y), \underline{\theta}_{\text{MBG}(Y)})$

$$\text{pMBG}(Y, G_1, \underline{\theta}_1) = \text{pMBG}(Y, G_2, \underline{\theta}_2)) \Rightarrow (F(G_1, \underline{\theta}_1) = F(G_2, \underline{\theta}_2). \quad (7.18)$$

Definition 7.2.3. *In case of a stable distribution $p(Y, \underline{X})$, the feature (sub)graph selection problem (FGS) denotes the identification of a Markov Blanket subgraph $MBG(Y, G)$, where DAG G denotes a perfect map of distribution p (i.e., it includes the identification of a Markov Blanket set $\underline{X}' \subseteq \underline{X}$ w.r.t. p and Y, and a Bayesian network substructure over $\underline{X}'$ representing the dependencies between these variables, excluding incoming edges into the parents of Y).*

7.3 The bootstrap confidence measure

The *bootstrap* approach to induce confidence measures for Bayesian network features was investigated as an alternative to the Bayesian approach to support statistical inference from small sample [96, 95]. An important motivation was to avoid the Monte Carlo simulations usually necessary in the Bayesian approach by using a simple resampling scheme and optimization.

The bootstrap is a general purpose, computationally intensive statistical inference method using resampling to assess the accuracy of a statistical estimate given a finite sample [83, 125]. We discuss it here as we refer to it only in this context, but it is a general statistical methodology and applicable with arbitrary model classes (or without as a nonparametric bootstrap). Assume a fixed i.i.d. sample $D_N = \{\underline{X}_1, \ldots, \underline{X}_N\}$ and let us denote $\hat{\theta}(D_N)$ the statistical estimate of interest and θ_0, the unknown true parameter. For a given sample size N its distribution, particularly its deviation $\hat{\theta}(D_N) - \theta_0$ is also of interest for constructing confidence intervals and hypothesis testing. The standard frequentist approach analytically derives its distribution, confidence intervals for restricted sets of sampling models and estimates (e.g., Gaussian data generation and mean estimate). Note that if we had access to the generative model $p(X|\underline{\theta}_0))$, we could sample it for any complex estimate. The standard Bayesian

approach would define a probabilistic model for the observations $p(X|\theta)$ with prior $p(\theta)$ providing a distribution for the estimate $\int p(\hat{\theta}(D_N)|\theta)p(\theta)\,\mathrm{d}\theta$, which can be analyzed or sampled to explore. The central idea of nonparametric bootstrap is the characterization of the distribution of the unobservable deviation $\hat{\theta} - \theta_0$ with the following distribution $\hat{\theta}^*(D_N^*) - \hat{\theta}(D_N)$, where the data set D_N^* of N samples (the boostrap replicate) is drawn uniformly from the observed D_N with replacement. That is, given a fixed sample D_N we define a bootstrap sample distribution over the finite (!) number of possible data sets D_N^*, which allows the assessment of the accuracy of the estimate $\hat{\theta}(D_N)$ by the distribution of $\hat{\theta}^*(D_N^*)$. In general, the bootstrap for $\hat{\theta}(D_N)$ is called consistent if

$$p(\hat{\theta}^*(D_N^*) - \hat{\theta}(D_N)) \to p(\hat{\theta}(D_N) - \theta_0) \text{ as } N \to \infty \text{ in distribution.} \tag{7.19}$$

For example, the *ideal (nonparametric) bootstrap estimate* of the variance $\mathrm{var}_{p(D_N)}(\hat{\theta}(D_N))$ is defined as $\mathrm{var}_{p(D_N^*)}(\hat{\theta}^*(D_N^*))$ (see [83]), which can be shown to provide a consistent estimate [83]. Because of the large number of bootstrap data sets with size N, the ideal bootstrap estimate is approximated by its Monte Carlo estimate using B number of randomly drawn bootstrap data sets $D_{b,N}^*$ for $b = 1, \ldots, B$ and the corresponding quantities $\hat{\theta}_b^*(D_{b,N}^*)$ as follows

$$\hat{\mathrm{var}}_B(\hat{\theta}^*) = \sum_{b=1}^{B} (\hat{\theta}_b^* - \hat{\theta}_{(.)}^*)^2/(B-1) \text{ where } \hat{\theta}_{(.)}^* = \sum_{\mathrm{i}=1}^{\mathrm{B}} \hat{\theta}_{\mathrm{b}}^*/\mathrm{B}. \tag{7.20}$$

The Monte Carlo estimate of the ideal bootstrap estimate itself has a variance, which is asymptotically $c_1/N^2 + c_2/NB$, so relatively low number of bootstrap replicates suffices in practice [83]. This also indicate that the distribution of $\hat{\theta}_b^*(D_{b,N}^*)$ is more spread than of the target $\hat{\theta}(D_N)$, so it cannot be used directly (e.g., for constructing quantiles for $\hat{\theta}(D_N)$).

Now we can turn to the application of the bootstrap to induce confidence measures for model structures and its properties. This is not without problems as its first application in the model space of phylogenetic trees has shown (phylogenetic trees represent evolutionary relationships between entities corresponding to its nodes [80]). We will follow the terminology and explanations from that field [134, 86, 35, 82, 7]. Assume that the i.i.d. data set D_N is generated from an unknown Bayesian network model $M_0 = (G_0, \underline{\theta}_0)$ and a fixed algorithm $\mathcal{C}$ induces the model structure $\hat{G}_{\mathcal{C}}(D_N)$, more exactly our hypothesis space are the observation equivalence classes of DAGs $G^{\sim}$. Because the estimate is a model structure without a semantic metric, we cannot define confidence intervals for models with an accuracy parameter, so the probably approximately correct (PAC) terminology is only partly applicable [246]. This frequentist definition of a confidence value is the probability of exact model induction with data sets of size N

$$p(D_N : \hat{G}_{\mathcal{C}}^{\sim}(D_N) = G_0^{\sim} | M_0, N). \tag{7.21}$$

The essence of the argument for the assessment of Eq. 7.21 with bootstrap is as follows (adapted for discrete valued Bayesian network learning from [86, 82,

83]). By assuming the naive table representation with $d = \prod_i |X_i|$ entries we can interpret a complete (!) data set as corresponding empirical relative frequencies for the complete configurations denoted with $\hat{\underline{\theta}}$, which geometrically is located on the d-dimensional simplex. Note that for a fixed size N, this determines both Bayesian network learning scores, so we can write $\hat{G}^{\sim}(\hat{\underline{\theta}})$. Disregarding that this statistical estimate changes non-continuously across boundaries, it looks like a standard bootstrap problem to assess its accuracy, that is to estimate the probability that $\hat{\underline{\theta}}$ is in the same region as $\underline{\theta}_0$ (i.e., $\hat{G}^{\sim}(\hat{\underline{\theta}}) = G_0^{\sim}(\underline{\theta}_0)$). For a given fixed data set D_N and corresponding $\hat{\underline{\theta}}$, this is estimated using the bootstrap frequencies $\hat{\underline{\theta}}^*$, similarly to the standard case when the distribution of $\hat{\theta}(D_N) - \theta_0$ is assessed with the distribution of $\hat{\theta}^*(D_N^*) - \hat{\theta}(D_N)$. So the probability of exact model induction for an induced model $\hat{G}_{\mathcal{C}}^{\sim}(D_N)$ given a data set D_N theoretically can be characterized with the bootstrap probability and approximated with its Monte Carlo estimate

$$p(D_N^* : \hat{G}_{\mathcal{C}}^{\sim}(D_N^*) = \hat{G}_{\mathcal{C}}^{\sim}(D_N)|D_N) \approx \frac{1}{B}\sum_{b=1}^{B} 1(\hat{G}_{\mathcal{C}}^{\sim}(D_{b,N}^*) = \hat{G}_{\mathcal{C}}^{\sim}(D_N)). \quad (7.22)$$

For phylogenetic trees with model structure T it is shown that the posterior for $T = \hat{T}$ using uninformative prior is nearly equal to the bootstrap probability for $\hat{T}^* = \hat{T}$ (called the "poor man's" Bayes posterior [125]).

We can proceed analogously for the structural features for Bayesian networks. The ideal confidence value is the probability of the induction of the structural feature of the underlying essential graph $F(G_0) = f_0$ with data set of size N [96, 95]

$$p(D_N : F(\hat{G}_{\mathcal{C}}^{\sim}(D_N)) = f_0|M_0, N). \quad (7.23)$$

This quantity is called "accuracy" in phylogenetics [134, 86]. As noted in [35, 86], this concept is still applicable for a non consistent induction algorithm widely used in a domain as indicating non-repeatability by the lack of support from a well-accepted method. With consistent structure learning algorithms as in the case of Bayesian networks, this value will converge to 1 with increasing N. Though the theoretical background for the application of bootstrap is still unsolved, because of the discrete valued estimate and the consistency properties of the induction algorithm $\mathcal{C}$, in empirical experiments the bootstrap probabilities of features were adopted as assessing the confidence values for features in the induced model $F(\hat{G}_{\mathcal{C}}^{\sim}(D_N)) = f_{D_N}$ given a data set D_N [96, 95].

$$p(D_N^* : F(\hat{G}_{\mathcal{C}}^{\sim}(D_N^*)) = f_{D_N}|D_N). \quad (7.24)$$

This is also backed by the arguments for phylogenetic trees. The bootstrap probabilities are approximated with their Monte Carlo estimates,

$$\frac{1}{B}\sum_{b=1}^{B} 1(\hat{G}_{\mathcal{C}}^{\sim}(D_{b,N}^*) = f_{D_N}), \quad (7.25)$$

with Monte Carlo variance rapidly decreasing with B, as mentioned above. However, the variation of the bootstrap probabilities depending on D_N in case of phylogenetic trees led to the concept of "repeatability" and its classical investigations empirically [134] and analytically [86]. An important clarification of a potential misuse of bootstrap was that the quantity

$$p(D_N^* : F(\hat{G}_{\mathcal{C}}^{\sim}(D_N^*)) = f_0 | D_N) \tag{7.26}$$

is not approximating the accuracy (i.e., the probability of induction of "true" features in Eq. 7.23). As suggested [86], a bootstrap probability p for an induced feature can be interpreted as a 1-p-value for the hypothesis that the feature is not present. For phylogenetic trees, a (computationally intensive) correction of the bootstrap probability for its use in the standard hypothesis testing framework is suggested in [82].

For Bayesian networks the bootstrap approach was applied for the following structural features: compelled edges $CompE(X_i, X_j|G)$ (as direct causal relation), Markov blanket membership $\mathrm{MBM}(X_i, X_j|G)$ (as pairwise relevance), pairwise precedence $X_i \prec_G X_j$ (as causal relation) [96, 95] (see results for partly parametric features [203]). The bootstrap probabilities in Eq. 7.24 for the features were interpreted as "support from a given algorithm" [96] and later in testing various induction algorithms as the assessment of the confidence of the induced feature as defined in Eq. 7.23. The experiments were conducted on a gold standard model as reference, which allowed the generation of multiple data sets for proper evaluation of the bootstrap, and on data sets from a genomic and text domain as well.

In summary, earlier works provided an empirical support for the applicability of the bootstrap for Bayesian network features with the following conclusions [96, 95]. It yields a cautious, conservative estimate (no false-positive error) for the features, but its applicability seems sensitive to the domain (e.g., the selection of a confidence threshold for reporting the features), and to the optimization algorithm. Certain pairwise features can be more reliably estimated, especially the pairwise Markov blanket relation (MBM), which can be explained by the topological robustness of this feature (i.e., a given relation can occur in large number of DAGs). The induced confidence measures were reported visually (as colors and thickness of the Bayesian network edges) to support efficient interpretation of the result of statistical inference from small amounts of data with large number of variables. Another use of the induced confidence measure also gave promising results, to support the second-phase learning of full Bayesian network models and subnetworks using the feature confidences as soft and hard constraints.

However, the relation of the bootstrap approach to the Bayesian approach is subtle w.r.t. the induced confidence measures for Bayesian network features, despite that under specific conditions the bootstrap probabilities approximate the corresponding posteriors [80, 82]. The Bayesian approach is capable to provide updated beliefs — the posterior — for an arbitrary fixed structural feature $F(G) = f_0$ given the observations D_N, either by Monte Carlo sampling

or sometimes analytically. This posterior practically can be approximated by the set of models $\mathcal{G}_{\mathcal{C}}^{\mathrm{HPD}}$ with high posteriors identified using an optimization algorithm $\mathcal{C}$ with some heuristic randomization (to correct its bias for local minima):

$$\begin{aligned} p(F(G) = f_0|D_N) &= \sum_G 1(F(G) = f_0)p(G|D_N) \\ &\approx \frac{1}{\sum_{G\in\mathcal{G}_{\mathcal{C}}^{\mathrm{HPD}}} p(G|D_N)} \sum_{G\in\mathcal{G}_{\mathcal{C}}^{\mathrm{HPD}}} 1(F(G) = f_0). \end{aligned} \tag{7.27}$$

Conversely, the bootstrap approach can provide confidence values for features in the frequentist, hypothesis testing framework by the bootstrap probabilities (i.e., by its Monte Carlo estimates):

$$\begin{aligned} p(F(\hat{G}_{\mathcal{C}}^{\sim}(D_N)) = f_0|M_0, N) &\approx p(F(\hat{G}_{\mathcal{C}}^{\sim}(D_N^*)) = f_{D_N}|D_N) \\ &\approx \frac{1}{B}\sum_{b=1}^{B} 1(\hat{G}_{\mathcal{C}}^{\sim}(D_{b,N}^*) = f_{D_N}). \end{aligned} \tag{7.28}$$

Indeed, as the similarity of the final sums suggests in Eq. 7.27 and Eq. 7.28, the bootstrap can be conceived as a heuristic method using perturbed data sets to generate a good subset of models $\mathcal{G}_{\mathcal{C}}^{\mathrm{HPD}}$ with high posteriors around the maximum a posteriori or maximum likelihood Bayesian network structure. But it cannot be used in general as an approximation to the sampling distribution $p(D_N|M_0, N)$, consequently to sample $p(\hat{G}_{\mathcal{C}}(D_N))|M_0, N)$ or to approximate the posterior $p(G|D_N)$, particularly not in the small sample case, which is the primary goal of learning Bayesian network features.

Furthermore, as the learning of Bayesian network is NP-hard, the computational complexity of the heuristic algorithms used in practice is comparable to the computational complexity of the application of Monte Carlo methods for Bayesian networks with computationally efficient sampling. In fact, after the investigation of the bootstrap approach [96, 95], the authors also reported an efficient Bayesian approach for inducing Bayesian confidence measures for certain Bayesian network features, which is applied in this thesis and described in the next section.

7.4 On the advantage of feature posteriors

After the overview of BN features, certain frequentist identification methods, and the bootstrap methodology to induce confidence measures, we now turn to the Bayesian approach.

The main disadvantage of the frequentist identification methods is that the significance level, if there is any or which in principle what could be derived with general aggregation methods of significances, is not model-based. Furthermore, the methods are fragmented by the type of the features (i.e.,

there are dedicated algorithms for the identification of local causal features ($RCEdge(X, Y)$), relevant variables and their subsets ($\mathrm{MB}(X), \mathrm{MBM}(X, Y)$) or subtheories ($G' \subseteq G$)).

The bootstrap methodology provides a model-based confidence value, its asymptotic behavior for increasing sample size is guaranteed with a consistent induction algorithm, although there are no theoretical results for its application on structural features for small sample size and it can be applied uniformly for arbitrary features. Furthermore, as it includes a model identification for each bootstrap replicates, its computational complexity can be considerable (e.g., compared to Bayesian Monte Carlo methods).

The introduction of Dirichlet parameter priors with parameter independence for Bayesian networks by Spiegelhalter et al. [227] (conjugate for the multinomial sampling, see Sections 3.2.1.2) provided an efficiently computable closed form for the posterior for the parental sets and for the structure conditional on a given ordering. Based on this, in the beginning of the 1990's the full Bayesian approach was proposed and advocated in a seminal paper [40]. In this paper Buntine proposed the posterior knowledge base view and analysis of the properties of the Bayesian network model conditioned on a given ordering. He also developed a construction method of an approximate posterior offline knowledge base to support theory (i.e., prior) refinement and full scale Bayesian inference. In [57], Cooper et al. discussed the general use of the posterior over Bayesian network structures as an inductive probabilistic knowledge base (i.e., to compute the posterior of arbitrary model properties). However this work had not proposed method to carry out the Bayesian inference. In [178], Madigan et al. proposed an MCMC scheme to approximate such Bayesian inference using the space of DAGs and PDAGS (utilizing also the orderings of the variables). They also developed the Ockham window algorithm for the construction of a small, selective set of models to support exploration of the posterior and inference with it. In [133], Heckerman considered the application of this full Bayesian approach to causal Bayesian networks (under the Causal Markov Condition). The DAG-based MCMC method was improved by Castelo et al. [112]. In [66], Dash et al. reported a method to perform exact full Bayesian inference in a restricted case of naive Bayesian classifiers. In [97, 98], Friedman et al. reported another MCMC scheme utilizing the ordering of the variables (hence its name, ordering-based MCMC method), which used a closed form for the ordering-conditional posterior of Markov blanket membership, beside the earlier closed form for parental membership. In [154], Koivisto et al. reported a method to perform exact full Bayesian inference over modular features in $\mathcal{O}(n2^n)$ time. Note that the treatment of the submodels as independent hypotheses differs from our approach, which treats them as aggregates of compatible complete models. It would include the assumption of the existential uncertainty of the domain objects represented by the random variables (for the discussion of treating orderings as sets of compatible DAGs or as separate objects, see 7.5.2.2).

Before discussing these methods and their application for complex features, first we summarize the properties of the Bayesian approach and open issues.

1. *Normativity.* The Bayesian approach is a normative, model-based combination of prior and data, so the inputs and the outputs are probabilities conditional on the observed data, which are applicable in the Bayesian decision-theoretic framework. Consequently, its application and interpretation in the small sample region is unconstrained.

2. *Probabilistic knowledge base.* The feature posteriors can be embedded into a probabilistic knowledge base, possibly with textual enrichment as in the case of ABN-KBs. An important question particularly for complex features is the efficient or approximate representation of the distribution over the feature space.

3. *Probabilistically linked model spaces and induced priors.* The feature posteriors can be used to induce priors for linked model spaces. For classifiers, see Chapter 10 and for the comparison of learning dual-Bayesian networks and the two-phased learning of Bayesian networks from literature data and clinical data, see Section 8.1).

4. *Optimally selected feature complexity.* The induced posteriors for the features are dependent in general. A solution followed in the thesis is the definition of a semantically important complex feature, (i.e., subtheory), which includes many dependent simpler features and estimate its posterior distributions.

5. *Integrated estimate and search method.* An already investigated and solved question is the estimation of a moderate number of posterior values (expectations) (e.g., pairwise features such as edge relation or Markov blanket membership with $\mathcal{O}(n^2)$ cardinality). However, the number of values of a complex feature can be exponentially large (e.g., the number of Markov blanket subsets is $\mathcal{O}(2^n)$), so search methods have to be integrated into the Monte Carlo inference methods to find feature values with relevant posterior.

7.5 MC methods for a feature posterior

As we discussed in Section 7.1, there are two approaches to the use of BN features. The first approach (reported in [40, 180, 131, 96, 95, 97, 98]) is based on a set of simple features to construct a fragmentary representation for the distribution over the complete domain model from multiple, though simple aspects using various interdependent marginals, such as edge probabilities. The other approach is based on a complex feature (or subtheory), which is a focused representation from a restricted, but still comprehensive point of view. In our case, this is the MBG feature to support classifier construction.

In both cases we have to use Monte Carlo methods to perform the Bayesian inference, because of the lack of analytical formulas for the posterior of the

features. So first, we summarize MC methods: the most direct DAG/PDAG-MCMC method and a latter developed method, the so-called ordering-based MCMC method to estimate the posterior of a limited number of features.

7.5.1 The DAG-based MCMC methods

The basic task is the estimation of the expectation of a given random variable $F(G)$ over the space of DAGs with a specified confidence level.

$$\hat{F} \approx \bar{F} = \mathrm{E}_{p(G|D_N)}[F(G)]. \tag{7.29}$$

In Eq. 3.44), we derived an efficiently computable closed formula for the (unnormalized) posterior of DAGs or for PDAGs in case of likelihood equivalent priors and our standard assumptions, such as complete data, discrete domain variables, multinomial local conditional distributions and Dirichlet priors at the parametric level. As the posterior over DAGs cannot be sampled directly in general and the construction of an approximating distribution to use in importance sampling is frequently not feasible, the standard approach is to use MCMC methods, such as the Metropolis-Hastings algorithm over the DAG or PDAG space (see Section 2.3.1.2).

The first application of *DAG-based MCMC* methods for BN feature estimated the posterior of compelled edges [178]. It investigated two proposal distributions. The first constructs a candidate by perturbing directly the edges with insertions, deletions and reversals. The second constructs a candidate by perturbing the partial ordering of the variables and then perturbing the edges to be compatible with this candidate ordering.

7.5.2 The ordering-based MCMC methods

The DAG-based MCMC method for estimating a given expectation is generally applicable, but its statistical properties frequently can be improved by specializing it to a certain type of features. In this section we consider the *ordering-based MCMC* method, which is a hierarchic, semi-analytic MCMC method [97]. We shall see in Section 7.7 that this method can be utilized also to integrate the estimation and the search process in the case of large numbers of features.

7.5.2.1 The ordering-conditional feature posteriors

Assuming modular structure priors, parameter independence, and modularity and complete data, the structure posterior has the following product form:

$$p(G, D_N) = \prod_i^n p(D_N | \operatorname{pa}(X_i, G)) p(\operatorname{pa}(X_i, G)).$$

The ordering-based MCMC method relies on the following two uses of this product form [40, 66, 97]. First, we note that the set of DAGs compatible with an ordering $\prec$ can be constructed as the Descartes product of sets of parental

sets compatible with the ordering, so combining this with the product form of the probability of DAG G we have

$$\begin{aligned} p(D_N|\prec) &= \sum_{G\in\mathcal{G}^{k(n),\prec}} p(D_N,G|\prec) \qquad (7.30)\\ &= \sum_{G\in\mathcal{G}^{k(n),\prec}} \prod_i^n p(D_N|\operatorname{pa}(X_i,G))p(\operatorname{pa}(X_i,G)|\prec)\\ &= \prod_i^n \sum_{\operatorname{pa}(X_i,G)\sim\prec} p(D_N|\operatorname{pa}(X_i,G))p(\operatorname{pa}(X_i,G)|\prec), \end{aligned}$$

where $\operatorname{pa}(X_i,G)\sim\prec$ denotes the compatibility of a parental set $\operatorname{pa}(X_i,G)$ with ordering $\prec$. Second, for an *ordering-modular feature* $F(G)=f$ defined as $\bigwedge_1^n C_i(f,\prec,\operatorname{pa}(X_i,G))$, where C_i is true for some parental sets possibly conditionally on a given ordering $\prec$, we have

$$\begin{aligned} p(f,D_N|\prec) &= \sum_{\substack{G\in\mathcal{G}^{k(n),\prec}\\ F(G)=f}} p(D_N,G|\prec) \qquad (7.31)\\ &= \sum_{\substack{\operatorname{pa}(X_i,G)\sim\prec\\ F(G)=f}} \prod_i^n p(D_N|\operatorname{pa}(X_i,G))p(\operatorname{pa}(X_i,G)|\prec)\\ &= \prod_i^n \sum_{\substack{\operatorname{pa}(X_i,G)\sim\prec\\ C_i(f,\prec,\operatorname{pa}(X_i,G))}} p(D_N|\operatorname{pa}(X_i,G))p(\operatorname{pa}(X_i,G)|\prec). \end{aligned}$$

This gives the following proposition (the generalization of Th. 7.2.1).

Proposition 7.5.1. *For an ordering-modular feature $F(G)=f$ defined as $\bigwedge_1^n C_i(f,\prec,\operatorname{pa}(X_i,G))$, the ordering conditional posterior is decomposed as*

$$\begin{aligned} p(f|D_N,\prec) &= \frac{p(f,D_N|\prec)}{p(D_N|\prec)} \qquad (7.32)\\ &= \prod_i^n \frac{\sum_{\substack{\operatorname{pa}(X_i,G)\sim\prec\\ C_i(f,\prec,\operatorname{pa}(X_i,G))}} p(D_N|\operatorname{pa}(X_i,G))p(\operatorname{pa}(X_i,G),f|\prec)}{\sum_{\operatorname{pa}(X_i,G)\sim\prec} p(D_N|\operatorname{pa}(X_i,G))p(\operatorname{pa}(X_i,G)|\prec)}\\ &= \prod_i^n p(C_i(f,\prec,\operatorname{pa}(X_i,G))|D_N,\prec). \end{aligned}$$

□

The possible special ("complementer") value without such form can be managed by appropriate summations for the other feature values. Note that if the maximum number of parents is bounded by k, then the ordering conditional feature posterior in Eq. 7.32 can be computed in polynomial time $\mathcal{O}(n^{k+1})$ in contrast to the exponential number of DAGs compatible with an ordering involved in the summations in Eq. 7.30, 7.31 [97].

7.5.2.2 Advantages of ordering-based MCMC

The existence of the unnormalized posterior for the orderings and the normalized ordering-conditional posterior for a feature allows semi-analytic ordering-based MC methods with advantageous properties w.r.t. DAG-based MC methods.

First, consider the statistical effect of using orderings instead of DAGs and ignore the effect of the MC method used. By assuming a binary feature $F(G)$ and using the identity $\mathrm{E}[X] = \mathrm{E}_Y[\mathrm{E}_X[X|Y]]$ the target quantity can be rewritten as

$$\mathrm{E}[F(G)|D_N] = \mathrm{E}_{p(\prec,D_N)}[\mathrm{E}[F(G)|\prec, D_N]], \tag{7.33}$$

where the random variable $p(F(G)|\prec, D_N) = \mathrm{E}[F(G)|\prec, D_N]$ has variance $\mathrm{var}_{p(\prec|D_N)}(\mathrm{E}[F(G)|\prec, D_N])$. We can decompose it as follows, which directly follows from the identity $\mathrm{var}(X) = \mathrm{E}_Y[\mathrm{var}(X|Y)] + \mathrm{var}_Y(\mathrm{E}[X|Y])$ [108].

Proposition 7.5.2. *The variance of a binary feature $F(G)$ $\mathrm{var}_{p(G|D_N)}(F(G))$ using the augmented space of $\mathcal{G} \times \{\prec\}$ with the distribution $p(G|\prec)p(\prec)$ is the sum of its mean variance and the variance of its mean:*

$$\begin{aligned} &\mathrm{var}_{p(G|D_N)}(F(G)) \\ &\quad = \mathrm{E}_{p(\prec|D_N)}[\mathrm{var}(F(G)|\prec, D_N)] + \mathrm{var}_{p(\prec|D_N)}(\mathrm{E}[F(G)|\prec, D_N]). \end{aligned} \tag{7.34}$$

□

Consequently, the availability of the ordering conditional posterior for a feature allows the cancellation of the term $\mathrm{E}_{p(\prec|D_N)}[\mathrm{var}(F(G)|\prec, D_N)]$ in the ordering-based MC approach compared to a DAG-based method with identical DAG posteriors. It can be a significant reduction because of the asymptotic behavior of the two terms. The expected variance of the ordering conditional probability of a feature is the expectation of the variance of a Bernoulli random variable with parameter $p(F(G)|\prec, D_N)$. In contrast, the other term can be close to zero, if the ordering-conditional posterior of a feature has a similar value for the orderings compatible with the essential graph generating the observations.

The decrease of the variance is not simply the consequence of "collapsing" the $\mathcal{G}(n)$ space into the space of orderings with smaller cardinality of $n!$, but of the augmentated state space with the orderings $\mathcal{G} \times \{\prec\}$ and the analytic marginalization of the ordering conditional DAGs in the case of ordering modular features (for the general effects of hierarchical approaches and collapsing the state space by analytical marginalization, a.k.a. Rao-Blackwellisation, on MC sampling, see [107]).

However, note that Proposition 7.5.2 treats the DAG space as part of an extended space and the explicit, autonomous use of the orderings in the joint distribution $p(G|\prec)p(\prec)$ can introduce a bias (cf. the implicit use of the orderings as sets of compatible DAGs with an induced distribution from $p(G)$). If the uniform distribution $p(\prec)$ is used as non-informative, then it has a bias towards DAGs compatible with many orderings. For example the empty graph is $n!$ times more probable than any complete graph. However, this bias is not related to standard measures of model complexity (i.e., to Ockham principle)

as the number of compatible orderings is different for observationally equivalent DAGs (e.g., a Markov chain with different, but observationally equivalent orientations, see Example 3.1.2). An interesting direct consequence is the following proposition.

Proposition 7.5.3. *The induced prior* $p(G) \propto \sum_{\prec \sim G} p(\prec)$ *from a uniform* $p(\prec)$ *violates the structural prior equivalence (see Section 3.1.5.2.4).* □

A computationally expensive solution to maintain uniformity over the DAGs is to weight the DAGs through $p(G|\prec)$ properly.

Second, let us compare the ordering-based MC method against the DAG-MC method computationally. Assume that the posteriors of the ordering-conditional parental set are available in $\mathcal{O}(1)$ time (they can be precomputed in $\mathcal{O}(Nn^k)$ time and stored in $\mathcal{O}(n^{(k+1)})$ space, which is either directly acceptable or can be significantly decreased by caching only the high-scoring parental sets). Let $P(n)$ denote the time complexity of the drawing a sample or proposal, which is typically $\mathcal{O}(n^2)$, and $F(n)$ the time complexity of the target feature $F(G(n))$, which is $\mathcal{O}(1)$ for edges, $\mathcal{O}(n)$ for the MBM, MBG and MB features. The unnormalized posterior $p(G, D_N)$ can be computed in $\mathcal{O}(n)$ (assuming the precomputation and storage of the local scores). Thus the overall time complexity of one step of DAG-based MC method is $\mathcal{O}(n^2)$. For the ordering-based MC method this is $\mathcal{O}(n^{k+1})$, but it evaluates n^{kn} or $2^{\mathcal{O}(kn\log(n))}$ DAGs in one step.

Furthermore, to perform exact full Bayesian inference over modular features a dynamic programming method can be used over subsets instead of the naive enumeration of the orderings [154]. This method reduces the super-exponential $\mathcal{O}(n!)$ to $\mathcal{O}(n2^n)$ time, but it requires $\mathcal{O}(n2^n)$ space.

7.5.2.3 Estimating edge and pairwise relevance

In the proposal of the ordering-based MCMC method and in subsequent applications the setting was the following [97, 98]. The ordering prior $p(\prec)$ was uniform. The ordering-conditional structure prior $p(G|\prec)$ was a modular prior with uniform weights for the size of the parental sets up to a limit k and with uniform weights for the parental sets with a given size. The parameter independence and modularity were assumed, and the $\mathrm{BD_{eu}}$ parameter prior was used. The MCMC method in the ordering space used two kinds of operations in the proposal distribution: the replacement of pairs and the circular (modulo) shifting of the ordering. The number of variables was 35 in a medical domain, 100-1000 in the genetic and text-mining domains. The target features were the edges ($X_i \to X_j$), the pairwise relevance relations ($\mathrm{MBM}(X_i, X_j)$), the pairwise precedence relations ($X_i \prec X_j$) and the pairwise causal relations ($X_i \dashrightarrow X_j$). There is a closed form for the ordering-conditional posterior, except for the existence of a directed path between two nodes. By noting that the edge feature $f_{X_i \to X_j}$ is an ordering-modular feature and for a given ordering only one clause is relevant in Eq. 7.32, its ordering-conditional posterior is as follows:

$$p(f_{X_i \to X_j} | D_N, \prec) \tag{7.35}$$
$$= \frac{\sum_{\substack{X_i \in \mathrm{pa}(X_j,G) \\ \mathrm{pa}(X_j,G) \sim \prec}} p(D_N | \, \mathrm{pa}(X_j,G)) p(\mathrm{pa}(X_j,G) | \prec)}{\sum_{\mathrm{pa}(X_j,G) \sim \prec} p(D_N | \, \mathrm{pa}(X_j,G)) p(\mathrm{pa}(X_j,G) | \prec)}.$$

The ordering-conditional posterior of the Markov Blanket Membership feature $f_{\mathrm{MBM}(X_i,X_j)}$ given $\prec$ can be derived by noting that for a given $\prec$ the clauses in the conjunctive normal form for the false value are as follows (assuming $X_i \prec X_j$): earlier parental sets are irrelevant (empty for $X_i \prec X_j$), X_i is not parent of X_j (the clause for X_j includes the parental sets without X_i), and there is no common child of X_i and X_j (the clauses for variables after $X_j \prec X_l$ include the parental sets without X_i and X_l)

$$p(f_{\neg f_{\mathrm{MBM}(X_i,X_j)}} | D_N, \prec) \tag{7.36}$$
$$= p(X_i \notin \mathrm{pa}(X_j,G) | D_N, \prec) \prod_{l=j+1}^{n} p(X_i, X_j \notin \mathrm{pa}(X_l,G) | D_N, \prec),$$

where

$$p(X_i \notin \mathrm{pa}(X_j,G) | D_N, \prec) = \frac{\sum_{\substack{X_i \notin \mathrm{pa}(X_j,G) \\ \mathrm{pa}(X_j,G) \sim \prec}} p(D_N | \, \mathrm{pa}(X_j,G)) p(\mathrm{pa}(X_j,G) | \prec)}{\sum_{\mathrm{pa}(X_j,G) \sim \prec} p(D_N | \, \mathrm{pa}(X_j,G)) p(\mathrm{pa}(X_j,G) | \prec)}$$
$$p(X_i, X_j \notin \mathrm{pa}(X_l,G) | D_N, \prec) = \frac{\sum_{\substack{X_i, X_j \notin \mathrm{pa}(X_l,G) \\ \mathrm{pa}(X_j,G) \sim \prec}} p(D_N | \, \mathrm{pa}(X_l,G)) p(\mathrm{pa}(X_l,G) | \prec)}{\sum_{\mathrm{pa}(X_l,G) \sim \prec} p(D_N | \, \mathrm{pa}(X,G)) p(\mathrm{pa}(X_l,G) | \prec)}.$$

The summations involve a polynomial number of terms if the parental set is bounded by k. For approximations using a restricted set of parental sets with high probability, see [97].

7.6 Decision over features using MC estimates

In the previous overview of estimation methods of the posteriors of pairwise features, we ignored that the estimated feature posteriors are usually used jointly and we simplified the problem to the estimation of a single posterior. However, the number of target features can be as high as $10^4 - 10^6$ features even for a given type of pairwise features and moderate domain complexity with $100-1000$ variables. For complex features the number of feature values is exponential in the number of variables. Such a high number of feature values makes for example the manual analysis of the estimated edge posteriors intractable. It is thus a typical expectation that the MCMC method should estimate the posteriors uniformly well for all the n^2 features or over a predefined set of features rated a

priori as highly relevant. Another typical expectation in bioinformatics is that the estimates allow the correct ranking of the features or at least the selection of the most probable K feature values. These expectations indicate that the problem of the joint usage of the estimated posteriors in case of large number of features requires an additional level of analysis of the overall MCMC process. In a formal approach we will define an additional frequentist decision-theoretic level over the Bayesian layer of posteriors and their MC estimates. We analyze the effect of feature cardinality on the error of selecting the most probable features.

7.6.1 The Most Probable Features problem

We consider the case of a single complex feature with set of values $\mathcal{F}$, when the unknown feature posteriors form a single multinomial distribution $\mathcal{P} = p(F|D_N)$. The *decision problem of feature selection* includes the feature posteriors $\mathcal{P}$ as the unknown parameters, the event space consists of M (possibly dependent) samples D'_M given by a MC method $\mathcal{A}$ as a sampling distribution, and the set of actions consists of the report of the estimates and selections of the parameters. The decision rule $\delta(D'_M) = (I, \hat{\mathcal{P}}_M)$ in general can give a binary vector I indicating the selection and a scalar vector $\hat{\mathcal{P}}_M$ containing the estimates $\hat{p}_M(f|D_N)$.

If the overall estimation is important, then general distance measures such as $L_2(\mathcal{P}, \hat{\mathcal{P}}_M)$ can be adopted as loss function. However, frequently the overall estimates or rankings of the feature values are irrelevant and only the selection of feature values with high posteriors is important.

Definition 7.6.1. *The* Most Probable Features *problem (MPFs) consists of the selection of a predefined K number of feature values $f \in \mathcal{F}$ with high posteriors $p(f|D_N)$, which minimize the following loss based only on $I \in \mathcal{I}^K$ ($\mathcal{I}^K$ denotes the set of $|\mathcal{F}|$ dimensional binary vectors with exactly K ones)*

$$L(I) = L(\mathcal{P}, I) = \sum_i I_i L(s_i), \text{ where } L(s_i) = 1 - \mathcal{P}_i. \tag{7.37}$$

Note that the estimates of the selected feature values are secondary and not involved in the loss function, and with this decomposable loss function this problem is not a set selection problem. The Most Probable Features problem with the Markov Blanket subset feature generalizes the feature subset selection problem and reformulates it in the Bayesian framework. With the Markov Blanket subgraph feature it generalizes and reformulates the feature subgraph selection problem Def. 7.2.3.

7.6.2 Effect of feature cardinality in MPFs

First, assume that the MC estimates of the posteriors are available for all the feature values and let us investigate the statistical consequences of using these estimates of the feature posteriors in the most probable feature selection problem with loss Eq. 7.37. That is we neglect momentarily the computational aspects of the search of the most probable features, and the integrated estimate and

search problem. Specifically, we investigate the effect of the cardinality of the feature values $|\mathcal{F}|$ on the mean error of the selected set of features.

Theorem 7.6.1 ([189]). *Let us assume that we solve the K Most Probable Features problem in Def. 7.6.1 using an i.i.d. data set D'_M containing M samples from the feature posterior $\mathcal{P} = p(F|D_N)$ and applying the following decision rule $\delta(D'_M) = I^*_M$ defined as $I^*_M = \arg\min_{I\in\mathcal{I}^\kappa} L(\hat{\mathcal{P}}_M, I)$ (i.e., we select the most probable feature values). The loss function is defined in Eq. 7.37. Let $\hat{L}(I), \hat{L}(s_i)$ denote the corresponding estimated losses based on $\hat{\mathcal{P}}_M$, $I^* = \arg\min_{I\in\mathcal{I}^\kappa} L(\mathcal{P}, I)$ denotes an optimal set, and $I^*_M = \arg\min_{I\in\mathcal{I}^\kappa} L(\hat{\mathcal{P}}_M, I)$ denotes an empirically† optimal set. The error is defined as $1/K(L(I^*_M) - L(I^*))$. Then the sample complexity and the expected error of the selection of the K most probable features are proportional to the logarithm of the number of feature values $|\mathcal{F}|$:*

$$p(\frac{1}{K}|L(I^*_M) - L(I^*)| \geq \epsilon) \leq \delta, \text{ if } M \geq 2/\epsilon^2(\log(2|\mathcal{F}|) + \log(1/\delta)), \quad (7.38)$$

$$\mathrm{E}_{p(D'_M)}[\frac{1}{K}(L(I^*_M) - L(I^*))] \leq \sqrt{\frac{\log(2|\mathcal{F}|) + 1}{M/2}}. \quad (7.39)$$

Proof. We proceed analogously as in the case of selecting the best (binary) classifier, in fact we treat each feature value as a classifier and this theorem is the generalization of the earlier results for selecting the single best classifier [73].

$$\begin{aligned}
&\tfrac{1}{K}(L(I^*_M) - L(I^*)) \\
&= \tfrac{1}{K}(L(I^*_M) - \hat{L}(I^*_M) + \underbrace{\hat{L}(I^*_M) - \hat{L}(I^*)}_{\leq 0} + \hat{L}(I^*) - L(I^*)) \\
&\leq \tfrac{1}{K}(L(I^*_M) - \hat{L}(I^*_M) + \hat{L}(I^*) - L(I^*)) \\
&\leq \tfrac{1}{K}|L(I^*_M) - \hat{L}(I^*_M)| + |\hat{L}(I^*) - L(I^*)| \\
&\leq 2\max_{f\in\mathcal{F}} |p(f|D_N) - \hat{p}_M(f|D_N)|. \quad (7.40)
\end{aligned}$$

It means that if we can estimate uniformly well the probabilities of the features, then we can bound the error of the selected set of features. Using the Hoeffding inequality [73], we get for ϵ accuracy and δ confidence

$$\begin{aligned}
&p(\frac{1}{K}|L(I^*_M) - L(I^*)| \geq \epsilon) \\
&\leq p(\max_{f\in\mathcal{F}} |p(f|D_N) - \hat{p}_M(f|D_N)| \geq \epsilon/2) \leq 2|\mathcal{F}|e^{-M\epsilon^2/2} \leq \delta,
\end{aligned}$$

which shows that the sample complexity is

$$M \geq 2/\epsilon^2(\log(2|\mathcal{F}|) + \log(1/\delta)). \quad (7.41)$$

†We use the empirical term w.r.t. the stochastic simulations as well.

Furthermore, the expected average error of the selected set of features can be bounded as follows using the inequality $\mathrm{E}[Z] \leq \sqrt{\frac{\log(ce)}{2M}}$ (which holds if $p(Z \geq \epsilon) \leq ce^{-2M\epsilon^2}$ for all $0 \leq \epsilon$ and some $0 \leq c$) [73]:

$$\mathrm{E}_{p(D'_M)}[1/K|L(I^*_M) - L(I^*)|] \leq \sqrt{\frac{\log(2|\mathcal{F}|)+1}{M/2}}. \tag{7.42}$$

□

Note that the best K-term approximation of $\mathcal{P}$ in L_1 is the K MAP feature posterior represented by I^*.

This result was derived assuming an i.i.d. sample from the feature posterior. Analogic results for estimates based on dependent MCMC samples can be derived using MCMC variants of the Hoeffding inequality (e.g., see [117]).

7.7 Integrating estimation and search of MBGs

Until now we have assumed that the estimates of the posteriors are available for all the feature values. As discussed below this assumption is implicitly fulfilled by DAG-MC methods, but it is computationally prohibitive for ordering-based MC methods. The DAG-MC methods perform an implicit feature selection by generating a sample $D'_M = G_1, \ldots, G_M$, which can be used to construct a feature-tree containing the maximum M number of distinct feature values usually in $\mathcal{O}(Mn^2)$ time to compute the non-zero single-feature scores in $\mathcal{O}(M)$, and to select the K optimal feature values in $\mathcal{O}(M \log(K))$ time. This total $\mathcal{O}(M(n^2 + log(K)))$ time and $\mathcal{O}(Mn^2)$ space complexity is usually acceptable in practice, although the additional costs of confidence estimation methods, the extra cost of achieving convergence for features that are not part of the solution and the space requirement suggest some selection or search method to process only the promising features (ideally only the finally reported K features).

On the contrary, the issue of an integrated feature selection method within the ordering-based MC method is relevant, because an ordering-based MC method does not generate implicitly a feature set, as usually an exponential number of features are compatible with an ordering. The alternative approaches are as follows: (1) we treat estimation embedded in a search method, (2) we perform an implicit estimation by sampling, precomputing, and storing to support the subsequent search, or (3) we perform an integrated estimation and search method. We investigate these options in turn focusing on the MBG feature.

First, we consider the separation of estimation and search (cases (1) and (2)). The time complexity of the computation of the posterior of an ordering $p(\prec |D_N)$ and an ordering-conditional posterior $p(f| \prec, D_N)$ of a modular or ordering-modular feature is $\mathcal{O}(n^{k+1})$, where the effect of the real sample size in computing the likelihood terms for a parental set is $\mathcal{O}(Nk)$. We will assume that this polynomial number of scores for the parental sets (or at least for the high-scoring sets) is cached in $\mathcal{O}(n^{k+1})$ space. We also consider the advantages

of precomputing ordering-conditional factors for subsequent feature search. For example, the sets of parental sets for a fixed ordering and for a given MBG feature value $S_i(f, \prec)$ can be either completely independent of the feature value (i.e., containing all the parental sets compatible with the ordering), completely determined by the MBG value (i.e., containing the parental set specified by it) or they can be dependent on both the ordering and the MBG value. However, this last option means less than n distinct sets of parental sets for each ordering (despite the exponential number of feature value, see Eq. 7.16). This shows that in the case of MBG feature we can precompute also n ordering-conditional factors with $\mathcal{O}(1)$ computational overhead and store in $\mathcal{O}(Mn)$ space together with the $\mathcal{O}(n^{k+1})$ ordering-free parental scores and M orderings in case (2). If the search process evaluates L number of feature value in cases (1) and (2), the overall time complexities are $\mathcal{O}(LMn^{k+1})$ and $\mathcal{O}(M(n^{k+1} + Ln))$ ($\mathcal{O}(n^{k+1} + n)$ corresponds to a separate ordering-based MCMC step).

Second, now we consider the embedding of search into the estimation to overlap them computationally and to decrease the number of estimated feature values L close to the number of selected feature values K (i.e., case (3)). This is particularly relevant if K is large (i.e., it is in the range of n^k), which is the case if our goal is the construction of an offline knowledge base for exploring the MBG space. Another reason is that features that are not part of the solution cause not only extra computational costs because of the computation of their estimates, but can delay the convergence of the MCMC simulation.

In such an integrated scheme the search method at step i can be based on the sequentially refined estimates of earlier selected features and on the currently available ordering-conditional posteriors $p(F|\ \prec_i, D_N)$. By noting that the extra cost of an additional feature statistics collection is negligible (i.e., L can be increased to n^k without having significant effect), a robust strategy applies a search method on $p(F|\ \prec_i, D_N)$ for collecting high-scoring features using constraints from the earlier selected features (e.g., threshold for the score). The selected features are estimated, convergence and confidence quantities are computed (note that automated methods are necessary for convergence diagnostics, such as described in Section 2.3.1.3). If the number of features grows over a given limit L, then they are pruned to maintain efficiency and space limits. In fact this approach can be conceived as a two phased sample-then-search method with a special search method exploiting the estimation steps and using increasing prefixes of an offline sample to decrease time complexity.

The search method for finding high-probability features can be any general search such as the deterministic greedy beam search or just the sampling of the ordering-conditional posterior $p(F|\ \prec_i, D_N)$ in each step i or an overpeaked $p(f|\ \prec_i, D_N)^\alpha$ with $1 < \alpha$. Note that the goals of exploring the space of feature values and estimating their posteriors are distinct for ordering-modular features.

To develop better estimate and search methods the following observations and constructs can be exploited. First, the product form of the ordering-conditional posterior of an ordering-modular feature allows a decomposed identification of the feature with maximal posterior for a given ordering $\prec_i$.

Lemma 7.7.1. *For an ordering-modular feature function F the most probable feature value f^* compatible with a given ordering $\prec$ can be found by independent optimizations per variable using the posterior $p(F|\prec, D_N)$.*

Proof. It is the direct consequence of the existence of decomposed ordering conditional posterior

$$f^* = \arg\max_{f\sim\prec} p(f|\prec, D_N) = \arg\max_{f\sim\prec} \prod_{i=1}^{n} p(S_i(f,\prec)|\prec, D_N) \quad (7.43)$$

$$= \prod_{i=1}^{n} \arg\max_{S_i(f,\prec)} p(S_i(f,\prec)|\prec, D_N). \quad (7.44)$$

The possible special ("complementer") value without such form can be managed by appropriate summations per variable. □

Furthermore, this decomposed form allows the sorting of the set of potential parental sets $S_i(F,\prec) = \{S_i(f,\prec) : \forall f \in \mathcal{F}\}$, which allows specialized search techniques in the space of $S_1(F,\prec) \times \ldots \times S_n(F,\prec)$.

Based on this observation we introduce the following concepts.

Definition 7.7.1. *The ordering conditional (truncated) MBG space for variable Y is the most probable subspace of $S_1(\mathrm{MBG}(Y),\prec) \times \ldots \times S_n(\mathrm{MBG}(Y),\prec)$ (the truncation in each dimension and the optional sorting is discussed below).*

An MBG state is represented by an $n' \leq n$ dimensional vector $\underline{s}$, where n' is the number of variables not preceding the target variable Y in the ordering $\prec$:

$$n' = \sum_{i=1}^{n} 1(Y \preceq X_i). \quad (7.45)$$

In each dimension, the range of the values are integers $s_i = 0, \ldots, r_i$ representing either separate parental sets or a special set of parental sets not including the target variable. This special value is present only for variables after the target variable and not for the target variable. So $|S_i(\mathrm{MBG}(Y,G),\prec)|$ is $\mathcal{O}(n^k)$, which implies that f^* in Lemma 7.7.1 from the potentially exponential number of features ($\mathcal{O}(n^{nk})$)) can be found in polynomial time $\mathcal{O}(n^{k+1})$, which drops to $\mathcal{O}(1)$ extra time factor if it is done in parallel with the ordering-based MCMC simulation. The product of the ordering conditional posteriors of the represented sets of parental sets gives the ordering conditional posterior of the represented MBG state. We assume that the conditional posteriors of the represented sets of parental sets are monotone decreasing w.r.t. their indices:

$$\forall s_i < s_i' : p(s_i|D_N,\prec) \geq p(s_i'|D_N,\prec). \quad (7.46)$$

Second, in the most probable features problem the loss of the selected features in Eq. 7.37 is a sum of non-negative terms, which allows an exact (!)

prefiltering (i.e., thresholds t_i to select only the potentially optimal features). Clearly, it is enough to process features with ordering-conditional posteriors above $\tau = \text{maxK}^{\text{th}}{}_{f \in F} \hat{p}_M(f|D_N)$ (where maxK^{th} denotes the Kth value in a set in decreasing ordering), because for a feature value f part of the set of K features with maximal MC-estimate

$$\tau \leq \hat{p}_M(f|D_N) = \frac{1}{M} \sum_{i=1}^{M} p(f| \prec_i, D_N) \leq \max_{i=1,\ldots,M} p(f| \prec_i, D_N). \tag{7.47}$$

Because such a threshold τ usually is not available a priori, a sample specific threshold τ_i can be used at sample i as the following lemma shows.

Lemma 7.7.2. *If for all MCMC sample $\prec_i$ $i = 1, \ldots, M$ a feature value f is always below a threshold $\tau_i = \text{maxK}^{\text{th}}{}_{f \in F_i^{\prec}} p(f| \prec_i, D_N)/M$, then f cannot be part of the set of K features with maximal MC-estimate, because there are at least K feature with larger estimate.*

$$(\forall_{i=1}^{M} \prec_i : p(f| \prec_i, D_N) < \tau_i) \Rightarrow (\hat{p}_M(f| \prec, D_N) \leq \text{maxK}^{\text{th}}{}_{f' \in F^{\prec_j}} p(f'|D_N))$$

Proof.

$$\begin{aligned} \hat{p}_M(f|D_N) = \frac{1}{M} \sum_{i=1}^{M} p(f| \prec_i, D_N) &\leq \max_{i=1,\ldots,M} p(f| \prec_i, D_N) \\ &< \text{maxK}^{\text{th}}{}_{f' \in F^{\prec_j}} p(f'| \prec_j, D_N)/M \\ &\leq \frac{1}{M} \sum_{i=1}^{M} p(f^*| \prec_i, D_N) = \hat{p}_M(f^*|D_N), \end{aligned} \tag{7.48}$$

where $j = \text{argmax}_{i=1,\ldots,M} p(f| \prec_i, D_N)$ and f^* can be any feature in the set

$$\{f'' \in F^{\prec_j} : \text{maxK}^{\text{th}}{}_{f' \in F^{\prec_j}} p(f'| \prec_j, D_N) \leq p(f''| \prec_j, D_N)\}.$$

□

Eq. 7.48 also shows that with small variance $\text{var}_{p(\prec_i|D_N)}(p(f| \prec_i, D_N))$ the threshold factor $\frac{1}{M}$ can be selected in practice to be smaller (i.e., when the maximum value is closer to the mean).

This truncation per orderings can be specialized for ordering-modular features to truncation per orderings and variables, because of their decomposed score in Eq. 7.43. In the case of $\text{MBG}(Y, G)$ feature, this specialized filtering can guide the truncation of the MBG space as follows. We can apply the thresholds per variable j at step i with a given ordering for limiting the $\mathcal{O}(n^k)$ number of set of parental sets to $r_{i,j}$. Furthermore, these can be sorted, which means an $\mathcal{O}(r_{i,j} \log(r_{i,j}))$ extra time factor if it is done in parallel with the ordering-based MCMC simulation). This allows a uniform-cost search or a cost-limited depth-first search. A corresponding estimation and search algorithm based on the orderings and on the ordering-conditional MBG spaces is reported in Section 8.5.1.

Chapter 8

Analysis and fusion

The availability of formalized prior domain knowledge, literature and statistical data calls for an integrated analysis. We present their separate analysis, their cross-comparison for validation and discovery, and their fusion. We also report the application of new concepts and methods, such as the ordering-based MCMC over complex conditional features and the application of rank statistics, classification, causal measures, and annotations in the analysis.

The expert prior reported in Chapter 4 included parameter prior and various structure priors. The literature data and publication models reported in Sections 6.4 and 6.1 allow the reconstruction of the history of consensus beliefs. The availability of such heterogeneous information with different biases, limitations and costs poses two kinds of questions: about their differential analysis for knowledge discovery and about their fusion. We adopt the view that comparison of the sources is frequently as important as their fusion — given that it is a prerequisite for proper fusion. This was our motivation for developing many IT methods and the ABN-KB besides working on methods for the fusion of heterogeneous information.

The chapter starts with presenting methods for unified probabilistic fusion of expert prior knowledge, literature data, and medical data at the level of data, model features, and complete domain models. The next section evaluates the parameter prior mainly to assess its quality and its dependence on the prior structure. Section 8.3 summarizes work at the pairwise level with two goals: corroborate the potential of the compiled literature data in ovarian cancer and introduce the use of new quantitative measures for the evaluation of feature learning, such as rank statistics, and classification methods. Section 8.4 present results at the level of models that investigate the validity of expert structure priors as gold standard and corroborate the potential of the compiled literature data for fusion. Section 8.5 reports the learning of features, particularly the use of complex conditional features and the corresponding method. Finally, the effects of fusion are reported, particularly the effect of incorporating the expert priors and the text mined priors in Bayesian inference with medical data. The effect of priors on classification performance is reported in Chapter 10.

8.1 Fusion of expertise, literature, and data

The fusion of heterogeneous information resources, particularly the integration of electronic prior knowledge, such as knowledge bases and free-text with expertise and experimental data is of vital importance and induced many heuristic approaches. The pABN-KB defines a general framework for an integration of logical and free text prior knowledge (i.e., literature and expertise) with probabilistic prior knowledge and experimental data through the combination of model posterior and model-based probabilistic semantics (see Section 5.2 and 5.5.1). This chapter goes one step further by presenting a practically applicable Bayesian fusion of literature, experimental data and expertise based on the concept of literature data in Def. 6.1.1 and on the FTTC probabilistic model of publication in Def. 6.4.1. It results in a model posterior given the literature, experimental data and expertise, which can also be incorporated in a pABN-KB as a more refined, jointly derived probabilistic engine.

8.1.1 Fusion through linked models

The assumption of a probabilistic link between the domain model and the corresponding publication model allows the computation of the posterior over the (true) domain models given the literature data $D^L_{N'}$ as

$$p(G,\theta|D^L_{N'}) \;=\; \sum_{G^L} p(G^L|D^L_{N'}) \int_{\theta^L} p(G,\theta|G^L,\theta^L)\,\mathrm{d}p(\theta^L|G^L,D^L_{N'}), \quad (8.1)$$

or by keeping only the structures as

$$p(G|D^L_{N'}) \;=\; \sum_{G^L} p(G|G^L)p(G^L|D^L_{N'}). \quad (8.2)$$

Besides literature, the expertise can be incorporated as follows.

Theorem 8.1.1 ([26]). *In a given domain with causal models G, a real data set D_N, and sampling distributions $p(D_N|G)$, let $p(G)$ denote the expert belief, $D^L_{N'}$ denote the literature data representation of a given document collection (see Def. 6.1.1), and let G^L denote the corresponding FTTC literature Bayesian networks with its sampling distributions $p(D^L_{N'}|G^L)$ and its bijective relation $\mathcal{T}(G) = G^L$ (see Def. 6.4.1). Then the posterior is as follows*

$$p(G|D_N, D^L_{N'}) \propto\; p(G)p(D_N|G)p(D^L_{N'}|\mathcal{T}(G)). \quad (8.3)$$

If a more flexible probabilistic link $p(G^L|G)$ is allowed, then the posterior is

$$p(G|D_N, D^L_{N'}) \propto p(G)p(D_N|G)\sum_{G^L} p(D^L_{N'}|G^L)p(G^L|G). \quad (8.4)$$

Proof. We can proceed as follows by using the assumed naive Bayesian network formalization ($D_N \leftarrow G \rightarrow G^L \rightarrow D^L_{N'}$)

$$
\begin{aligned}
p(G|D_N, D^L_{N'}) &= \frac{p(D_N, D^L_{N'}|G)p(G)}{p(D_N, D^L_{N'})} && (8.5)\\
&= \frac{p(G)}{p(D^L_{N'})p(D_N|D^L_{N'})} p(D_N|D^L_{N'}, G)p(D^L_{N'}|G) && (8.6)\\
&= \underbrace{p(G)\frac{p(D^L_{N'}|G)}{p(D^L_{N'})}}_{p(G|D^L_{N'})} \frac{p(D_N|G)}{p(D_N|D^L_{N'})} && (8.7)\\
&\propto p(G)p(D^L_{N'}|G)p(D_N|G) && (8.8)\\
&\propto p(G)p(D_N|G)\sum_{G^L} p(D^L_{N'}|G^L)p(G^L|G) && (8.9)\\
&= p(D_N|G)p(G|D^L_{N'}), && (8.10)
\end{aligned}
$$

which shows that the prior is updated by the literature (data) and then by the (clinical) data in the Bayesian update scheme. Note that $p(G|D^L_{N'})$ can be conceived of a "posterior-prior", because it incorporates both the original structure prior $p(G)$ and the literature through the likelihood term $p(D^L_{N'}|G)$.

The assumption of a bijective relation between the domain model structures G and the publication model structures G^L ($\mathcal{T}(G) = G^L$) provides the posterior given the literature and possibly the clinical data as:

$$
\begin{aligned}
&p(G|D_N, D^L_{N'}) && (8.11)\\
&\quad\propto\quad p(G)p(D_N|G)p(D^L_{N'}|\mathcal{T}(G))\\
&\quad\propto\quad p(D_N|G)p(\mathcal{T}(G)|D^L_{N'}), && (8.12)
\end{aligned}
$$

showing the contributions of the literature, the clinical data, and the expertise. □

An interesting feature of this approach is that it integrates literature data and clinical data exactly (within the limits of the applied statistical natural language processing and the vector representation of the free-text). Nevertheless, it has a considerable cost on optimization or Bayesian computation due to the computation of the likelihood with literature data. Additionally, in integrated learning from heterogeneous sources, rescaling of belief for the sources is advisable to express our confidence in them.

8.1.2 Fusion through linked features

A possible solution is the approximation of the "posterior-prior" $p(\mathcal{T}(G)|D^L_{N'}, \xi)$ with the product of feature posteriors

$$
p(\mathcal{T}(G)|\xi^+) \triangleq p(\mathcal{T}(G)|D^L_{N'}, \xi) \approx \prod_i p(F_i(\mathcal{T}(G))|D^L_{N'}, \xi), \qquad (8.13)
$$

where a feature posterior $p(F_i(\mathcal{T}(G))|D^L_{N'}, \xi)$ can be seen as an approximation of the reconstructed posterior belief neglecting the parametric layer (see Eq. 6.3). Standard feature sets are the parental sets, the directed edges and undirected edges. The advantage of such feature posterior-prior is that it can be precomputed without the experimental data, analyzed, scaled if necessary (assuming interpretable features), stored offline and used as a structure prior beside the experimental data in MAP optimization or Bayesian computations without the additional run-time costs.

Further possibilities are the use of deviation priors

$$p(G|\mathcal{T}^{-1}(G^{L,MAP})), \text{ where } G^{L,MAP} = \arg\max_{G^L} p(G|D^L_{N'}), \tag{8.14}$$

and feature-deviation priors with literature-based reference model or feature posteriors (see Section 8.6 for an application).

8.1.3 Fusion of pairwise text-based scores and models

Finally we discuss the fusion of pairwise (therefore indirect!), symmetric text-based scores $R_{\text{Text}}(X_i; X_j)$ from Section 6.5, which is a model-free analysis of the literature data. As an approximation we use the prior from Eq. 3.19, which defines a prior belief for structure G based on the beliefs in direct influences w.r.t. its skeleton. Therefore $p(X_j \in \text{pa}(X_i)|\xi^+)$ (using p_{ij} as a shorthand notation) can be defined by the pairwise text scores:

$$p_{ij} \triangleq p(X_j \in \text{pa}(X_i)|\xi^+) \quad \sim \quad R_{\text{Text}}(X_i, \pi_{ik}). \tag{8.15}$$

Note that for all text scores $0 \leq R_{\text{Text}}(X_i, \pi_{ik}) \leq 1$ and that we guarantee a lower limit ϵ and an upper limit $1-\epsilon$ for all p_{ij} to avoid the a priori exclusion or inclusion of edges and consequently structures. This relative definition of edge probabilities can be refined to satisfy prior knowledge on higher-order statistics by an appropriate scaling to achieve a given expectation of the number of edges as described in Section 3.1.5.2.4.

8.2 Data-based evaluation of the small BN

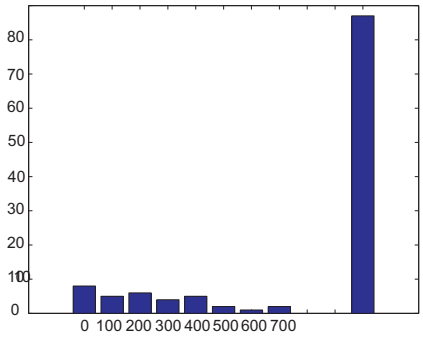
This section analyzes the elicited prior parameters, the parameter priors, and the elicitation structure based on the clinical data. The main purpose is the analysis of the quality of the parameter prior, the analysis of its hyperparameters and its sensitivity to prior transformation.

8.2.1 From prior parameters to hyperposteriors

As reported in Section 4.3.2, a domain expert specified point-valued prior parameters for a BN containing 11 variables.

Table 8.1 reports the percentage of elicited probabilities per variable in the posterior 95%, 99% and 99.9% credible regions using BD_{eu} priors and prior

virtual sample size equal to 1. The figure on the left reports the sample size of "strong divergence", from which the respective estimate in a sequential analysis is always outside of a given credible region (only for the variable ColorScore).



Variable	5%	1%	0.1%
Pathology	22.22	22.22	22.22
WallRegularity	36.67	26.67	16.67
Meno	33.33	33.33	0.00
Ascites	11.67	3.33	0.00
PapSmooth	63.33	44.44	17.78
Locularity	60.00	60.00	60.00
CA125	27.78	25.00	16.67
ColScore	10.8 3	4.17	1.67
PillUse	100.0	100.0	100.0
Bilateral	50.00	16.67	0.00
Volume	27.78	22.22	22.2

Figure 8.1: The comparison of the expert's estimates of the conditional probabilities to estimates from the data at different confidence level and with different sample sizes. (Left) The histogram of the sample sizes of "strong divergence" corresponding to the expert's probability estimates for the conditional probabilities of the ColorScore variable. Above the "strong divergence" sample size the respective estimate in a sequential analysis is always outside of a HPD region with given credibility. The rightmost column represents the percentages of the estimates without such threshold. Note that these perhaps failed in the sequential analysis temporarily. (Right) Percentage of elicited probabilities in the posterior 95%, 99% and 999% credible regions using $\mathrm{BD}_{\mathrm{eu}}$ prior.

For the Bayesian analysis we need also a parameter prior expressing the expert's confidence. In our case, the expert's experience includes the ultrasonographic examination of more than 10,000 cases [237] and it is a reasonable assumption that his prior belief over parameters can be approximated by using a single global prior virtual sample size (see Th. 3.1.6). In our earlier study with the IDO variables and data set, we estimated this value in the range of 10 to 100 partly based on the results of various prior transformations in classification using continuous variables as well (see Section 10.6). With the more specialized IOTA variables we expected its value to be comparable to the IOTA data set (i.e., in the range of 100 to 1000). Here we present a formal Bayesian inference about this hyperparameter by exactly computing its posterior $p(N'|M, D_N)$ using various domain models M and discretization schemes. Assuming a general uniform prior $p(N'|\xi)$ in $[1, 10000]$, the posterior is given by

$$p(N'|M, D_N) = \frac{p(D_N|M, N')\overbrace{p(M|N')}^{=p(M)}p(N')}{p(D_N|M)p(M)} \tag{8.16}$$

$$= p(N') \int p(D_N|\underline{\theta}, M, N')\mathrm{Dir}(\underline{\theta}|\underline{\theta}_0, M, N')\,\mathrm{d}\underline{\theta}, \tag{8.17}$$

where the integral has a closed form given in Eq. 3.34 and the hyperparameters

in Eq. 3.1.6. Fig. 8.2 shows the model log-likelihoods and the corresponding unnormalized posteriors.

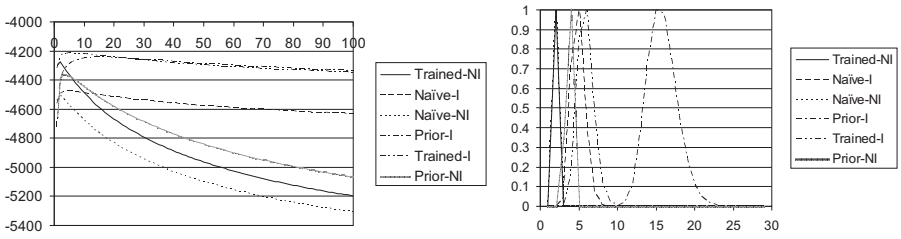


Figure 8.2: The posterior of the hyperparameter for the prior virtual sample size (i.e., the posterior belief for the count of the a priori seen samples after observing the data). The investigated Bayesian network models are composed of (1) the naive, best-inductive (Trained), and elicited prior structures and (2) noninformative — $\mathrm{BD_{eu}}$ — (NI) and informative (I) parameter priors. (Left) The model log-likelihoods in the $[1, 1000]$ interval with step size 10. (Right) The corresponding unnormalized posteriors over the virtual sample size in the $[1, 300]$ interval with step size 10 using uniform priors in $[1, 1000]$.

Because the analytic treatment of a probabilistic approach of prior virtual sample size is not possible, we adopt an approach to select an appropriate value [175]. The posterior confirms that a reasonable global prior sample size is around 150 for the prior structure, which is in the lower part of our expectation. Preliminary results indicate that it is partly the consequence of the adopted single prior sample size and that an analog posterior analysis per variable would give larger modes for their hyperparamaters except for some variables, such as PillUse or PapSmooth. The mode of the hyperparameter posterior is around 70 for a MAP structure containing these variables (see Fig. A.4) and around 50 for the respective naive BN. Note that its mode is around 40 and 20 for a non-informative uniform prior $\mathrm{BD_{eu}}$ as well. The interpretation of these results can be helped by the following intuitive explanation (beside the standard counting interpretation of the Dirichlet hyperparameters). Assuming an i.i.d. case with finite, discrete values, the prequential score is a logarithmic cumulative score summing losses until convergence equal on average to the cross-entropy corresponding to the actual estimates. Then it sums losses equal on average to the entropy. A larger prior sample size N' with its conservative bias may help to decrease the initially accumulated loss by ensuring smaller variance for the step-by-step updated estimates in the BFS, which explain the non-informative case, though it can delay the convergence. The advantageous effect of initially good estimates in a BN with local multinomial models lasts until they are not updated in all parental configuration, including configurations with small probabilities, to which a larger prior sample size can put more weight. This explains that in simpler models such as in the naive BN or in BNs optimal to the data smaller prior sample size is enough for the same bias effect.

This explanation can be also helpful to interpret the prequential comparison of models with different priors, e.g. it is informative to find the sample size from that the difference in loss accumulation disappears.

8.2.2 Evaluation of parental sets and configurations

First we report the prequential performance of the model and its decomposition to the contributions of each variable using an informative and a noninformative parameter prior, see Fig. 8.3. This shows a strong beneficial effect of the prior in general until 200 samples. The effects of the prior becomes flat after 300, although it remains significantly positive. This range is reasonable w.r.t. the complexity of the prior BN (e.g., number of parameters, number of parents, non-extremity of the conditional probabilities). The most influential variables are the ColorScore, Ascites, and CA125 (+) and PapillationSmooth (-).

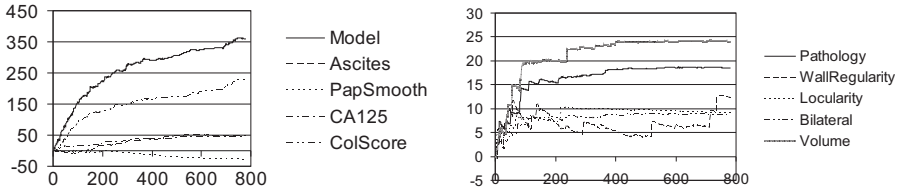


Figure 8.3: The advantage of the expert parameter estimates for each parental set per variable. The vertical axis shows the sequential predicted data log-likelihood using the expert updated estimates relative to a reference model of the updated BD_{eu} priors. The PillUse and Meno variables are not reported, because their absolute values are below 1.5. The horizontal axis gives the sample size.

Next Fig. 8.4 presents a parallel result comparing again the parental set monitors using an informative and a noninformative parameter prior with prior sample size 150 and 30 respectively, which shows again the strong positive effect of ColorScore, Ascites and CA125.

8.2.3 Evaluation of models and transformed priors

After the prequential analysis of the elicited prior parameters and parameter priors, we investigate now their transformation to other model classes, related to Chapter 10 on Bayesian classifiers with informative priors. We discuss two other BN models including the same 11 variables. A Naive BN, because of its relevance for classification and a maximum a posteriori Bayesian network over the eleven variables given the IOTA-1.2 data set (see A.4) as an objective "best" structure reference. Fig. 8.5 reports the effect of prior parameters in the Naive BN model and the performance of the Naive BN as a domain model and not as a classifier (!). It shows that the transformed parameter prior in the Naive BN has a strong beneficial effect until 100 samples and has no significant effect after

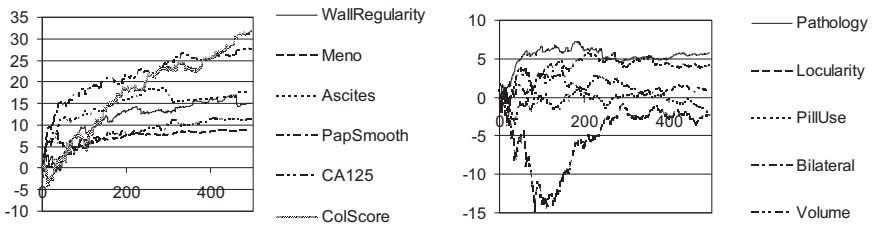


Figure 8.4: The advantage of the expert parameter estimates for each parental set per variable. The vertical axis shows the sequential predicted data log-likelihood using the expert updated estimates with the value of virtual sample size 150 relative to a reference model of the updated $\mathrm{BD}_{\mathrm{eu}}$ priors with the value of virtual sample size 30. The horizontal axis is the sample size.

200 cases. The smaller scale of the values w.r.t. original model is compatible with the different model complexity. From the point of view of the structure, the elicitation model structure is significantly better, only the Locularity variable with 3 parents needs a longer convergence period.

Finally, Fig. 8.6 reports the performances of the BN models from the combinations of the naive, best inductive, and elicited structures and of the noninformative and informative parameter priors. It shows the insufficiency of the naive BN structure and the advantage of the MAP model. Its performance is influenced by the parameter prior slightly less than the original structure, but still it determines whether it is better or not than the original structure with an informative prior. It also shows the beneficial effect of the prior parameters in each model (i.e., of the transformed parameter prior in the sense of Section 10.2.1).

8.3 Analysis of local scores

To corroborate the potential of the compiled literature data and the associative text scores $R^{L}_{\mathrm{Text}}(X;Y)$ (see Section 6.5) we compared them against pairwise data scores based on the IOTA-1.1 data set using rank statistics and classification methods[16]. We introduced similar associative data scores $R_{\mathrm{Data}}(X;Y)$ to quantify the pairwise informational relevance of X and Y. Under the assumptions that the stochastic variables are discrete and the cases in the data set are complete, a natural choice is to use the mutual information:

$$R^{\mathrm{MI}}_{\mathrm{Data}}(X;Y) \triangleq I(X;Y). \tag{8.18}$$

In the Bayesian approach a symmetric, pairwise data score $R^{\mathrm{BD}}_{\mathrm{Data}}(X;Y)$ can be defined analogously to the complete structure score (see Eq. 3.34, 3.44), which expresses the probabilities of the individual pairwise structures $p(Y \rightarrow X|D_N)$:

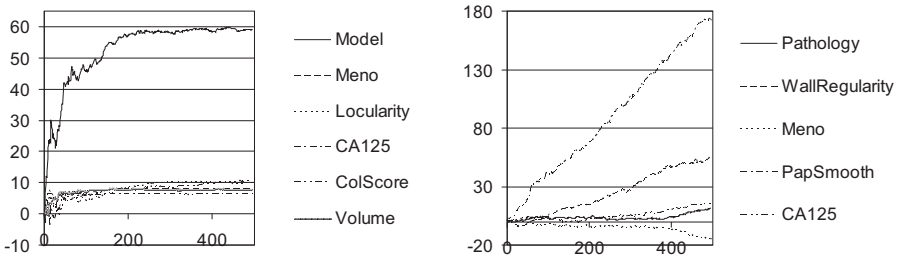


Figure 8.5: (Left)The advantage of the transformed expert parameter estimates in the naive model for each parental set per variable. The vertical axis shows the sequential predicted data log-likelihood using the expert updated estimates relative to a reference model of the updated $\mathrm{BD}_{\mathrm{eu}}$ priors. The variables Pathology, WallRegularity, Ascites, PapSmooth, PillUse, and Bilateral are not shown, because their value are less than 10. (Right) The advantage of the expert parameter estimates and original parental sets relative to the parental sets in the naive model with the transformed informative prior. The vertical axis shows the sequential predicted data log-likelihood using the expert updated estimates in the original model relative to the naive model with the transformed parameter estimates. The variables Ascites, Locularity, ColorScore, PillUse, Bilateral, and Volume are not shown, because their value are less than 10. The horizontal axis is the sample size.

$$R^{\mathrm{BD}}_{\mathrm{Data}}(X;Y) \propto \prod_{j=1}^{r_Y} \prod_{k=1}^{r_X} \Gamma(N^{YX}_{jk} + \frac{1}{r_X r_Y}). \tag{8.19}$$

Here r_X, r_Y denote the number of discrete values of variables X and Y and N^{YX}_{jk} denotes the number of times we observe value j for variable Y and value k for variable X in the data D_N.

In the analysis, we also included the expert priors $R_{\mathrm{Expert}}(X;Y)$ for pairwise relevance (see Section 4.3.3.2), though we expected it to be biased toward direct dependencies.

In general, we can characterize the R_{Expert} as an expert reference, the annotation similarity R_{ASIM} as a kind of textual expression of expert belief, the co-occurrence relation R_{COOC} as an unbiased literature relation, the corelevance relation R_{COREL} as a mixture of expert belief and literature, and finally the data scores R_{Data} as objective pairwise references.

The main goal of this analysis was to understand the characteristics and usability of the text scores in learning Bayesian networks. First we compare the constructed text scores against the expert score R_{Expert} and the data scores $R^{\mathrm{BD}}_{\mathrm{Data}}$ and $R^{\mathrm{MI}}_{\mathrm{Data}}$. In the comparison, we applied two quantitative evaluation methods: the efficiency of detecting the pairwise relations from the expert and the Spearman rank correlation.

The relation detection means that we try to find back a set of important relations (specified by the medical expert) using a score for these pairwise relations

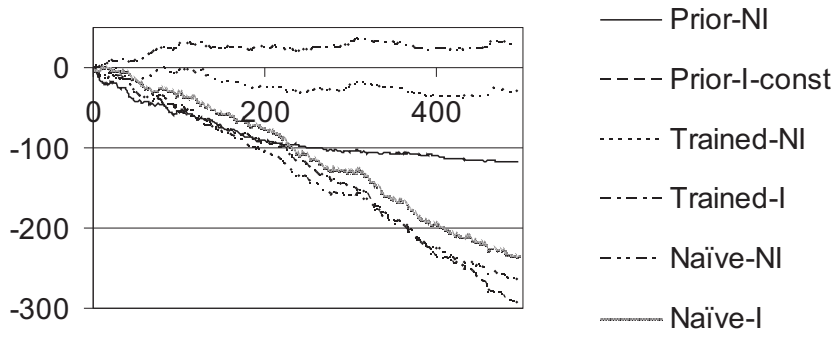


Figure 8.6: The evaluation of the combinations of (1) the naive, best-inductive (Trained) and elicited prior structures and (2) noninformative (NI) and informative (I) parameter priors. The vertical axis shows the sequential predicted data log-likelihood relative to a reference model of the elicited structure with informative priors. The horizontal axis is the sample size.

and some threshold. (We use the sets S^r and S^h as defined in Section 4.3.3, S^m is omitted for simplicity, S^r_P and S^h_P are the respective subsets containing only the relations corresponding to the variable *Pathology*.)

To quantitatively evaluate different text scores and understand their relations, we computed the Area Under the ROC curve (AUC) to detect the relevance relations identified by the expert (see Section 4.3.3.2) and the Spearman rank correlation coefficient R_S with the expert score and with the data scores. The first column of Table 8.1 shows the AUC values for detecting the S^h, S^m, and S^r relations. The second column of Table 8.1 shows the specificity values for detecting these sets corresponding to 50% sensitivity. The third column of Table 8.1 shows the sensitivity values for detecting these sets corresponding to 50% specificity. The upper triangle of Table 8.2 presents the Spearman rank correlation coefficients for all pairs of the expert score, text scores and data scores as introduced above and in Sections 4.3.3, and 6.5. Beside the Spearman rank correlation coefficients for all the relations, the lower triangle of Table 8.2 shows the Spearman rank correlation coefficients for the relations of the variable *Pathology*. Bold, underscore, and bold underscore typesettings indicate significant monotonic relationship between the ranks with $p < 0.05$, $p < 0.001$, and $p < 0.001$ respectively.

First, using the AUC values and sensitivity-specificity values from Table 8.1, we examine which of the text prior or the data can select better the relations from the expert. We expect that the domain is known enough, thus we expect the highest correspondence between the prior and the data scores (we expect the text scores to be less accurate due to the noise and bias). Surprisingly, the text scores performed better than expected; for example the $R^{MI,C3}_{\text{COREL}}$ achieved an AUC value of 82.01 and $R^{\text{MI}}_{\text{Data}}$ achieves AUC=85.95 for selecting

Table 8.1: The AUC values for detecting important expert relations using the different text scores and the data scores (S^h contains only the most important relations identified by the expert, S^r contains a broader range of relevant relations as described in Section 4.3.3, S^r_P and S^h_P are their respective restrictions to the pairwise relations involving the variable *Pathology*). The specificity column presents the specificity values corresponding to 50% sensitivity (i.e., it shows the percentage of *not* relevant relations that are correctly classified as *not* relevant when we demand that 50% of the relevant relations are correctly detected). The sensitivity column presents the sensitivity values corresponding to 50% specificity (i.e., it shows the percentage of relevant relations that are correctly detected when we allow only 50% of the *not* relevant relations to be incorrectly classified as relevant). In each column, the three best values are indicated with bold.

	Area under the ROC curve (%)				Specificity (%)		Sensitivity (%)	
Settings	S^r	S^h	S^r_P	S^h_P	S^r	S^r_P	S^r	S^r_P
$R^{\mathrm{MI},C_3}_{\mathrm{COREL}}$	**82.01**	**93.24**	**78.26**	**95.83**	**90.43**	71.43	**90.74**	**82.61**
$R^{\mathrm{MI},C_0}_{\mathrm{COREL}}$	75.17	88.79	68.32	**91.67**	80.53	71.43	86.42	73.91
$R^{\mathrm{AND},C_3}_{\mathrm{COREL}}$	**82.10**	92.68	**78.26**	79.17	89.44	**100.00**	**90.74**	**82.61**
$R^{\mathrm{AND},C_0}_{\mathrm{COREL}}$	75.71	89.86	67.70	**90.97**	81.85	71.43	83.95	**82.61**
$R^{\mathrm{MI},C_3}_{\mathrm{COOC}}$	61.61	66.95	54.04	37.50	73.27	71.43	66.05	52.17
$R^{\mathrm{MI},C_0}_{\mathrm{COOC}}$	64.95	72.05	65.84	42.36	81.52	85.71	72.84	73.91
$R^{\mathrm{AND},C_3}_{\mathrm{COOC}}$	67.36	68.70	60.87	39.58	84.49	71.43	76.54	65.22
$R^{\mathrm{AND},C_0}_{\mathrm{COOC}}$	64.58	72.15	63.35	42.36	73.60	85.71	69.75	69.57
R_{ASIM}	65.83	88.48	75.78	88.89	80.20	**100.00**	67.28	69.57
$R^{\mathrm{BD}}_{\mathrm{Data}}$	75.99	**95.64**	**91.30**	75.69	**94.39**	**100.00**	77.16	**91.30**
$R^{\mathrm{MI}}_{\mathrm{Data}}$	**85.95**	**97.53**	**93.17**	72.92	**94.72**	**100.00**	**93.21**	**91.30**

the S^r relations. Although the data scores are slightly better, the differences are not statistically significant. The opposite behavior of S^r_P is investigated below. Another unexpected result is that $R^{MI,C3}_{\mathrm{COREL}}$ outperforms the R_{ASIM} relation (AUC=65.83), although the corelevance methods is a mixture of experts belief and literature, while the annotation similarity is closer to the expert belief.

Second, we examine the effect of increasing the size of the document collection from C_3 to C_0, which basically means a broader scope with less domain specificity and thus a higher noise level. As Table 8.1 shows, the (name) co-occurrence-based scores perform better on a larger collection—that is, they gain more from the larger number of publications than they lose from the fact that the documents are less domain-specific (e.g., AUC=61.61 for C_0 versus AUC=64.95 for C_3 of the $R^{\mathrm{MI}}_{\mathrm{COOC}}$ for the set S^r). This is probably caused by the scarcity of names (i.e., the lack of a nomenclature). Conversely, the corelevance methods perform better on the smaller, more specific collection C_3 (e.g., AUC=75.17 for $R^{\mathrm{MI},C_0}_{\mathrm{COREL}}$ versus AUC=82.01 for $R^{\mathrm{MI},C_3}_{\mathrm{COREL}}$ for the S^r set). It means that the vector representation and the applied relevance measure cannot cope with the broader scope of the corpus, while is still much better than the simpler co-occurrence methods.

Third, we examine the effect of detecting the "most relevant" relations S^h and all the relevant relations S^r. As we expected, the "most relevant" relations are more easy to identify for all the text scores and data scores in the case of S^r

Table 8.2: The Spearman rank correlation coefficients for the cross-comparison of the expert score, the text scores, and the data scores (because of symmetry, the upper triangle presents the coefficients for comparing all the relations and the lower triangle presents the coefficients for comparing the relations related to the variable *Pathology*). The level of significance is indicated by underscore ($p < 0.05$), bold ($p < 0.001$), and bold underscore ($p < 0.001$).

settings	$R^{\mathrm{MI},C_3}_{\mathrm{COREL}}$	$R^{\mathrm{MI},C_0}_{\mathrm{COREL}}$	$R^{\mathrm{MI},C_3}_{\mathrm{COOC}}$	$R^{\mathrm{MI},C_0}_{\mathrm{COOC}}$	R_{ASIM}	$R^{\mathrm{BD}}_{\mathrm{Data}}$	$R^{\mathrm{MI}}_{\mathrm{Data}}$	R_{Expert}
$R^{\mathrm{MI},C_3}_{\mathrm{COREL}}$		**0.726**	0.101	**0.111**	**0.508**	**0.385**	**0.408**	**0.507**
$R^{\mathrm{MI},C_0}_{\mathrm{COREL}}$	**0.787**		0.028	0.081	**0.555**	**0.363**	**0.346**	**0.413**
$R^{\mathrm{MI},C_3}_{\mathrm{COOC}}$	-0.042	-0.117		**0.766**	-0.022	**0.139**	**0.193**	**0.175**
$R^{\mathrm{MI},C_0}_{\mathrm{COOC}}$	0.021	0.003	**0.684**		0.035	**0.179**	**0.268**	**0.237**
R_{ASIM}	**0.672**	**0.677**	-0.109	-0.006		**0.427**	**0.271**	**0.297**
$R^{\mathrm{BD}}_{\mathrm{Data}}$	0.572	0.473	0.010	0.160	0.541		**0.629**	**0.471**
$R^{\mathrm{MI}}_{\mathrm{Data}}$	0.513	0.439	0.037	0.223	0.534	**0.968**		**0.546**
R_{Expert}	**0.627**	0.527	-0.119	0.009	0.537	**0.640**	**0.650**	

versus S^h. It also holds for the data scores, which means that on average the expert score is in close correspondence with what the data says. Interestingly, this trend is mixed in the case of the relations including variable *Pathology* (S^r_P versus S^h_P), in which the data scores are less effective to select the most relevant variables than a broader scope of related variables. A preliminary evaluation has shown that the expert ranking of certain factors as "most relevant" and "moderately relevant" is responsible for this, for example the top-rated papillation related variables were rated lower by the data. Furthermore, the co-occurrence scores R_{COOC}, which can be seen as objective literature scores beside the objective data scores, are similarly less effective to select the most relevant variables than to select the broadest scope of variables. Note that this is not the case for the annotation-based score R_{ASIM}, which reflects the expert's textual ranking. However, in a detailed analysis of the ranking of the expert, data and literature, the limitations of the pairwise approach should be taken into consideration also, because the variables are strongly dependent — which makes it difficult for the expert to select pairwise relations.

Finally, we examined the effect of using the mutual information (MI) and the co-occurrence (AND) formulas. Because the name co-occurrence method in our domain is prone to generating extreme relations (i.e., with uncommon variable names that never occur), the corelevance method is more appropriate for this investigation, but as Table 8.1 illustrates we could not find a significant difference or qualitative difference along this dimension.

The other quantitative method for the comparison of the scores is the comparison of the correspondence of their ranking by the Spearman ranking coefficient R_S (note that the scaling of the scores defined in Eq. 3.20 is monotonic, so does not influence ranking). Table 8.2 presents all the cross-comparisons, both for all of the relations and for only the *Pathology* relations (the AND options are not shown for simplicity, because they are not different from the MI case). From Table 8.2, we can conclude that the expert score R_{Expert} is significantly, strongly rank correlated with the data scores, so its reference status is corroborated (see

Equation 8.20). Similarly, the text scores, more specifically the corelevance R_{COREL} and the annotations similarity R_{ASIM} are significantly, strongly rank correlated with the prior and somewhat weakly with the data scores (see Equation 8.21). Furthermore, the corelevance R_{COREL} and the annotation similarity R_{ASIM} are really better rank correlated with the expert score than with the "objective" literature-based co-occurrence score (see Equation 8.22). However, contrary to our expectations, the corelevance relation R_{COREL} outperforms the annotation similarity R_{ASIM} (see Equation 8.22), which indicates that the annotations does not reflect completely the expert prior and can be refined in this respect using the literature by the corelevance method. Finally, the R_{COREL} score is strongly rank correlated with R_{ASIM} but not with R_{COOC}, and similarly R_{ASIM} is not rank correlated with R_{COOC} (see Equation 8.23). This conclusions can be grouped and summarized as follows (by indicating the strength of a rank correlation in increasing order with $\sim$, $\approx$, $\simeq$, and $\cong$):

1. *Quality of expert prior.* The expert score R_{Expert} strongly rank correlates with the data scores $R^{\text{BD}}_{\text{Data}}$ and $R^{\text{MI}}_{\text{Data}}$:

$$R^{\text{BD}}_{\text{Data}} \cong R^{\text{MI}}_{\text{Data}} \text{ and } R_{\text{Data}} \simeq R_{\text{Expert}}. \tag{8.20}$$

2. *Quality of text-based priors.* The text scores R_{Text} strongly rank correlate with the expert score R_{Expert} and weakly with the data scores R_{Data}:

$$R_{\text{Text}} \simeq R_{\text{Expert}} \text{ and } R_{\text{Text}} \approx R_{\text{Data}}. \tag{8.21}$$

3. *Subjectivity of text scores.* The annotation-based score R_{ASIM} is the most subjective (i.e., closest to the expert prior R_{Expert}):

$$R_{\text{ASIM}} \simeq R_{\text{Expert}},\ R_{\text{COREL}} \approx R_{\text{Expert}}\ \text{, and } R_{\text{COOC}} \sim R_{\text{Expert}}. \tag{8.22}$$

   In other words, the hybrid corelevance method R_{COREL} is between the expert (subjective) R_{ASIM} and the literature (objective) R_{COOC}:

$$R_{\text{COOC}} \sim R_{\text{ASIM}},\ R_{\text{COOC}} \approx R_{\text{COREL}} \text{ and } R_{\text{ASIM}} \approx R_{\text{COREL}}. \tag{8.23}$$

8.4 Analysis at the model level

After the comparison of the local scores, we continue with the model level analysis. We discuss questions of compatibility of expert priors (w.r.t. clinical data), sufficiency of clinical data for learning complete domain models (w.r.t. expert prior), and validity of literature BNs (w.r.t. data and expert priors). Again, as with local scores, we try to cross-validate the expert priors and data, to determine references (e.g., by selecting between G^H, G^M, and G^R expert structures), and to use them for evaluation of literature models.

8.4.1 Structure priors vs. clinical data

We investigate the compatibility of the clinical data and structural priors, particularly the multiparental priors (see comments on the relation of domain and multiparental relations in Section 4.3.3.1). Our goal is twofold: to explore the validity of the structure priors (e.g., by estimating confidence based on the sample size), and to explore the sufficiency of the data w.r.t. learning multiparental relations and domain models over this set of variables.

In Section 8.2.2 we already applied the *node monitors*, specifically the *mechanism monitor* to track the performance of a parental set. Because of the availability of a total ordering, we use again this monitor with two extensions. First, taking advantage of the polynomial-time computable posterior for a modular feature given an ordering and a threshold over the parental set size, we investigated ordering conditional posteriors with uniform priors (i.e., normalized likelihood). Fig. 8.7 reports the sequential ordering-conditional posteriors of the parental sets in G^H.

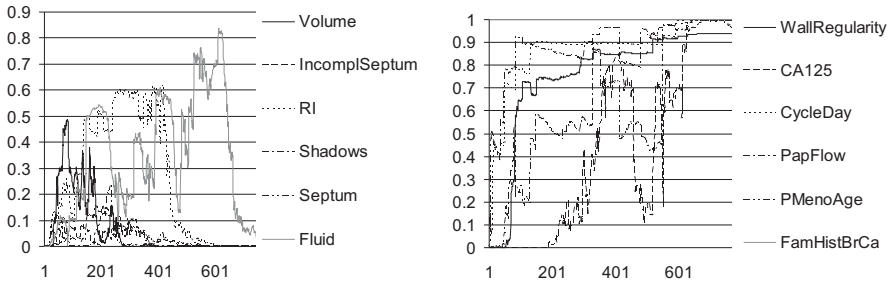


Figure 8.7: The temporal evolution of the belief in the local models (i.e., parental sets) in the expert's G^H Bayesian network; inferred from growing amount of clinical data (IOTA-1.2), $\mathrm{BD}_{\mathrm{eu}}$ prior, uniform structure prior, and given the expert's total causal ordering. The majority of posteriors are highly varying below 200 samples, but most of them are stabilized above 500 samples. (Left) The sets (i.e., children) with posteriors below 0.05 for the complete IOTA-1.2 data set (the variables Meno, Hysterectomy, PapSmooth, Solid, TAMX, PI, ColScore, and PillUse, Parity, Locularity, Echogenicity, Shadows are omitted, because their values are less than 0.05 for all sample sizes and for sample sizes larger than 200). (Right) The sets (i.e., children) with posteriors above 0.95 (the variables ReprYears, Famhist, Age, and PostmenoY are omitted, because their values are larger than 0.95 for sample sizes larger than 200). The vertical axis shows the posterior, the horizontal axis is the sample size.

Second, taking advantage of the ABN-KB we defined "knowledge-based" modular features F_i (called "ABN-node-monitor") for each variable X_i based on the multiparental and pairwise prior relations. We found the following features (expressions) particularly useful

$$F_i(\mathrm{pa}) \quad = \quad c \geq (|\Delta(S_i^{H|M|R}, \mathrm{pa}) \cap S_i^{h|m|r|n}|), \tag{8.24}$$

where c denotes an arbitrary threshold. Fig. 8.4.1 reports the sequential posteriors of more than one variable difference in the parental sets of the variables in the expert's G^M model given the expert's total causal ordering.

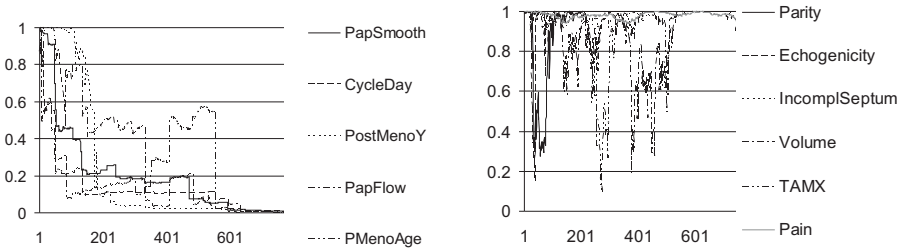


Figure 8.8: The temporal evolution of the posteriors of more than one variable difference in the parental sets of the variables in the expert's G^M model. The posteriors are computed with $\mathrm{BD}_{\mathrm{eu}}$ priors, with noninformative structure priors and conditionally on the expert's total ordering. The majority of posteriors are highly varying below 200 samples, but most of them are stabilized above 500 samples. (Left) The sets (i.e., children) with posteriors below 0.05 for the complete IOTA-1.2 data set (the variables ReprYears, Famhist, and FamhistBrCa are omitted, because their values are less than 0.05 for sample sizes larger than 200). (Right) The sets (i.e., children) with posteriors above 0.95 (the variables PillUse, Pathology, Locularity, PI, ColScore, Bilateral, CA125, Fluid, RI, PSV, Septum, HormTherapy, Shadows and Meno, Hysterectomy are omitted, because their values are larger than 0.95 for all sample sizes or for sample size larger than 200).

Note that such ABN-node monitors can be combined into a semantic model (global) monitor, expressing for example the posterior probability of deviation smaller than a specified threshold for all the nodes from the reference structure.

The usage of posteriors of such complex statements is an exact and informative method for evaluating the compatibility of the data and the prior knowledge, but it requires either strong assumptions or computational resources. In a simplified approach we can investigate their compatibility by comparing only a MAP BN against the prior knowledge. We defined and applied the following scoring function based on the prior pairwise edge rating $S^{h/m/r/n}$

$$L_{KB}(\hat{G}) = \sum_i \lambda_i \mathrm{EdgeDiff}(G_i, h|m|r|n, \mathrm{Logical}, \mathrm{Orientation}, \mathrm{DiffType}), \quad (8.25)$$

where the EdgeDiff() function returns the number of the edges that have a given h, m, r, n status and *logical* pairwise status in the ABN-KB KB and differ in $\hat{G}$ and G_i (where G_i are reference structures, such as the $G^{H/M/R}$ structures from Section 4.3.3.1). It can respect the orientation and the type of the difference (+/-) depending on the setting of bOrientation and DiffType. Fig. 8.9 shows the learning curve of the edge difference between clinical data-based maximum

a posteriori Bayesian networks and the expert overall G^M Bayesian network model w.r.t. the $S^{h/m/r/n}$ rating.

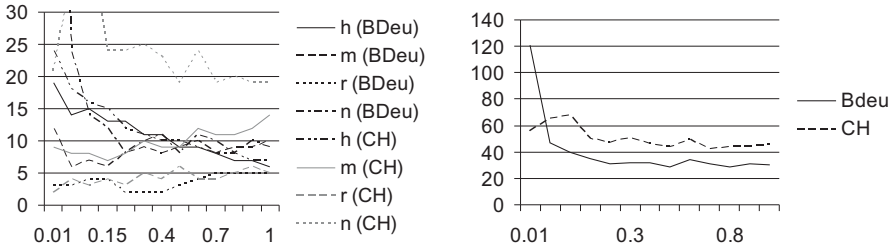


Figure 8.9: The temporal evolution of the typed $h|m|r|n$ (left) and overall (right) edge-differences between the expert overall G^M Bayesian network model and clinical data-based maximum a posteriori Bayesian networks labelled as BDeu and CH. The MAP Bayesian network was trained with noninformative $\mathrm{BD_{eu}}$ and CH parameter priors, uniform structure priors, and using the IOTA-1.2 data set. The decomposition by types shows that most of differences are rated as not relevant ones, and only the relations rated as highly relevant have monotonically improving scores.

Another type of semantic comparison of a prior structure and a MAP BN uses the causal interpretation of the BNs. It is based on the comparison of the *Causal edge (E)*, *Causal path (P)*, *(Pure) Confounded (Conf)*, and *Independent (I)* pairwise relations in the models (see Section 3.1.3.4). For example, Table 8.3 shows these differences between the G^M and a MAP network in a matrix containing the number of relations of a given type in the models.

Table 8.3: Detailed causal comparison of prior and data based BNs using the *Causal edge (E)*, *Causal path (P)*, *(Pure) Confounded (Conf)*, and *Independent (I)* pairwise relations. The most important differences are between the I vs. Conf. vs. P/E. We used the $D^{PM^H_R}$ data set, the $\mathrm{BD_{eu}}$ parameter priors and noninformative structure priors and exhaustive search to 3 parents with K2 greedy continuation over 10^6 random ordering.

	I	Conf	P	E
I	6	2	0	4
Co	0	66	50	16
P	54	486	336/0	50/0
E	0	10	40/0	70/0

Scalar scores can be derived from this matrix similarly by summing the elements with different weights as in Eq. 8.25 [58, 258], for example a simple acasual indicator is the number of extra and missing pairwise independencies ($I+/I-$). Table 8.4 reports the typed and causal differences between the prior G^H, G^M, and G^R and the MAP BNs based on clinical data.

The experiments providing these results confirmed that the G^H network is

Table 8.4: Typed and causal differences between the prior G^H, G^M and G^R and the clinical data based MAP BN $\hat{G}^{\prec}_{CH}$ counting the extra(+) and missing (-) edges with $S^{h/m/r/n}$ rating and the difference w.r.t. *Independent (I)* pairwise relation. We used the CH parameter prior, uniform structure priors and exhaustive search to 5 parents with K2 greedy continuation given the total ordering (see A.5). The last three rows shows the undirected edge differences excluding the pairs with logical relations.

	h+	m+	r+	n+	h-	m-	r-	n-	I+	I-
G^H	8	15	3	24	3	0	0	0	540+472	0
G^M	1	7	3	24	6	18	1	0	194+140	0
G^R	1	7	0	24	6	22	19	0	4+8	0+6
G^H	1	3	2	32	3	0	0	0	-	-
G^M	0	1	2	25	6	18	1	0	-	-
G^R	0	1	0	25	6	22	19	0	-	-

Table 8.5: Typed and causal differences between the prior G^H, G^M and G^R and the clinical data based MAP BN $\hat{G}_{CH}$ counting the extra(+) and missing (-) edges with $S^{h/m/r/n}$ rating and the difference w.r.t. *Independent (I)* pairwise relation. We used the CH parameter prior, uniform structure priors and exhaustive search to 3 parents with K2 greedy continuation using 10^6 random orderings (see A.6). The last three rows shows the undirected edge differences excluding the pairs with logical relations.

	h+	m+	r+	n+	h-	m-	r-	n-	I+	I-
G^H	6	13	4	23	6	0	0	0	544+468	0
G^M	0	7	4	23	10	20	1	0	192+142	0
G^R	0	7	0	23	10	24	18	0	4+8	0+6
G^H	1	2	2	28	3	0	0	0	-	-
G^M	0	1	2	20	6	19	1	0	-	-
G^R	0	1	0	19	6	23	18	0	-	-

a sound, but incomplete reference structure. The G^M and the G^R network can provide valuable prior information, but their use as a gold standard for evaluating learning methods requires caution, because none of them can be reconstructed exactly. However the manual investigation exposed that the structural prior is more reliable w.r.t. the classification aspects in the clinical diagnostics, which further supports its use as a reference in Bayesian learning of complex conditional features. In conclusion, neither these structure priors nor MAP domain models based on some part of the clinical data can serve as an exclusive gold standard, but both can provide a point of view for evaluation. A reasonable choice is the selection of the pair of the G^M or G^R prior structure and the $\hat{G}_{\mathrm{BD_{eu}}}$ MAP model.

Finally, we discuss the validity of our general limit of the parental set size (4), which is a high-level structural prior fundamental in both optimization and MC methods. This is smaller than five in the G^M network and the MAP structures have not refuted it. We tested this structural assumption against the clinical data in the Bayesian framework as well by computing the posterior distribution

Table 8.6: Typed and causal differences between the prior G^H, G^M, and G^R and the MAP BN $\hat{G}_{\mathrm{BD_{eu}}}$ based on clinical data counting the extra(+) and missing (-) edges with $S^{h/m/r/n}$ rating and the difference w.r.t. *Independent (I)* pairwise relation. We used the $\mathrm{BD_{eu}}$ prior, uniform structure priors, and exhaustive search up to 3 parents with K2 greedy continuation using 10^6 random orderings (see A.7). The last three rows show the undirected edge differences excluding the pairs with logical relations.

	h+	m+	r+	n+	h-	m-	r-	n-	I+	I-
G^H	6	8	3	18	5	0	0	0	588+352	0
G^M	0	4	3	18	9	22	1	0	200+62	0
G^R	0	4	3	18	9	22	1	0	4+2	0+72
G^H	1	1	2	19	2	0	0	0	-	-
G^M	0	0	2	12	5	20	1	0	-	-
G^R	0	0	2	12	5	20	1	0	-	-

over the parental set sizes for each variable given the total ordering, which also confirmed the validity of this bound.

8.4.2 Evaluating literature models

The evaluation of the literature data based on the domain model and the application of BNs involves many options. The selection of the source and derivation of the literature data (i.e. the selection of the corpus ME or PM possible restricted by the H, M, R journal rating and the selection of the co-occurrence or co-relevance method with a threshold); the selection of reference structures based on the prior knowledge or selection of a setting for learning a BN using the clinical data set; the selection of a setting for learning a BN using the literature data set. We analyzed most of the combinations, where the selection was partly influenced by the experiments in Section 8.3 (i.e., focused corpus with the corelevance method) and in Section 8.4.1 (the G^M and G^R prior structures) and partly by the result of these manual exploration. We will report results under the following conditions that reflect the characteristics of the results and show the effects of the narrower or wider corpus, parameter prior, total ordering, and expert or data reference. We will use the corelevance method with a threshold of 0.01, the more focused PM corpus and optionally the highly relevant journal filter. As a reference we will use the G^M prior structure and a MAP BN ($\hat{G}_{\mathrm{BD_{eu}}}$ based on clinical data, see A.7). For learning literature BNs, we use the prior total ordering in learning from literature data $D^{PM_R^R}$ and from $D^{PM_R^H}$ (in the later case with both CH and $\mathrm{BD_{eu}}$ parameter priors). There is no ordering constraint in one case when learning from $D^{PM_R^H}$. The models are respectively $\hat{G}^{\prec}_{\mathrm{BD_{eu}}}(D^{PM_R^R})$ (Fig. A.8), $\hat{G}^{\prec}_{\mathrm{BD_{eu}}}(D^{PM_R^H})$(Fig. A.9), $\hat{G}^{\prec}_{CH}(D^{PM_R^H})$(Fig. A.10) and $\hat{G}_{\mathrm{BD_{eu}}}(D^{PM_R^H})$(Fig. 6.2). Table 8.7 reports the typed and causal comparison of these reference structures and the literature BNs.

These results show that the difference between the literature models and the prior G^M model is considerably larger than that of models based on clinical

Table 8.7: Typed and causal comparison of literature based BNs against the prior structure G^M (first 4 rows) and a MAP structure based on clinical data (see A.7). The counts of the extra(+) and missing (-) edges with $S^{h/m/r/n}$ rating and the difference w.r.t. *Independent (I)* pairwise relation are reported.

	h+	m+	r+	n+	h-	m-	r-	n-	I+	I-
$\hat{G}^{\prec}_{\mathrm{BD_{eu}}}(D^{PM^R_R})$	3	12	6	38	7	19	1	0	196+198	0
$\hat{G}^{\prec}_{\mathrm{BD_{eu}}}(D^{PM^H_R})$	0	5	3	33	10	21	1	0	144+238	0+10
$\hat{G}^{\prec}_{CH}(D^{PM^H_R})$	0	2	8	15	14	20	1	0	60+118	128+86
$\hat{G}_{\mathrm{BD_{eu}}}(D^{PM^H_R})$	0	2	2	46	24	21	1	0	216+116	24+14
$\hat{G}^{\prec}_{\mathrm{BD_{eu}}}(D^{PM^R_R})$	10	14	4	35	5	3	1	15	24+108	0
$\hat{G}^{\prec}_{\mathrm{BD_{eu}}}(D^{PM^H_R})$	7	9	2	30	8	7	2	15	22+104	0+16
$\hat{G}^{\prec}_{CH}(D^{PM^H_R})$	5	7	7	12	10	7	2	15	18+54	264+106
$\hat{G}_{\mathrm{BD_{eu}}}(D^{PM^H_R})$	2	5	2	43	17	6	3	15	74+48	70+20

data, the number of different edges is roughly twofold. However the decomposition of the different edges shows that the errors w.r.t. the typed h, m, r pairwise relations are comparable for the BNs based on clinical data and literature data, which is in line with our expectation that these relations are better reported and represented in the literature. Another interesting feature is that the differences between the literature BNs and the G^M model are comparable to their differences against the clinical data based BN. This is compatible with the conclusion from the pairwise investigation of the prior, clinical data and the literature data (see Section 8.3) that the expert priors, the clinical data and the literature with the derived models and relations reflect three different points of view of the domain, though they are inevitably linked, particularly w.r.t. the basic, already established relations. This excludes the usage of one of them as a gold standard to validate the others and strengthens our view that our knowledge rich analysis (focusing on complex model properties and full-scale probabilistic fusion) is necessary.

8.5 Feature learning

In Section 7.7, we overviewed many approaches to estimate and find complex BN features. For complex features for classification, we proposed the use of the MBG space within ordering-based methods. The ordering and the ordering conditional MBG spaces can be used independently by deterministic heuristic searches and by MC methods. Fitting to the OC domain we will report the following combination: (1) the reference total ordering or MCMC method over the unconstrained orderings and (2) heuristic search in the MBG space. Besides this we report results about simple pairwise features as well.

8.5.1 An estimation and search method for MBGs

The usage of ordering for learning BN features relies on the polynomial-time approximation of the ordering posterior and on the polynomial-time approximation of the ordering conditional posterior for ordering-modular BN features discussed in Section 7.5.2. The usage of ordering means an overlapping clustering of the DAG space and an analytic solution in each cluster corresponding to an ordering (i.e., over the DAGs compatible with a given ordering). An expectable effect of the smoother posterior of the orderings and the analytic ordering conditional posterior are better convergence and confidence properties for the MC methods (see Section 7.5.2.2). Another advantage is that, in case of an ordering-modular complex feature, its ordering-conditional posterior can be used in search methods for finding high-scoring values both to guide the integrated search-estimate method and to approximate the ordering conditional posterior of non ordering-modular features.

Furthermore, for ordering-modular BN features, specifically for complex classification BN features in Section 7.7, we proposed the use of the ordering conditional MBG space 7.7.1. The purpose of the MBG space is twofold: selecting high-scoring MBGs for estimation and updating the estimates of already selected MBGs.

In the ordering space we will use the following options. The total ordering is used to generate sequential results (i.e., to compute the posterior for increasing size of the data, to compute a baseline using all the clinical data, and to calibrate the settings of the methods working in the MBG space). The prior partial orderings and the informative priors over the orderings are not used. For the unconstrained orderings a general purpose MCMC method is used.

The overview of the method implemented is shown in Alg. 1. This algorithm is aimed at solving the Feature Subset Selection problem, the Feature subGraph Selection problem, and the Most Probable Features problem with the Markov Blanket subset feature, the Markov Blanket subgraph feature (see Def. 7.2.3 and Def. 7.6.1). Besides searching and estimating MAP MBG values for a given variable, it also estimates the posteriors of simple conditional features, such as edge and MBM relations, the MB feature of the given variable, and the posteriors of prespecified ABN-KB sentences. The method can be parameterized to use different training proportions of the input data set with various averaging schemes, so it provides a *Bayesian, four-level, sequential relevance analysis* at the levels of Markov Blanket Memberships, Markov Blanket sets, Markov Blanket graphs, and complete Bayesian networks.

For notational simplicity only the MCMC case is shown, but the implemented method includes deterministic enumeration methods and the importance sampling over the orderings as well.

Alg. 1 on the one hand reflects the relative independence of the search and the estimation process (e.g., without the expand option it simply estimates the posteriors of the a priori specified MBGs). But on the other hand it also shows their overlap (e.g., the selected high-scoring MBGs can be used for expanding the set of the estimated feature values and for restricting the computationally

Algorithm 1 Ordering based search and MC estimation of conditional features

Require: [ordering prior]set of allowed orderings;
Require: [noninformative structural prior $p(G|\prec,\xi^{-})$]]limit of the maximum parental set size k, uniform parental prior, "uniform over and within sizes" parental prior;
Require: [informative structural prior $p(G|\prec,\xi^{+})$] a priori excluded/included edges and a priori edge probability matrix;
Require: [parameter prior $p(\underline{\theta}|G)$] prior point specification $\underline{\theta}_0$ and prior virtual sample size N';
Require: [targets] the target variable Y, the set of a priori specified MBGs S^{MBG}, the set of ABN-KB sentences S^{α};
Require: [targets]the number of the most probable MBGs and MBs to be reported K, K'
Require: [settings] $R, \rho, L^S, \rho^S, L^U, \rho^U$,bExpand,bPartialUpdate,$L^T$,$M$;
Require: [learning curve] training proportion and averaging scheme;
Ensure: Estimates of the posteriors of directed and undirected edges, MBM relations, and the ABN-KB sentences in S^{α};
Ensure: K MAP MBGs with their estimates;
Ensure: K' MAP MBs with their estimates;
Ensure: MCMC convergence and confidence estimates for the posteriors of the elements in S^{BN}, S^{MBG}, S^{MB};
Cache ordering-free parental posteriors $\Pi = \{\forall\ i, |\operatorname{pa}(X_i)| < k : p(\operatorname{pa}(X_i)|D_N)\}$
Initialize MCMC, the MBG-tree $\mathcal{T}$, MBM and edge posterior matrices $\mathcal{R}$, $\mathcal{E}$;
Insert the induced a priori MBGs in $\mathcal{T}$ and store them in a set S^{MBG};
Store the a priori specified BN and MB sets S^{BN}, S^{MB};
for $l = 0$ to M **do** {the sampling cycle}
  Draw next ordering using the "flip-flop" and "cutting" operators;
  Cache ordering specific common factors Ψ
  $p(|\operatorname{pa}(X_i)| \leq k| \prec_l)$ for all X_i
  $p(Y \notin |\operatorname{pa}(X_i)| \leq k| \prec_l)$ for $Y \prec_l X_i$;
  Compute edge posteriors $p((\rightarrow X_i, X_j)|D_N)$ and update $\mathcal{E}$;
  Compute pairwise relevances $p(\mathrm{MBM}(X_i, X_j)|D_N)$ and update $\mathcal{R}$;
  Compute $p(\prec_l |D_N)$;
  if bExpand **then**
    Construct ordering conditional MBG-Subspace(Π, Ψ, R, ρ)=Φ
    if bPartialUpdate **then**
      S^S, S^U=Search(Φ,L^S, ρ^S,L^U, ρ^U);
    else
      S^S=Search(Φ,L^S, ρ^S);
    for all mbg $\in S^S$ **do**
      if mbg $\notin \mathcal{T}$ **then**
        Insert($\mathcal{T}$, mbg) with counter n(mbg)=0;
    if $L^T < |\mathcal{T}|$ **then**
      $\mathcal{T}$=PruneToHPD($\mathcal{T}$,L^T);
  for all mbg $\in \mathcal{T}$ **do**
    increase counter n(mbg)=n(mbg)+1;
    if $\neg$ bPartialUpdate or bPartialUpdate and mbg $\in S^U$ **then**
      $\hat{p}(\mathrm{mbg}\,|D_N) += p(\mathrm{mbg}\,|\prec_l, D_N)$;

costly update of the estimates). This is supported by the expand and partial-update options (bExpand, bPartialUpdate) and their independent parameters L^S, ρ^S and L^U, ρ^U regulating the construction of the sets for expansion and update (S^S, S^U).

The identification of high-scoring MBGs is based on the observation that the ordering conditionally MAP MBG can be found in $\mathcal{O}(n^{k+1})$ time with a negligible constant increase only. Similarly, there is only a $R\log(R)$ extra multiplicative factor for the construction of the MBG-space including the maximum R values for the n' dimension representing the most probable sets of parental sets as described above. Additionally, we allow the restriction of the MBG subspace separately for each dimension to values less than R by requiring that the corresponding posteriors are above the $\exp(-\rho)$ ratio of the respective MAP value, which can be calibrated as described in Section 7.7.

We experimented with two search algorithms in the constructed MBG subspace. The first method simply used the hypercube defined by the R, ρ values, but its exponential nature makes it not feasible. The second is a uniform-cost search starting from the ordering conditional MAP MBG, which stops after the expansion of L^S number of states or if the most probable MBG in its search list (fringe) drops below $\exp(\rho^S)$ ratio of the ordering conditional posterior of the starting (MAP) MBG. Because the expansion is more costly computationally than the update, the algorithm continues with L^U, ρ^U parameters if necessary (we assume that bPartialUpdate implies bExpand). However the full, exact update has an acceptable run-time cost, if the size of the estimated MBG set L^T is below 10^6]. This L^T value ensures that the newly inserted MBGs are not pruned before their estimates reliably indicate their high-scoring potential and still allows an exact update. In larger domains this balance can be different and the analysis of this question in general is for future research. So subsequently we will report results using always a false value for the bPartialUpdate and we will discuss in the next section that $R = 20, \rho = 4$ and $L^S = 10^4, \rho^S = 10^{-6}$ are reasonable choices. Another aspect of the R, ρ, L^S, ρ^S parameters is that complex feature values with high posterior are not necessarily ranked high based on the ordering conditional posterior (see comments for Eq. 7.47 and 7.48). However in practice for a relatively peaked posterior over the orderings, the globally high-scoring feature values are determined by the high-scoring feature values for the most significant orderings, which are evaluated by a standard MCMC simulation. Note that in the case of applying the algorithm for a single ordering the parameters L^S, ρ^S are functionally equivalent to the parameters L^U, ρ^U. In this case the estimates of the posteriors of non-ordering-modular features, such as the MB feature based on the collected high-scoring MBGs are always underestimated and the increasing value of the equivalent L^S, L^U, L^T parameters in the limit ensures their exact posteriors (in practice this underestimation can be counterbalanced by normalization).

In the standard settings used in this section the length of the burn-in and MCMC simulation was 10^4, the probability of the pairwise replace operator was 0.8, the parameter prior was the $\mathrm{BD}_{\mathrm{eu}}$ and the structure prior was uniform prior for the parental sets with size less than k. The maximum number of parents

was 4, which is consistent with the prior and the posterior of the parental set size conditional on the ordering $\prec_0$. In general, for the reported MB and MBG values after burn-in with this setting the single-chain convergence test from Geweke comparing averages has a z-score of approximately 0.1, the R value of the multiple-chain method of Gelman-Rubin convergence test with 5 chains drops below 1.01 (see Section 2.3.1.3). The variances of the MCMC estimates of the feature values drop below 10^{-5}.

The method implemented can incorporate parameter priors, structure priors, and priors over the orderings including the dichotomy of the variables to causes vs. influencing factors and applying the ordering-based MCMC in these spaces. In this section we will focus on the tabula rasa application of the method.

8.5.2 The exact treatment of the orderings

If the set of a priori allowed ordering $S^{\prec}$ is very restricted, the exact posteriors for an ordering-modular feature F, such as edge, MBM and MBG features can be computed analytically for any prespecified value f.

This is not possible for non-ordering modular features, such as the MB feature, and the identification of high-scoring values is an additional issue in case of complex features. In the OC domain a total ordering is available, whose primary usage was to compute a baseline using all the clinical data and to calibrate the parameters of Alg. 1 that are used independently in the outer cycle working in the ordering space, such as the ordering-based MCMC.

The investigated parameters includes the R, ρ, and L^T parameters specifying the constructed MBG subspace and the maximum number of estimated MBGs, the L^S and ρ^S parameters controlling the expansion of the estimated MBGs. Related parameters are the partial update option (bPartialUpdate) and its additional parameters L^U, ρ^U. All these parameters depend on the peakness of the ordering conditional posteriors of the parental sets, or more exactly of the sets of parental sets constructing the MBG subspaces. Table A.4 shows the ordering conditional distributions of the parental sets for the total ordering $\prec_0$ and the posteriors of the corresponding MBG subspace. We can observe that the peakness (rate of decrease) ensures that the ordering conditional posterior of the tenth most probable value is less than a hundredth ($\exp(-5) < 0.01$) of the MAP value (see last column). Furthermore, the aggregated value in the MBG space are mostly present (see differences in the last column). In fact, it is frequently the most probable value (see differences in the first column). Exceptions are the strong indicator variables, such as CA125 and PapFlow, for which the parental sets without the target variable has insignificant posteriors. This suggests that it is sufficient to set the parameter R above 10. We used 20 because the CH parameter prior is more widespread and we used $\rho = -20$ corresponding to a cut-off ratio less than $10e-8$.

The ranked ordering conditional posteriors of the most probable complete BN structures, MBGs and MB sets are reported in Fig. 8.10.

The ranked posterior curves indicate that the parameter L^S has to be larger than 10^3, and the parameter L^T as well. We computed the ordering conditional

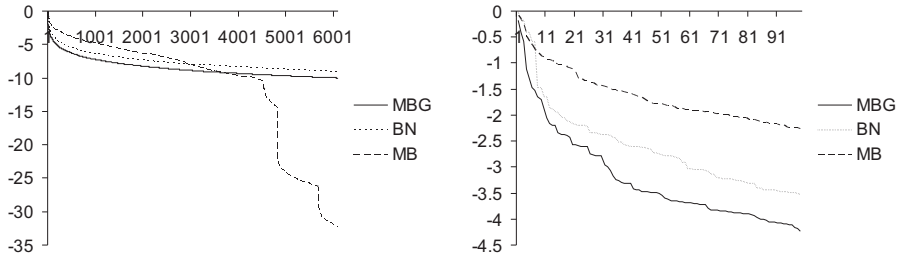


Figure 8.10: The rate of decrease of the posteriors of the most probable MB sets, MBGs, and BN structures. In each cases the posteriors are insignificant for objects with ranks larger than 100. The horizontal axis shows the relative logarithm of the ordering conditional posteriors of the most probable complete BN structures, MBGs and MB sets w.r.t. the respective MAP value. The vertical axis corresponds to the appropriate ranks.

posterior of the set S^S containing the most probable MBGs returned by the uniform cost search with different sizes. To make it computationally feasible we restricted the number of parents to two, which is still a realistic selection in case of the $\mathrm{BD_{eu}}$ parameter prior (see Fig. A.6). The ordering conditional posterior $p(\mathrm{mbg} : \mathrm{mbg} \in S^S | \prec_0, D_N)$ is 0.76 (10^3), 0.81(10^4), 0.93 (10^5) and 0.9911 (10^6) for the respective set sizes indicated in parenthesis. We used the values $L^S = 10^3$ and $L^T = 10^5$, which proved to be sufficient compared to checks with $L^S = 10^4$ and $L^T = 10^6$. Furthermore, we performed the same analysis for some random orderings, which indicated even more peaked posteriors.

8.5.2.1 Posteriors of Markov blanket memberships

The MBM(Y, G) feature is a symmetric, pairwise and observationally equivalent feature. Table A.5 reports the posteriors of the MBM(Pathology,X_i) features for the combinations of CH/$\mathrm{BD_{eu}}$ parameter priors and single/unconstrained orderings. The qualitative analysis of the MBM posteriors shows that the (7,10)/(8,12) variables have posteriors less than 0.05 and larger than 0.95 for the expert's ordering and (10,7)/(11,10) in the unconstrained case (the values corresponding to the thresholds (< 0.05,$0.95 <$) are separated by ',', and the '/' indicates the use of CH/$\mathrm{BD_{eu}}$ priors respectively). The effect of the parameter priors w.r.t. the 0.05 and 0.95 thresholds is negligible, as the only differences are the Locularity in the fixed, and the TAMX and Hysterectomy in the unconstrained ordering case. However the effect of the parameter priors at the quantitative level is significant as the L_1 difference of the posteriors is larger than 0.05 for 9 variables in the case of fixed and for 10 variables in the case of free orderings. Whereas this issue is for future research, to simplify the exposition we shall report results using the $\mathrm{BD_{eu}}$ parameter priors, which provides more compact models in our experiments, except its anomaly in the small sample re-

gion. The effect of the restriction to the expert's total ordering is considerable, because 8 variables have different status w.r.t. the 0.05 and 0.95 thresholds (Age, Meno, HormTherapy, PapSmooth PI TAMX, Hysterectomy Solid) and the L_1 difference of the posteriors is larger than 0.05 for 11 variables in the case of $\mathrm{BD}_{\mathrm{eu}}$ priors.

We investigated the relation of the MBM posteriors to the prior domain knowledge only in an overall comparison using the expert's rating of pairwise $S^{h/m/r}$ dependencies (see Section 4.3.3.2) and the induced Markov blankets of the multiparental $S^{H/M/R}$ dependencies (see Section 4.3.3.1, 4.3.3.2), which represent associative, pairwise priors and model-based priors. Table 8.8 reports the AUC values of MBM posteriors based on clinical and literature data w.r.t. the $S^{H/M/R}$ and $S^{h/m/r}$ relations. The relatively high AUC values indicate success particularly for the Markov blanket induced by the G^M structure, which was concluded as the most reliable prior structure (see Section 8.4.1, 8.4.2). For the $S^{h/m/r}$ priors, its performance is below that of the pairwise approaches based on data (see Table 8.1), which can be explained by the fact that these elicited expert priors represent indirect pairwise dependencies, in contrast to the model-based MBM relations.

Table 8.8: The learnability of the expert's opinion that a given variable is relevant for the preoperative diagnostics of ovarian cancer. We used the posteriors of the MBM(Pathology,X_i) features as scores to discriminate the expert's $S^{h|m|r}$ and $S^{H|M|R}$ relations; and reported the corresponding AUC values. High AUC values (above 0.8) indicate both the sufficiency of the data and the good quality of the prior as gold standard (e.g., the S^h and S^M cases). The posteriors are computed for single (FixO.) and unconstrained (MCMC) orderings using the maximum parental set size 4 and the IOTA-1.2 data set. The last line (Lit-FixO./BD) report the analog AUC values in case of $\mathrm{BD}_{\mathrm{eu}}$ priors and the $PM_{R_0}^{CR}$ Pubmed corpus.

	S^h	S^m	S^r	S^H	S^M	S^R
FixO./CH	.809	.691	.675	.731	.871	.759
FixO./BD	.828	.660	.700	.709	.886	.831
MCMC/CH	.721	.649	.642	.762	.852	.757
MCMC/BD	.726	.642	.742	.716	.856	.831
Lit-FixO./BD	.737	.687	.679	.827	.686	.646

To provide more information about the posteriors we computed the posteriors of the membership of each variable in the Markov blanket set of the Pathology variable for increasing data size using the original temporal sequence of the cases (i.e., the posterior learning curves). Note that these are different from the pure likelihood-based prequential curves (see Section 2.4.1 and 3.4), as these are combined with the prior and normalized in each step. Because of computational reasons this is computed only for the expert's total causal ordering. We classified the variables as simple vs. complex and positive vs. negative depending on the rate of convergence in this sequential evaluation and their final status using all the clinical data (the only exception is the Age variable, see below). This characterization is based on the fact that the rate of convergence

of an MBM feature represents the average rate of convergence of the domain models that makes it true. For example if MBM(Pathology,X_i) is true only in complex models requiring large sample size for a significant posterior including the limiting equivalence class as well, then the feature is a complex, positive feature with slow increasing rate. Fig. 8.11, 8.12 shows the respective curves.

These classification of the variables allows a qualitative evaluation of the sufficiency of the sample size of the clinical data set with 782 cases w.r.t. these MBM features. Because of the general non-monotonicity of the sequential posterior curves with multiple maxima we cannot exclude the possibility of the changes of the trends for further cases, but the reported results suggest that all variables would continue its convergence to the 0/1 value according to its status at using the complete data set. Additional computations using 2 and 3 as maximum parental set sizes also confirmed this opinion.

The MBM pairwise features in general are not independent, and in Fig. 8.12 we can observe an illustrative example. This figure shows the MBM features with slow convergence to 0, except the MBM(Pathology,Age), which has a slowly increasing trend. Because the Age and the Meno variables are the most significant potential parents in this ordering and semantically they are quite redundant, the observable complementary characteristics from 200 samples indicates their antagonist dependence (the sum of their posterior is in the $[0.95, 1.1]$ range).

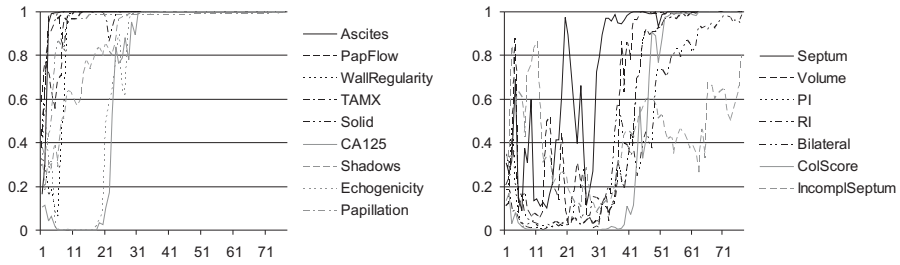


Figure 8.11: The temporal evolution of the belief — inferred from growing amount of clinical data — that a given variable is relevant for the preoperative diagnostics of ovarian cancer. Belief in relevance is represented by the posterior of the MBM feature, thus the figure shows the sequential posteriors of MBM(Pathology,.) features with fast/slow convergence to 1 given the expert's total causal ordering. In both groups the majority of posteriors are highly varying below 200 samples, but they are stabilized above 500 samples. The posteriors are computed with $\mathrm{BD}_{\mathrm{eu}}$ priors, with noninformative structure priors and conditionally on the expert's total causal ordering. The horizontal axis is the sample size with step size 10.

Finally we report the sequential posteriors of the MBM(Pathology,.) features using the temporal sequence of publications between 1980 and 2005 in the large PubMed corpus. Analogously to the clinical data case we tried to classify the variables w.r.t. rate of convergence and assumed limiting value as

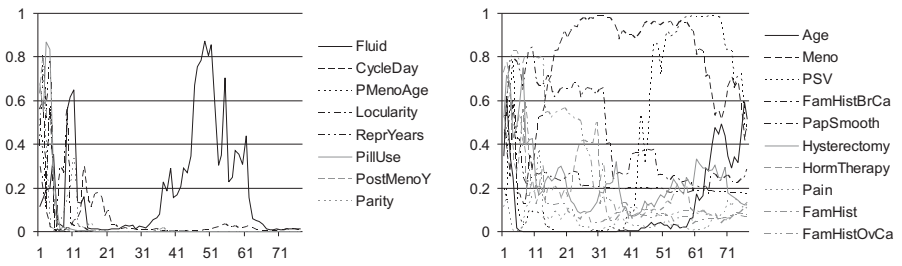


Figure 8.12: The temporal evolution of the belief — inferred from growing amount of clinical data — that a given variable is not relevant for the preoperative diagnostics of ovarian cancer. Belief in relevance is represented by the posterior of the MBM feature, thus the figure shows the sequential posteriors of MBM(Pathology,.) features with fast/slow convergence to 0 using the sequence of clinical samples and given the expert's total causal ordering. The posteriors are computed with $\mathrm{BD}_{\mathrm{eu}}$ priors, with noninformative structure priors and conditionally on the expert's total causal ordering. The horizontal axis is the sample size with step size 10.

MBM(Pathology,.) features with fast or slow convergence to 0 or 1, but we defined a class with "mixed status" instead of "fast convergence to 0" class (with one element). The comparison of the clinical data based "fast convergence to 0" class against the opposite literature based "fast convergence to 1" class revealed only 1 common item (Locularity). The comparison of the clinical data based "fast convergence to 1" class against the opposite literature based "slow convergence to 0" class revealed only 2 common items (Echogenicity and Shadows). These are semantically related and the different classifications probably indicate the relatively recent status of these diagnostic features.

8.5.2.2 Posteriors of MB sets and MB graphs

The pairwise $\mathrm{MBM}(Y, X_i)$ features are model-based, but they treat independently the variables X_i. If this assumption is acceptable for example because of the general assumption of a naive BN with parent Y, then the MBM posteriors $p(\mathrm{MBM}(Y, .)|D_N)$ can be used to approximate the posterior of Markov blanket sets as follows

$$p(\mathrm{MB}(Y){=}\mathrm{mb}|D_N) \approx \prod_{X_i \in \mathrm{mb}} p(\mathrm{MBM}(Y, X_i)|D_N) \prod_{X_i \notin \mathrm{mb}} (1 - p(\mathrm{MBM}(Y, X_i)|D_N)). \quad (8.26)$$

Similarly, the posterior of the Markov blanket spanning subgraph can be approximated using edge posteriors

$$p(\mathrm{MBG}(Y) = \mathrm{mbg}\,|D_N) \approx \prod_{e_{ij} \in \mathrm{mbg}} p(e_{ij}|D_N) \prod_{e_{ij} \notin \mathrm{mbg}} (1 - p(e_{ij}|D_N)). \quad (8.27)$$

Within this approximation high-scoring MB and MBG features can be found

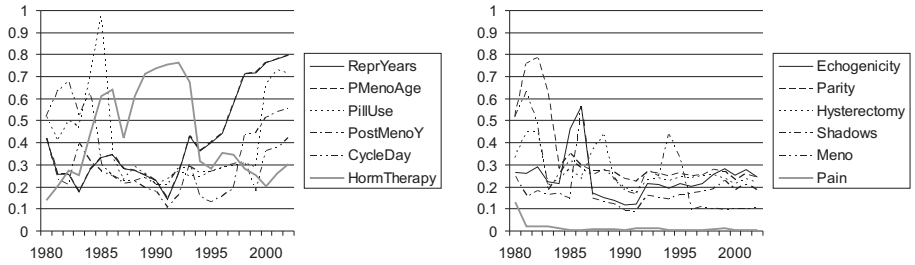


Figure 8.13: The temporal evolution of the collective belief — inferred from the literature — that a given variable is not relevant for the preoperative diagnostics of ovarian cancer. Belief in (pairwise) relevance is represented by the posterior of the MBM feature, thus the figure shows the sequential posteriors of MBM(Pathology,.) features with mixed status with fast convergence to 0 using the temporal sequence of publications between 1980 and 2005 in the large PubMed corpus binarized with corelevance, $\mathrm{BD}_{\mathrm{eu}}$ priors, noninformative structure priors and conditionally on the expert's total causal ordering.

in $\mathcal{O}(n^2)$ time by exhaustively collecting and combining the most probable pairwise feature values.

Without such strong assumptions, we can use the MBG-ordering based search-estimate algorithm described in Alg.1 to search for high-scoring complex features and estimate their posterior exactly in the limit. Fig. 8.14 shows the estimated ranked posteriors and their MBM-based approximations as in Eq. 8.26 for the 20 most probable MB set (w.r.t. estimated posteriors).

The MBM-based approximation performs relatively well, particularly w.r.t. ranking in the case of the expert's ordering $\prec_0$, but it performs poorly in the unconstrained case both w.r.t. estimates and ranks (note that it excludes relevant MB sets). As a cross-check we also computed the MB-based MBM posteriors using the 100 most probable MB set

$$p(\mathrm{MBM}(Y, X_i)|D_N) \approx \sum_{\mathrm{mb}\in S_{MB}^{100}} p(\mathrm{mb}\,|D_N) 1(X_i \in \mathrm{mb}), \tag{8.28}$$

which provided good estimates (see Table A.5), because of the relatively peaked MB posterior (see Fig. 8.14).

The 10 most probable MB sets are reported in Table A.7. The sum of the posteriors of these 10 sets (i.e., their coverage) is around 0.4. The variables with changing status are FamHist, IncomplSeptum, Pain, Locularity, PI, TAMX, FamHistBrCa, Shadows, ColScore, PI and PSV, which coincides approximately with the set of variables with MBM(Pathology,.) posterior in the range of $[0.1, 0.9]$. Note that because of the exact posterior interpretation, variables with MBM posterior close to 0.5 are affected earlier by increasing the coverage of the reported top MB sets. From a theoretical point of view this shows that the set of DAGs spanning the high-ranking MB sets are approxi-

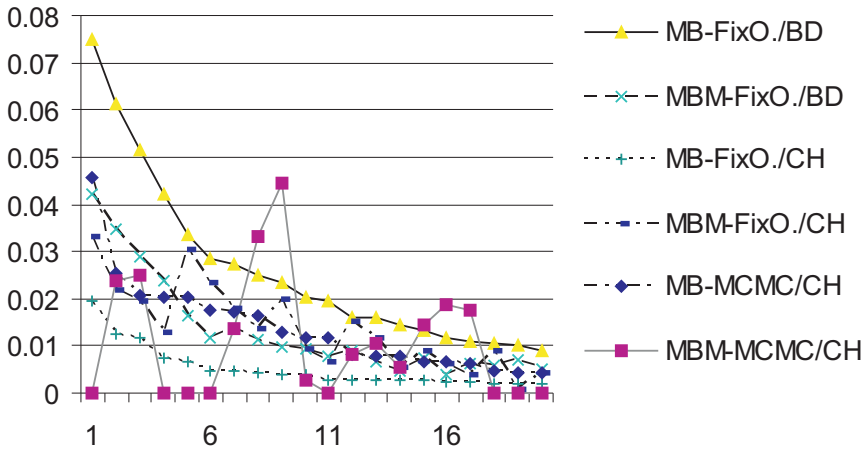


Figure 8.14: The MBM-based, pairwise approximations of the posteriors and ranks of the 20 most probable MB(*Pathology*) sets. In the general (unconstrained) case the MBM-based approximation give misleading results w.r.t. the multivariate analysis by excluding relevant MB sets (additionally to its seriously perturbed ranking). We used combinations of $CH/\mathrm{BD}_{\mathrm{eu}}$ parameter priors and single/unconstrained orderings (FixO./MCMC) with uniorm structure prior, the IOTA-1.2 data set, and the maximum parental set size 4. The vertical axis corresponds to the ranks of the MB sets w.r.t. the case of CH prior and unconstrained orderings.

mately representative. From a practical point of view this backs up the literal definition of the MBM feature that its posterior can be used to indicate the uncertainty of the relevance of a variable status in a model-based manner, but it performs poorly in a multivariate context to estimate or rank the overall MB sets (see Table 8.14). Table 8.9 shows the classification performance of high-scoring MB sets from Table A.7 and for a MB set based on literature using the $PM^{CR}_{R_0}$ Pubmed corpus, which corresponds to the settings reported in Table 8.8.

The relatively low misclassification rates, particularly the $.15 - .25$ values for the M^* MB set induced by the most relevant G^M structure indicate that the data based MB sets approximate well the expert's references. Interestingly, the most probable set based on the MBM posteriors reported in Fig. 8.14 under the title MBM-MCMC/CH has similarly high classification performance, which is partly the consequence of these simple performance measure (see Fig. 8.14 for the performance of the MBM posteriors to rank and estimate real posteriors of MB sets).

Finally we compared the $\mathrm{MBG}(Y, G^{\mathrm{MAP}})$ and $\mathrm{MB}(Y, G^{\mathrm{MAP}})$ feature values defined by the MAP BN structure G^{MAP} against the MAP MBG feature value $\mathrm{MBG}(Y)^{\mathrm{MAP}}$ and the MAP MB feature value $\mathrm{MB}(Y)^{\mathrm{MAP}}$ including the MB feature value defined by the MAP MBG feature value $\mathrm{MB}(Y, \mathrm{MBG}(Y)^{\mathrm{MAP}})$

Table 8.9: The sensitivity, specificity and misclassification rate of the most probable MB sets of the Pathology variable w.r.t. the $S^{h/m/r}$ relations and the Markov blankets induced by the $G^{H/M/R}$ structures (denoted with $(H/M/R)^*$). The MBM line reports the performance of the most probable set based on the MBM posteriors in the same settings. The Lit. line report the analog values in case of the literature data $D^{PM^R_R}$.

	Sensitivity						Specificity (%)						Misclassification rate (%)					
MB set	h	m	r	H^*	M^*	R^*	h	m	r	H^*	M^*	R^*	h	m	r	H^*	M^*	R^*
FixO_1	.87	.64	.6	.75	.77	.67	.26	.22	0	.46	.08	0	.21	.32	.35	.41	.18	.26
FixO_6	.87	.64	.57	.75	.77	.63	.21	.11	0	.42		0	.18	.29	.38	.38	.15	.29
MCMC_1	.67	.56	.53	.75	.68	.59	.32	.22	0	.38	.08	0	.32	.38	.41	.35	.24	.32
MCMC_2	.73	.6	.57	.75	.73	.63	.32	.22	0	.42	.08	0	.29	.35	.38	.38	.21	.29
Lit.	.6	.56	.53	.88	.55	.56	.37	.22	0	.35	.33	.14	.38	.38	.41	.29	.41	.38
MBM	.87	.68	.6	.88	.82	.67	.37	.33	.5	.5	.17	.29	.26	.32	.41	.41	.18	.32

$$\text{MBG}^{\text{MAP}} = \arg\max_{\text{mbg}} p(\text{mbg}\,|D_N)\ \ \text{MB}^{\text{MAP}} = \arg\max_{\text{mb}} p(\text{mb}\,|D_N) \tag{8.29}$$

First note that because of its goal in general the best scoring network found in the MCMC simulation is significantly worse than the best MAP structure found in optimization (with scores of -14294.48 vs. -14069.72). So we also performed the comparison using the best BN structure found in the MCMC simulation, beside the empirically best ordering conditional model (reported in Fig. A.5). The MAP MBG feature value $\text{MBG}(Y)^{\text{MAP}}$ differed significantly from both MAP domain model, as it was already suggested by the difference between the ordering conditional parental set space and MBG space reported in Table A.4. Different parental sets or variables in the MAP MBG are, for example the additional Meno variable (vs. Age in the domain models) and missing Solid and Fluid variables (against their child status in domain models). The MAP MB feature value $\text{MB}(Y)^{\text{MAP}}$ similarly differs from the MB sets defined by the MAP domain models for example w.r.t. the vascularization variables, such as PI. Interestingly, the MAP MB feature value also differs from the MB feature value defined by the MAP MBG feature value $\text{MB}(Y, \text{MBG}(Y)^{\text{MAP}})$, for example w.r.t. the Age vs. Meno variables and TAMX, PI, Solid variables. In conclusion these results together with the comparison against the simple feature-based analysis, such as MBM-based analysis, shows the relevance of the complex feature-based analysis.

Similarly to the pairwise case, to provide information about the sequential and sample size aspects of the posteriors we computed the sequence of posteriors of high-scoring MBG and MB feature values for increasing data size. We used the original temporal sequence of the cases and the expert's ordering $\prec_0$. Fig. 1.4 reports the MB case, Fig. 1.5 the MBG case and Fig. 1.4 reports the sequence of posterior of an empirically MAP domain model conditionally on ordering $\prec_0$ reported in Fig. A.5.

The trends of the ordering conditional sequential posteriors of the MBG and MB values indicate that the complex MAP feature values at this sample size are in a transitional phase when the posterior is not concentrated around a

single model and not even around a single complex feature value. On the one hand, for half of the sample size, the final MAP feature values have negligible posteriors (see Fig. 1.5 and 1.4) and the ranked posterior of complex feature values for the complete sample in Fig. 8.14 shows that there are approximately 20-30 feature values with posterior above 0.01 (both for the ordering $\prec_0$ and for unconstrained orderings). In fact this multivariate uncertainty can be expected based on the uncertainty on the more simple pairwise level of the MBM features (e.g., see Fig. 8.11 and 8.12). This shows the necessity of the exact estimation of the posteriors of complex feature values with known confidence values.

8.5.3 Applying MCMC methods over the orderings

After the discussion of the aspects of searching complex features given an ordering and the sample size aspects of feature posteriors given an ordering, this section describes the aspects of the estimation of simple and complex features. To determine a standard setting, we applied single chain and multiple chain convergence methods for one of the simple feature (MBM) related to classification and for two complex features (MB and MBG). We also quantify the MCMC sampling variance by reporting its estimates using the batch approach. The burn-in was 10000 which will be the standard settings in the Section, though the results proved to be robust for burn-in larger than 1000. The Z_G values for the $\log(p(\prec))$ for the 4 chains were less than 0.5363 and the $\hat{R}$ was 1.0611, which does not refute convergence in general, though it has to be tested for individual feature values as well (see Section 2.3.1.3).

For the pairwise features, the maxima of the Z_G values was less than 0.7286, the maximum of the $\hat{R}$ values is 1.12 and the square root of the variance of the posterior was less than 0.0794, see Table A.6.For the complex features, we report the convergence scores and the estimated posterior with their estimated variances for the most probable values. Table A.10 reports the most probable MBGs, Table A.11 reports their Z_G and $\hat{R}$ convergence values with the square root of the variance of the posterior averages. Table A.7 and Table A.8 reports the same for the most probable MB sets.

8.6 Effect of fusion

The Section 8.1 described multiple approaches to the fusion of expert prior, literature (data) and clinical data, including a particularly simple. It infers feature posteriors from the literature data and creates composite priors by integrating them with expert priors through hyperparameters. We apply this approach in this section using informative deviation priors with the most relevant G^H and G^M structures from the expert and various edge priors, such as the edge prior from the expert and edge priors derived from various literature data sets (as a test case we also included edge priors derived from the training part of the clinical data set). To quantify the overall effect of the incorporation of structure priors, we use the first half of the IOTA-1.2 data set as training sample to select

MAP BN models and report their (prior free) log-likelihood using the second part w.r.t. a MAP model with a reference prior. To minimize the effects of the optimization, the learning was performed using the expert's total ordering of the variables with exhaustive search up to five parents with a subsequent K2 greedy search. The expert's informative edge deviation priors are combined with the G^H and the G^M structure (i.e., the G^H or the G^M structure were applied as an a priori structure to define a deviation prior). The tested priors are the noninformative prior as the reference, the uniform edge deviation prior with $p = 0.1$ (denoted by the NI01 postfix in the graph legend), the expert's rating of edges without scaling and by scaling to 3 average parents as described in Section 3.1.5.2 ((denoted by the E and E3 postfixes)), the expert's rating of edges with 0.01, 0.3, 0.6, 0.9 (EXP963), the mutual information scaled to 3 average parent (MI), the Bayesian pairwise score from Eq. 8.19 scaled to 3 average parent (PW) and (EPW) the exact edge posteriors using the second-half of the IOTA-1.2 data set (EPW) and the literature data set $D^{\mathrm{PM}_R^R}$ (PM-R-R) with the expert's total ordering of the variables and the limit of four on the number of parents. Fig. 8.15 reports the effect of the expert's structure prior on learning MAP Bayesian networks for varying sample sizes. The effect of the structure priors on classification are reported in Section 10.7.

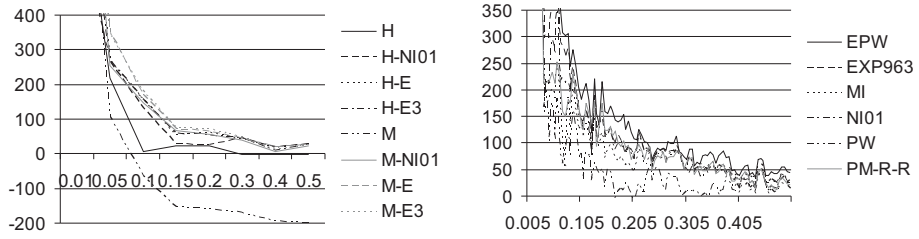


Figure 8.15: (Left) The effect of the expert's informative edge deviation priors w.r.t. the G^H (H) and the G^M structure (M). The tested priors are the uniform edge deviation prior with $p = 0.1$ (NI01), the expert's rating of edges without scaling and by scaling to 3 average parents (E and E3). (Right) The effect of various edge deviation priors w.r.t. the G^M prior structure. The tested priors are the expert's rating of edges (EXP963), the mutual information (MI) and Bayesian pairwise scores (PW), and the exact edge posteriors using the second-half of the IOTA-1.2 data set (EPW) and the literature data set $D^{\mathrm{PM}_R^R}$ (PM-R-R). The first half of the temporal sequence of the IOTA-1.2 data set is used in the structure learning incrementally (showed on the x-axis). The performance measure of the MAP Bayesian network is the relative likelihood score computed on the second half of the IOTA-1.2 data set w.r.t. MAP models with the reference, uniform prior.

Chapter 9

Bayesian classification

We overview the application of the Bayesian framework to model a dependency relation (i.e., its conditional application). We summarize performance measures for classification and regression. Then we discuss the application of Bayesian networks for classification. Finally logistic regression and multilayer perceptrons are summarized, particularly their relation to the Markov blanket subgraph feature.

In the *conditional approach* the primary interest is in modeling and understanding the uncertain dependency relation between the output variables Y and the input variables X. For ovarian cancer, this is the dependence of the *Pathology* variable. If the output variable is discrete the problem is called a classification problem, otherwise a regression problem. Other naming conventions for the output and input variables are the response, outcome, or dependent variables and predictor, explanatory, or independent variables (or covariates). If the dependency is uncertain, then the probabilistic approach can be adopted using the same reasoning as in Section 2.1, that is our goal is to model

$$p(Y_{N+1}|X_{N+1}, (X_1, Y_1), \ldots, (X_N, Y_N)). \tag{9.1}$$

Furthermore, an analogous parametric Bayesian representation can be derived for conditionally exchangeable distributions [72]

$$p(y_1, \ldots, y_N|x_1, \ldots, x_N) = \int (\prod_{i=1}^{N} p(y_i|\theta(x_i)))p(\theta(x))\,\mathrm{d}\theta(x), \tag{9.2}$$

where $p(\theta(x))$ is a prior over the parameters for a parametric function class specifying $p(y_i|\theta(x_i))$.

A significant computational challenge for Bayesian classification models, such as Bayesian logistic regression and its extensions, is the lack of conjugate prior, in contrast to the availability (and even necessity) of the Dirichlet parameter prior for Bayesian networks.

Clearly, the prior belief corresponding to a domain model determines the validity of a usually computationally and statistically simpler conditional mod-

eling, but a general structure and parameter prior $p(G), p(\theta|G)$ for a Bayesian network modeling the complete domain does not decompose to a prior for a conditional model $p(Y|X)$. That is, for $\{Y, X\} \subset V$

$$p(y|x) = \sum_G p(G) \int_\Theta p(y|x, \theta, G)\, \mathrm{d}p(\theta). \tag{9.3}$$

In general, a prior domain model can guide the whole process of constructing a classifier, for example by performing inferences about properties of the domain model relevant to construct the structure of the classification model. Another usage is that the modular parameter and structure priors can be used directly as building blocks in priors for various simpler classification models, or as sources for probabilistically inducing such priors. It can support the definition of real, informative priors for classification models — a largely open problem for many conditional model classes. It can also support the interpretation of posteriors from the conditional modeling. These issues, particularly the usage of the domain model in classifier construction, the issue of probabilistically linked model spaces and induced priors are central issues in this thesis and we will investigate the relation of classification models and their relation to the domain model in Section 9.5.

Bayesian network classifiers are specially restricted Bayesian networks. As such, they are eligible for the causal interpretation as well and can incorporate directly prior information. This made them a natural choice as classification models in this thesis. Logistic regression (LR) was applied, because of its biomedical interpretation and to provide a baseline with its established methodology. The multilayer perceptron model was selected, because of its derivation from the LRs and its universal approximating capacity as a parametric function.

To treat the binary classification task, we introduce the following notation: $\underline{\omega} \in \mathbb{R}^d$ denotes the model parameters, and when distinction is necessary, $\underline{\omega}$ denotes the parameters of the multilayer perceptron and $\underline{\theta}$ the parameters of the Bayesian network. We assume the existence of a labeled training set $D_N = \{(\underline{x}_k, y_k)\}_{k=1}^N, \quad (\underline{x}_k, y_k) \in \mathbb{R}^d \times \{0, 1\}$, where $\underline{x}$ is a real-valued l-dimensional input vector and y is the corresponding class label.

9.1 On the validity of the conditional approach

In the pragmatist Bayesian approach, the assumptions for the conditional representation in Eq. 9.2 means independent beliefs corresponding to the modeling of the dependence of the output variable Y on X (i.e., without modeling the overall domain). Let us assume that the distribution over the observables $p(Y, X)$ is defined by the prior $p(\theta)$ and the sampling distribution $p(Y, X|\theta)$, where the parameter θ is composed of (ϕ, ω), the parameter ϕ corresponds to X as $X \not\perp\!\!\!\perp \phi$ and $(X \perp\!\!\!\perp \omega|\phi)$ and ω corresponds to $Y|X$ as $Y \not\perp\!\!\!\perp \{X, \omega\}$ and $(Y \perp\!\!\!\perp \phi|\omega, X)$. The conditional approach assumes that $\omega \perp\!\!\!\perp \phi$ (i.e., decomposed priors, Fig. 9.1 shows the corresponding Bayesian network).

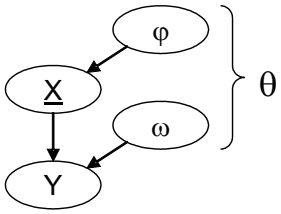


Figure 9.1: The Bayesian network representation of the assumptions of the Bayesian conditional modeling. The dependent and independent variables are denoted with Y and X, the corresponding parameters are ω and ϕ (jointly forming θ).

Decomposed priors ensures the basis of the conditional approach, because it implies decomposed posteriors for complete observations

$$\begin{aligned} p(\theta|x,y) \triangleq p(\omega,\phi|x,y) &\propto p(x,y|\omega,\phi)p(\omega,\phi) && (9.4)\\ &= p(y|x,\omega)p(x|\phi)p(\omega|\phi)p(\phi) && (9.5)\\ &= p(y|x,\omega)p(\omega)p(x|\phi)p(\phi) && (9.6)\\ &\propto p(\omega|x,y)p(\phi|x). && (9.7) \end{aligned}$$

That is, if we are interested only in the conditional model (i.e., in ω), then in the conditional approach

$$\begin{aligned} p(\omega|x,y) &\propto \int_{\phi} p(y|x,\omega,\phi)p(x|\phi)p(\omega|\phi)p(\phi)\,\mathrm{d}\phi && (9.8)\\ &= p(y|x,\omega)p(\omega) && (9.9) \end{aligned}$$

9.2 The Bayesian modeling of class probabilities

In binary classification problems, we are interested, for a given $\underline{x}$, in the class label, which are the observable quantities. Because of the uncertain dependency of Y on X, we shall model their distribution in the conditional, parametric Bayesian framework by a parametric *probabilistic regression* model or regression function $P(Y=1|\underline{x},\underline{\omega}) = f(\underline{x},\underline{\omega}) \in [0,1]$ and a prior distribution over the model parameters $\underline{\omega}$. Note that $\mathrm{E}[Y|\omega,X] = f(\underline{x},\underline{\omega})$, that is by defining the output values as in [138]:

$$y = f(\underline{x},\underline{\omega}) + \epsilon, \text{where } \epsilon = \begin{cases} 1 - f(\underline{x},\underline{\omega}) \text{ with probability } f(\underline{x},\underline{\omega}) \\ -f(\underline{x},\underline{\omega}) \text{ with probability } 1 - f(\underline{x},\underline{\omega}) \end{cases} \quad (9.10)$$

a corresponding regression problem can be defined, where the conditional model corresponds to the conditional mean with this Bernoulli error term, instead of a Gaussian one.

In this approach, the goal is the modeling of the uncertain dependency of Y on X through the mean regression function $\mathrm{E}_{p(\underline{\omega})}[f(\underline{x},\underline{\omega})]$ and not the direct

modeling of the optimal class label under a given fixed loss $L(y, \hat{y})$ through a decision function $g(x)$ with range $\{0, 1\}$. Furthermore, if the function class is interpretable, the goal is the modeling of the uncertainties at that level as well. In general, the regression approach allows a wider range of applications, but it is statistically harder because optimal decisions can be reached using an imprecise regression estimate in a threshold-based decision rule [73].

In the Bayesian regression approach to classification, the prior distribution can be transformed into the posterior distribution $p_{\underline{\Omega}}(\underline{\omega}|D_N)$ by using the observed data D_N and applying Bayes' rule:

$$p_{\underline{\Omega}}(\underline{\omega}|D_N) = \frac{p_{\underline{Y}}(y_1, \ldots, y_N|\underline{\omega}, \underline{x}_1, \ldots, \underline{x}_N) p_{\underline{\Omega}}(\underline{\omega})}{p(D_N)} \propto L(\underline{\omega}; D_N) p_{\underline{\Omega}}(\underline{\omega}).$$

$L(\underline{\omega}; D_N)$ denotes the likelihood function and $f(\underline{x}, \underline{\Omega})$ denotes the induced random variable on $[0, 1]$ for the predicted posterior class probability. Note that, in general, because of the nonlinearity of $f(.)$, $f(\underline{x}) = \mathrm{E}[f(\underline{x}, \underline{\Omega})]$ is not equal to $f(\underline{x}, \bar{\underline{\omega}})$, where $\bar{\underline{\omega}} = \mathrm{E}[\underline{\omega}]$ or with $f(\underline{x}, \underline{\omega}^{\mathrm{MAP}})$ where $\underline{\omega}^{\mathrm{MAP}} = \arg\max p(\underline{\omega}|D_N)$ (for an overview of approximations, see [36], p.405).

So this conditional, parametric Bayesian approach to binary classification, for a given $\underline{x}$, provides the random variable $f(\underline{x}, \underline{\Omega})$ corresponding to the probability $P(Y = 1|\underline{x}, \underline{\omega})$, where $\underline{\Omega}$ is a random parameter vector. In this way the distribution of $f(\underline{x}, \underline{\Omega})$ gives us uncertainty information about the predicted class probability.

9.3 Reporting as decision in Bayesian classification

We can define the following decision problems for a given $\underline{x}$ having the distribution of the class probability $P(Y = 1|\underline{x}, \underline{\omega})$ as $f(\underline{x}, \underline{\Omega})$. The outcome can be either the binary class label y corresponding to the observable quantity or a scalar class probability $p(y|\underline{x})$ in some imaginary reporting situation. The action can be the report of a predicted the class label $g(\underline{x})$ (defined by a decision function) or the scalar class probability $f(\underline{x})$ (defined by a regression function), possibly combined with the option of "no decision" and rejection, where $g(.)$ and $f(.)$ are based on $f(\underline{x}, \underline{\Omega})$.

9.3.1 Reporting the class label

If the outcome y and the reporting action $\hat{y}$ are binary, the loss is defined by a binary cost matrix $C_{y,\hat{y}}$, such as the misclassification rate ($C_{0,1} = C_{1,0} = 1$ and $C_{0,0} = C_{1,1} = 0$). The minimal loss decision is

$$\arg\min_{\hat{y}} C_{0,\hat{y}} P(Y = 0|\underline{x}) + C_{1,\hat{y}} P(Y = 1|\underline{x}), \text{ where } p(Y|\underline{x}) = \int p(Y|\underline{x}, \underline{\omega}) p(\underline{\omega}) \, \mathrm{d}\underline{\omega}, \tag{9.11}$$

which shows that only the mean class probabilities are present in the decision.

9.3.2 Reporting the class probability

If the outcome y is binary and the reported value is the conditional probability $\hat{p}(y|x)$, then the logarithmic loss is a standard choice,

$$L(y, f(\underline{x})) = y \log(f(\underline{x})) + (1 - y) \log(1 - f(\underline{x})), \tag{9.12}$$

which corresponds to the cross-entropy error function [36] and was characterized as a score function for the multinomial case used in the prequential analysis.

The reported conditional distribution allows a refined assessment of the performance of a probabilistic classifier using the Area Under the Receiver Operating Characteristic (ROC) Curve (AUC), which is an utility-independent performance measure particularly widespread in medical applications (see [123, 122]). It evaluates a score (test) function $t(x) : X \to \mathcal{R}$ in the case of a binary outcome by analyzing its sensitivity and specificity on a given data set for all effectively different utilities (i.e., for all threshold $\tau \in [0, 1]$ assuming optimal decisions) as follows. The score function in case of a probabilistic classifier is the regression function or the mean regression function $f(\underline{x})$. The *sensitivity* is defined as

$$\text{Sens}(\tau) = p(\tau \leq t(\underline{x})|y = 1) \approx \frac{\sum_{i=1}^{N} y_i 1(\tau \leq t(\underline{x}))}{\sum_{i=1}^{N} y_i}, \tag{9.13}$$

and *specificity* is defined as

$$\text{Spec}(\tau) = p(\tau \geq t(\underline{x})|y = 0) \approx \frac{\sum_{i=1}^{N} (1 - y_i) 1(\tau \geq t(\underline{x}))}{\sum_{i=1}^{N} 1 - y_i}. \tag{9.14}$$

For a finite sample (and because of the possible discreteness of X), the *Receiver Operating Characteristic* (*ROC*) Curve is defined over the effectively different values of the threshold parameter $(Sens(\tau_i), 1 - Spec(\tau_i))$ for $\tau_i \in [0, 1]$. For test statistics, see [123, 122, 236].

9.4 Bayesian network classifiers

First we discuss the case where the modeling of the dependence of Y on X cannot be separated from the modeling of the input variables X in the sense of Eq. 9.9 and possibly from the modeling of other domain variables $X \cup Y \cup Z$.

9.4.1 Domain models as classifiers

Using the Bayesian network representation, the conditional is

$$\begin{aligned}
& p(\underline{y}|\underline{x}, D_N) && (9.15)\\
&= \mathrm{E}_{p(G|D_N)}[\mathrm{E}_{p(\underline{\theta}|G,D_N)}[p(\underline{y}|\underline{x}, \underline{\theta}, G)]] && (9.16)\\
&= \sum_G p(G|D_N) \int p(\underline{y}|\underline{x}, \underline{\theta}, G) p(\underline{\theta}|G, D_N) \, \mathrm{d}\theta && (9.17)\\
&= \sum_{\mathrm{MBG}(Y)=\mathrm{mbg}} p(\mathrm{mbg}\,|D_N) \int p(\underline{y}|\underline{x}, \underline{\theta}, \mathrm{mbg}) p(\underline{\theta}|\,\mathrm{mbg}, D_N) \, \mathrm{d}\theta, && (9.18)
\end{aligned}$$

which requires efficient model averaging technique.

The use of the domain model as a classifier has many advantages, for example to handle missing values. However, the knowledge engineering (parameter and structural prior construction), statistical (sample collection), and computational aspects suggest the use of restricted domain models that are specialized to the classification of Y using X. A special case is the MBGs in Eq 9.18 exactly representing the relevant information from the domain model. Ideally, these restricted Bayesian networks have the following properties: (1) the optimization or the Bayesian inference in the model class is computationally more efficient, (2) it can gradually approximate the structural (dependency) and parametric information of an unrestricted Bayesian network w.r.t. classification $Y|X$, and (3) it preserves the interpretation of the Bayesian network, so it can incorporate direct or transformed (induced) priors from the unrestricted Bayesian network and its posteriors can be analyzed and interpreted.

Besides these general attempts to avoid the larger cost of domain modeling, there are specific problems in applying a domain model as classifier both in the Bayesian and the frequentist framework. Because of the centrality of the likelihood term both in the frequentist model selection (e.g., in the MDL score in Eq. 3.46) and Bayesian model selection (e.g., in the prequential framework in Eq. 2.33), let us consider first the decomposition of the complete data likelihood for a fixed model G, θ containing only $\underline{X}, Y$ [91]:

$$LL(G, \theta; D_N) = \log p(D_N|G, \theta) \tag{9.19}$$

$$= \quad CLL_Y(G, \theta; D_N) + \sum_{i=1}^{N} \log p(\underline{X}_i|G, \theta) \tag{9.20}$$

where

$$CLL_Y(G, \theta; D_N) \quad = \quad \sum_{i=1}^{N} \log p(Y_i|G, \theta, \underline{X}_i). \tag{9.21}$$

The first term is the conditional data log-likelihood $CLL_Y(G, \theta; D_N)$, which solely determines the classification. The other term is a price for not working with a conditional model class, which as we shall see either introduces biased parameters and statistical noise or causes ordering-dependency in the case of its absence. First, because of the generality of the Bayesian network (e.g., the presence of non-factorizing normalizing constants), the optimal parameters for classification $\arg\max_\theta CLL_Y(G, \theta; D_N)$ are not equal to the maximum likelihood parameters for the domain model $\arg\max_\theta LL_Y(G, \theta; D_N)$ nor with the mean parameters $\mathrm{E}_{p(\theta|G,D_N)}[\theta]$ for some prior $p(\theta)$, and their determination requires optimization. This is in contrast with the existence of efficiently computable closed forms for the other two cases. The only exception is if the output variable is a leaf node [91], such as in the noisy-OR classifier [252]. Second, the term $CLL_Y(G, \theta; D_N)$ is dominated by the $n - 1$ analog terms for the output variables, which could cause erroneous model selection and especially feature selection [91]. Now consider the exclusive use of the first term in the Bayesian

framework as suggested and applied in [227, 158, 160]. Let us factorize the marginal model likelihood as follows (using the notation $\underline{Y}_l = (Y_i)_{i=1,\ldots,l}$ and $\underline{\underline{X}}_l = (\underline{X}_i)_{i=1,\ldots,l}$)

$$p(D_N|G) = p(\underline{Y}_N, \underline{\underline{X}}_N|G) = \prod_{i=1}^{N} p(Y_i, \underline{X}_i|\underline{Y}_{i-1}, \underline{\underline{X}}_{i-1}, G) \tag{9.22}$$

$$= \prod_{i=1}^{N} p(Y_i|\underline{Y}_{i-1}, \underline{\underline{X}}_i, G) \prod_{i=1}^{N} p(\underline{X}_i|\underline{Y}_{i-1}, \underline{\underline{X}}_{i-1}, G). \tag{9.23}$$

The logarithm of the first product is identical to a prequential score called *conditional node monitor* or *MBG monitor* (see Eq. 3.38 and 3.39), but this score is ordering-dependent, so it cannot be used directly as a batch score only as a sequential cumulative score (for approximations, see [160]). It is so, despite the ordering independence of the *global monitor* (see 3.35) and the ordering independence of the cross-entropy score of a binary regression model M $f_\theta(x)$:

$$\log p(\underline{Y}_N|\underline{\underline{X}}_N, M) = \sum_{i=1}^{N} \log p(Y_i|\underline{Y}_{i-1}, \underline{\underline{X}}_i, M), \tag{9.24}$$

which holds because of the independence of the belief ω in a conditional model and X_i ($\omega \perp\!\!\!\perp X_i$) until Y_i is given according to Eq. 9.9.

9.4.2 The naive Bayesian network and its extensions

The *Naive Bayesian network* (N-BN) model with a single parent and not interconnected children has a long history of successful applications both for numeric and nominal variables and both in the regression, but particularly in the classification approach. The independence structure of a naive BN over the output variable $Y(= X_0)$ and the n potential input variables $X_1, \ldots, X_n$ satisfy the following constraint $X_i \perp\!\!\!\perp \{X_j : j \neq i\}|Y$ for all input variable X_i. The conditional independence assumption allows the well-known inference computable in $\mathcal{O}(n)$ time

$$\log P(y|\underline{X}) = \sum_{i=1}^{n} \log p(X_i|y) + \log P(y) - \log P(\underline{X}), \tag{9.25}$$

which shows that in the binary and in general in the nominal case, this is a linear classifier (discriminator). However, if the variables are continuous, nonlinear or disconnected regions can arise [79, 75]. The successful applications of the naive BN model as threshold-based classifiers on its regression estimate $p(Y = 1|x)$ in domains violating significantly its assumption prompted theoretical investigations [89, 75].

Because of the frequently untenable assumption of the N-BN model, various extensions were proposed to increase its representational power w.r.t. $p(Y|\underline{X})$:

the *Tree Augmented Naive Bayesian Network* (*TAN*) and the contextual multinet [91], the semi-naive and the *Augmented Naive Bayesian Network* classifier [75, 150, 174], the *Bayesian Network Augmented Naive Bayesian Network* (BAN) [48], and the *Hierarchical Naive Bayesian Network* [164]. We used the TAN and the BAN models because of their robust performance, low computational complexity and their ability to incorporate prior information.

The TAN model class (over the output variable Y and the $n-1$ potential input variables $X_1, \ldots, X_{n-1}$) is defined as complete N-BNs augmented with a complete tree over the input variables. This avoids the filtering of input variables and it allows globally and conditionally optimalized insertion of $n-1$ edges keeping the maximum number of parents $k \leq 2$ [91].

The Bayesian Network Augmented Naive Bayesian Network (BAN) model class is defined as complete N-BNs augmented with a general Bayesian network tree over the input variables, possibly with certain restrictions such as the maximum number of parents. In this case, there are no constructive methods, so general search methods have to be applied.

9.5 Logistic regression and its extensions

After the discussion of domain model based classifiers, now we overview a conditional parametric model class applicable if Eq. 9.2, 9.9 can be assumed, including the logistic regression and multilayer perceptron. These are well-performing parametric regression models over nominal and continuous inputs as well, which makes them an ideal candidate for investigating the support of their construction using annotated Bayesian networks. The interpretation of the logistic regression model allows a more direct support for its construction methodology, whereas the multilayer perceptron with its increased expressive power requires special supportive techniques, both discussed in Chapter 10, particularly to construct informative parameter priors. In each model class, we used the hybrid Monte Carlo Markov Chain to perform predictive inference (see Section 2.3.2).

9.5.1 Logistic regression

Logistic regression is a standard choice for the investigation of the structural and parametric aspects of the conditional relation of an output variable Y from the input variables $\underline{\underline{X}}$ based on observational data or even data from case-control studies. In case of binary output with values $y, \bar{y}$ and binary inputs with values $x_i, \bar{x}_i$, the model without interaction terms (LR-I) includes the odds ratios corresponding to the inputs x_i, $i = 1, \ldots, n$ and a bias (or intercept) term Ψ_0 ($x_0 \triangleq 1$):

$$\Psi_i = \frac{P(y|x_i)P(\bar{y}|\bar{x}_i)}{P(\bar{y}|x_i)P(y|\bar{x}_i)} \triangleq \exp^{\beta_i} \tag{9.26}$$

and defines for a given $\underline{x}$ the odds $P(y|\underline{x})/P(\bar{y}|\underline{x})$ as their multiplicative combination

$$P(y|\underline{x})/P(\bar{y}|\underline{x}) = \prod_{i=0}^{n} \Psi_i^{x_i} \quad (9.27)$$

$$\log(P(y|\underline{x})/P(\bar{y}|\underline{x})) = \sum_{i=0}^{n} \beta_i x_i \quad (9.28)$$

$$P(y|\underline{x}) = \sigma(\sum_{i=0}^{n} \beta_i x_i), \quad (9.29)$$

in which $\sigma()$ denotes the logistic sigmoid function $\sigma(x) = 1/(1+e^{-x})$. In the general case the mixture of binary and continuous inputs is allowed and higher-order interaction terms:

$$P(y|\underline{x}) = \sigma[\sum_{i=0}^{n}(\beta_i x_i + \sum_{j=1}^{n}(\beta_{i,j} x_i x_j + \sum_{k=1}^{n}(\beta_{i,j,k} x_i x_j x_k + \ldots)))], \quad (9.30)$$

which shows that, based on the LR-I model, a linear discriminating function can be defined.

The Eq. 9.29 for the logistic regression model can be derived under various assumptions [138, 36, 125]. The most general view is that the formula in Eq. 9.30 can be seen as a regression model in a two-class problem assuming binomial noise as in Eq. 9.10, in which case model fitting means the maximization of the conditional likelihood (i.e., the cross-entropy error function):

$$\begin{aligned} p(D_N|\underline{\beta}) &= \sum_{i=1}^{N} y_i \log(\hat{p}(y=1|\underline{x}_i, \underline{\beta})) + (1-y_i)\log(1-\hat{p}(y=1|\underline{x}_i, \underline{\beta})) \\ &= \sum_{i=1}^{N} y_i \log \frac{(\hat{p}(y=1|\underline{x}_i, \underline{\beta}))}{(1-\hat{p}(y=1|\underline{x}_i, \underline{\beta}))} + \sum_{i=1}^{N} \log(1-\hat{p}(y=1|\underline{x}_i, \underline{\beta})) \\ &= \sum_{i=1}^{N} y_i \underline{\beta}\underline{x}_i - \sum_{i=1}^{N} \log(1+\exp(\underline{\beta}\underline{x}_i)), \end{aligned} \quad (9.31)$$

or the L_2 error (see [36],p.247 for conditions on error functions for interpreting the output as probabilities)

$$p(D_N|\underline{\beta}) = \sum_{i=1}^{N}(y_i - p(y=1|\underline{x}_i, \underline{\beta}))^2. \quad (9.32)$$

Another derivation is based on the assumption that the class-conditional distributions of the independent variables $p(\underline{X}|Y)$ are normal with equal covariance matrices, $N(\underline{\mu}, \underline{\underline{\Sigma}})$. In this view, model fitting can be seen as a result of a maximum likelihood estimation based on the class probabilities $\hat{p}(y)$ and $\underline{\hat{\mu}}, \underline{\underline{\hat{\Sigma}}}$,

which determines the logistic regression parameters or as a result of finding an optimal linear discriminant function.

A similar derivation of a logistic regression model, which we use later, is based on the assumption of independent binary features (i.e., on the Naive BN). Applying the Bayes rule to invert Y and $\underline{X}$, we get a formula with parameters that appear directly in a Naive Bayes model (i.e., $\Psi_i = \frac{P(x_i|y)P(\bar{x}_i|\bar{y})}{P(\bar{x}_i|y)P(x_i|\bar{y})}$), and we can express Ψ_0 also in terms of the parameters of the Naive Bayes model by rewriting Eq. 9.27:

$$\frac{P(y|\underline{x})}{P(\bar{y}|\underline{x})} = \frac{P(y)\prod_{i=1}^{n} p(x_i|y)}{P(\bar{y})\prod_{i=1}^{n} p(x_i|\bar{y})} \tag{9.33}$$

$$= \frac{P(y)}{P(\bar{y})}\prod_{i=1}^{n}\frac{P(x_i|y)}{P(x_i|\bar{y})}^{x_i}\frac{P(\bar{x}_i|y)}{P(\bar{x}_i|\bar{y})}^{1-x_i} \tag{9.34}$$

$$= \prod_{i=1}^{n}\underbrace{\frac{P(x_i|y)P(\bar{x}_i|\bar{y})}{P(x_i|\bar{y})P(\bar{x}_i|y)}}_{\Psi_i}^{x_i}\underbrace{\frac{P(y)}{P(\bar{y})}\prod_{i=1}^{n}\frac{P(\bar{x}_i|y)}{P(\bar{x}_i|\bar{y})}}_{\Psi_0} \tag{9.35}$$

Multiple observations are in order. First, clearly an LR model without interaction terms and a Naive BN model can be conditionally equivalent and the parameter transformation is local and transparent, but not one-to-one (for the bijective relation between the conditional part of a Noisy-OR classifier and LR model, see [252]). However, the LR model is basically conditional with $n+1$ parameters, a special case of the causal local dependency model [130]

$$p(Y = 1|\underline{X}) = f(g_1(X_1), \ldots, g_n(X_n)), \tag{9.36}$$

whereas the Naive BN models is the joint distribution with $2n+1$ parameters (the input distribution is the difference).

9.5.2 The relation between MBG and LR models

If a distribution contains additional dependencies w.r.t. a Naive BN, then the induced conditional distribution cannot be represented by an LR model without interaction terms. More specifically we can state the following constraints on the LR model in case of a complete domain model represented by a BN. First note that assuming complete data $\underline{x}$ the conditional aspects are completely represented by $\mathrm{MBG}(Y, G)$, so it is enough to consider the parametric constraints of $\mathrm{MBG}(Y, G)$ instead of G on the LR model (see Proposition 7.2.1 and Proposition 7.2.2). About the MBG contraints we can state the following two lemmas.

Lemma 9.5.1. *A Markov Blanket subgraph* $\mathrm{MBG}(Y, G, \underline{\theta})$ *with binary variables can be transformed to a BAN Bayesian network that is equivalent w.r.t. the conditional* $p(Y|\underline{X}, G, \underline{\theta})$.

Proof. Let us define a BAN model over the variables $X_i \in \mathrm{MBG}(Y, G)$ that $Y = X_0$ is the root node, the children of Y and their parental sets are identical, and the $X'_i \in (\mathrm{pa}(Y, G) = \underline{X}')$ are converted to a child clique of X_0 (i.e., to a completely connected subgraph, in which each X'_i is connected to Y), treating them as an aggregated mega-variable. By setting the new conditionals to $p(\underline{X}'|Y) \propto p(Y|\underline{X}')/p(Y)$ (and, for example, $p(Y) = p(\neg Y) = 0.5$), we have that

$$\begin{aligned} p(Y|\underline{X}, \mathrm{MBG}(Y)) &\propto p(Y|\underline{X}')p(\underline{X}^c|Y, \underline{X}') \\ &\propto p(\underline{X}'|Y)p(Y)p(\underline{X}^c|Y, \underline{X}') \propto p(Y|\underline{X}, BAN), \end{aligned} \tag{9.37}$$

where $\underline{X}^c = \underline{X} \setminus (\underline{X}' \cup \{X_0\})$. □

This MBG-to-BAN transformation is not minimal (e.g., in case of a special $p(Y|\underline{X}')$ it is possible that only one of the X'_is has to be connected to Y and the others are only directed into this node).

Lemma 9.5.2. *A BAN Bayesian network $(G, \underline{\theta})$ with binary variables can be transformed to logistic regression model with interaction terms that is equivalent w.r.t. the conditional $p(Y|\underline{X}, G, \underline{\theta})$.*

Proof. Let us the odds as follows

$$\frac{p(Y|\underline{X})}{p(\bar{Y}|\underline{X})} = \frac{p(Y)}{p(\bar{Y})} \prod_{i=1}^{n} \frac{p(X_i|\,\mathrm{Pa}(X_i))}{\bar{p}(X_i|\,\mathrm{Pa}(X_i))} = \frac{p(Y)}{p(\bar{Y})} \prod_{i=1}^{n} Q_i(X_i, \mathrm{Pa}(X_i)), \tag{9.38}$$

where $\bar{p}$ denotes $Y = 0$ and the factor Q_i is the product of all the corresponding $2^{|\,\mathrm{Pa}(X_i)|+1}$ configurations, selecting always the active as $k = |\,\mathrm{Pa}(X_i)|$:

$$\begin{aligned} Q_i(X_i, \mathrm{Pa}(X_i)) &= \frac{p(\bar{x}_i|\,\mathrm{pa}(X_i) = \underline{0})}{\bar{p}(\bar{x}_i|\,\mathrm{pa}(X_i) = \underline{0})}^{(1-X_i)(1-\mathrm{pa}(X_i)_1)\ldots(1-\mathrm{pa}(X_i)_k)} \\ &\vdots \\ &= \frac{p(x_i|\,\mathrm{pa}(X_i) = \underline{1})}{\bar{p}(x_i|\,\mathrm{pa}(X_i) = \underline{1})}^{(X_i)(\mathrm{pa}(X_i)_1)\ldots(\mathrm{pa}(X_i)_k)} \end{aligned} \tag{9.39}$$

By collecting the exponents, this directly defines the coefficients for the single terms and the interaction terms, also showing that the largest interaction term with variable X_i is maximized by the largest clique size after moralizing the BAN network (i.e., connecting all the parents and dropping the orientations). □

Again, this is sufficient, but not minimal, for example because of causal local conditional models and contextual conditional independencies in the BAN or because of the encoded unstable distribution.

In summary the two step mapping in Lemma 9.5.1 and Lemma 9.5.2 of the BN (MBG) to an LR model shows that the conditional independencies, as high-level common constraints on the model spaces, create links between the structure of a BN model and the parametric structure of an LR model. It

allows the use of BN features in the LR model construction and interpretation, as reported in Section 10.4.

Regarding the optimization of the model parameters, note that in the LR model the parameters are optimized by maximizing the conditional likelihood of the data (e.g., we use the scaled conjugate gradient method for the LR and MLP models [190]; for an overview of methods, see [6]), whereas in the BN model the parameters are optimized by maximizing the likelihood of the complete data by setting the observed frequencies. More significant differences arise in the Bayesian framework as different priors are used in the LR model (Laplace or Gaussian) and in the BN model (Dirichlet).

In the hypothesis testing framework, the model fitting and selection is problematic both from the computational point of view and more seriously from the statistical point of view in case of large number of potential inputs and of a small sample size. The first problem of the search for potential candidates can be automated using standard forward and backward stepwise scheme with heuristics [223, 215, 216]. But the large number of tests to explore the huge hypothesis space (2^n) worsens the second problem to ensure an overall significance level. Consequently, the proper exploitation of the prior domain knowledge can be crucial both from the computational and from the statistical point of view, but in the hypothesis testing framework the use of prior knowledge is problematic. For the construction of informative priors to support the Bayesian application of the MLP and LR models, see Chapter 10.

9.5.3 The multilayer perceptron extension

Another view of the LR model is that it is a baseline model of further extensions to allow adaptive, unconstrained higher-order interaction terms resulting in the *multilayer perceptron* (*MLP*). A multilayer perceptron defines a complex nonlinear input-output mapping defined by consecutive layers of summation and elementary nonlinear mappings forming a feedforward structure without feedback, ensuring arbitrary approximation capacity even with one hidden layer containing a sufficient number of units [129, 36, 79]. For example an MLP with one hidden layer with L units is given as follows:

$$f(\underline{x}, \underline{\omega}) = \sigma[\sum_{i=1}^{L}(\omega_i \ \tanh[\sum_{j=1}^{|\underline{X}|}(\omega_{ij}x_j + \omega_{i0})])],$$

in which the activation or transfer function is a hyperbolic tangent or the logistic function $\sigma(x) = 1/(1+e^{-x})$; and $\underline{\omega}$ contains all the parameters including the bias parameters ω_{i0}. The application of the logistic function ensure that $f(\underline{x}, \underline{\omega}) : \mathcal{R}^d \rightarrow [0, 1]$.

Chapter 10

Bayesian classifiers with a prior domain model

We discuss the modeling of the uncertain dependency relation of an output variable from a set of input variables, if informative prior knowledge is available formalized using Bayesian networks. We show how can we use a distribution $p(G, \theta)$ over Bayesian networks as a probabilistic knowledge base for this purpose, specifically how can we link this distribution to the distribution over a parametric classification model $p(S, \omega)$. This specification of a joint distribution can be seen as prior transformation in our context. We then report the comparison of the different prior transformation methods by presenting the performance of learning methods for Bayesian networks and multilayer perceptrons.

The modeling and understanding of the dependency relation of output variables from the input variables frequently can be done at least as an approximation in the conditional Bayesian approach without complete domain modeling. However, the validity of this approximation is frequently not explored and the application and the fair evaluation of conditional model classes is hindered by the lack of methods for incorporating prior knowledge into black-box classifiers. This is in contrast with the availability of techniques and frequently resources for constructing probabilistic domain models, in which such classification submodels are grounded. We argue for the coexistence of domain and classifier models, particularly for the supportive role of a domain model through the whole process of classifier construction. We propose a two-step hybrid method for probabilistically linking these model spaces, specifically for inducing informative priors for classifiers. First, we review the analytic transformation of priors for Bayesian Network classifiers. Then we describe our developed conditional distance minimization method for inducing informative parameter priors for parametric black-box classifiers, such as logistic regression and multilayer perceptron models from Bayesian networks. We compare this method with prior transformation methods based on a virtual sample and other transformation techniques. Then we present the joint posteriors of various conditional features

and performance measures using the domain model, which also allows inducing structure priors for properties of regression models, such as for its inputs and for its complexity. Next, we evaluated the classification performance of Bayesian network classifiers and logistic regression models with informative priors.

10.1 Reasons for the dual representation

In the thesis we make the following assumptions about the classification task we consider:

1. The classification is binary with continuous and nominal input variables and the prediction of the class probability and of an uncertainty measure about this probability is advantageous.

2. The size of the sample is small or medium with respect to the learnability of the problem and missing data are infrequent.

3. A large amount of prior knowledge is available about the domain, the variables, the dependencies between variables and the quantification of these dependencies.

These assumptions are inspired partly by the ovarian tumor problem and our mathematical derivations will be formulated in this context. It is however important to stress that the methods proposed can straightforwardly be extended to multiclass classification and to regression.

If standard statistical tools, such as logistic regression, do not give satisfactory results, we need to use more complex models like data-driven black-box methods (such as MLPs, decision trees and kernel-based methods) or more knowledge-oriented white-box methods (such as BNs).

In the case of black-box methods, the possibilities to incorporate prior knowledge in the model or in the learning process are limited, even though this incorporation is frequently essential. It is generally confined to the selection of the input variables, of the model structure and of the learning algorithm and to the management of missing values. In the Bayesian context, an inherent problem for black-box parametric models is that it is not possible to directly construct an informative prior distribution.

In the case of white-box methods, particularly for Bayesian networks, the possibilities to incorporate domain knowledge in the model are greatly enhanced. A prior distribution for the parameterization of a given model structure or for the model structures can be constructed by established methods (see Sections 3.1.5.1 and 3.1.5.2). However, the sample complexity of parameter learning is in practice frequently higher than for black-box models. The full scale Bayesian inference or general structure learning is hindered by the superexponential cardinality of the structure space and the high sample complexity (see Section 3.5). Additionally, the general structure learning methods — the data dependent terms and regularization terms — are optimal for learning the joint

distribution. It means they are more appropriate to solve a much harder task than is necessary in a standard classification (see Section 9.4.1). As a solution, special Bayesian network classifiers with restricted structures were suggested (see Section 9.4). Another approach is to combine the predictions of multiple models, such as the conditional model and the domain model [253] or to combine the regression model into the domain model [94]. Paradoxically, an additional problem for domain models, because the prior domain models is usually formalized using a discretization scheme for the continuous variables, is that the learning process has to refine this jointly with learning the structure and parameter priors (for an integrated BN learning scheme, see [93]). Finally in general domain models the computational complexity of inference is typically much higher both in the frequentist and Bayesian framework.

Beside the research on Bayesian network classifiers, the sequential application of the white-box and black-blox techniques arises from the standpoint of black-box learning. The appearance of learning theories [247, 37, 73] made it possible to formalize how the incorporation of domain knowledge in inductive techniques reduces the statistical complexity of learning (in the classical statistical context [126, 109, 2] and in the Bayesian context [127]). On the practical side, Abu-Mostafa [2] and Niyogi et al. [198] reported methods for exploiting a priori known regularities and symmetries in the input space. Another approach, the Knowledge-Based Artificial Neural Network, used the prior knowledge for selecting an appropriate multilayer perceptron architecture [241]. This method formalizes the domain knowledge in propositional logic to construct the structure of a multilayer perceptron. Further works reported results about the inductive refinement of the initial network structure, the extension of the translation and transformation of symbolic rules into a feedforward artificial neural network (for surveys about Knowledge-Based Neurocomputing, see [222, 53]). Sowmya et al. [225] extended the symbolic paradigm for the transformation of domain theories into a feedforward neural network by proposing Bayesian networks with certain local models for knowledge modeling. Another proposal [193] similarly emphasized the appropriateness of Bayesian networks for prior knowledge formalization and described a mapping of Bayesian networks onto stochastic neural networks to support parallel computations. The potential of the Bayesian framework using Bayesian networks as domain models for supporting the construction of a classifier, particularly for inducing priors for black-box classifiers was described in [10]. The methodology proposed in this paper generalizes the earlier symbolic methods and the frequentist Bayesian network based methods to the Bayesian framework.

Because small or medium size samples are frequent in medical problems and a good prior Bayesian network was available [28], we decided to investigate the incorporation of prior knowledge into a multilayer perceptron as an evaluation of a general methodology for such problems. Partly because general structure-learning algorithms have failed in our preliminary experiments to achieve a good quantitative performance, while multilayer perceptrons reached nearly the performance of expert diagnosticians [237], but only in the large sample region (i.e., for all the cases occurred in two years in a large, referral medical center).

Our two-step methodology combines certain complementary advantages of Bayesian networks and multilayer perceptrons by (1) formalizing the prior knowledge with the Bayesian network and (2) incorporating the formalized knowledge into the Bayesian learning and inference of the multilayer perceptron (see Fig. 1.6). The use of a black-box method in the inductive phase provides a computationally and statistically efficient solution to refine jointly the a priori structure, the prior over the parameters, and the a priori discretization corresponding to the prior Bayesian network. The existence of a prior domain model also allows a more refined management of missing values [12].

Furthermore, if a distribution over causal Bayesian networks is available as a probabilistic knowledge base, then we can link this distribution to the distribution over a parametric classification model space $p(S, \omega)$, which allows the exploration of the validity of the conditional model class, inducing parameter and structure priors for it and supporting its interpretation. Indeed, it shows that the Bayesian network methodology is not only an alternative to the "black box approach" of classifier construction, but it provides a general complementary tool to support the complete process of classifier construction.

10.2 Parameter priors for Bayesian classifiers

In this section we will assume (1) that we have a parameter prior $p(\theta|G_0)$ formalized for a transparent, domain model G_0, whose fixed model structure is restrictive to ease knowledge elicitation and (2) that our goal is to derive a parameter prior $p(\omega|S)$ for a parametric black-box model with an arbitrary structure S, which is more powerful w.r.t. modeling the conditional distribution. More specifically, we assume a parameter prior $p(\theta|G_0)$ for given Bayesian network structure G_0 and our goal is to define and analytically compute or numerically approximate an informative domain model-based posterior $p(\omega|S)$ for any MLP structure S. In this context such a link between parameter spaces for different models corresponds to the transfer of prior information from a semantically transparent model class to a "black-box" model class, but these methods can be equally interpreted as general transformation methods supporting fair Bayesian comparison of models by ensuring the same priors for each model [195].

The methods can be divided into two groups. The first group contains the so-called prior or virtual sample based methods, which assume that the prior knowledge $p(\theta|G_0, \xi^+)$ can be expressed as the posterior update of a prior $p(\theta|G_0, \xi^-)$ with an observed single prior virtual data set $D^+_{N'}$ or more generally that the prior can be expressed as a distribution $p(D^+_{N'})$ over data sets with size N'. We will summarize the conditions for this assumption. This assumption allows the usage of such a prior (e.g., the prior data set) to update the target parameter prior $p(\omega|S, \xi^-)$ to a posterior $p(\omega|S, \xi^-, D^+_{N'})$ as well, and use this virtual posterior as an informative prior $p(\omega|S, \xi^+)$.

The second group contains methods based on mapping $\mathcal{T} : \mathbf{\Theta} \rightarrow \mathbf{\Omega}$ that transforms a prior distribution $p(\theta)$ over the Bayesian network parameter space into a prior probability distribution $p(\omega)$ over the black-box model parameter

space. An illustration of this idea is a bijective mapping between the conditionally relevant parameters of a discrete Bayesian network and the parameters of a logistic regression model (see Section 9.5.2). The outline of the general mapping is the following: the black-box model $f_\omega(\mathbf{x})$ is used for approximating the conditional distribution of the output class $P(c_1|\mathbf{x})$ conditioned on the input $\mathbf{x}$, which is defined by the Bayesian network. Thus we can define a mapping from every Bayesian network parameter $\theta \in \mathbf{\Theta}$ to the "best" approximating black-box function parameter $\omega \in \mathbf{\Omega}$. Note that this is not a marginalization because of its approximative nature. This method can be seen as the asymptotic version of a prior sample method, though the role of the data set here is purely technical, outside the Bayesian context (e.g., the sample size does not express confidence). Then we either directly use this induced distribution in Bayesian inferences or approximate $p(\omega)$ with a mixture of Gaussians $\sum_{i=1}^{L} \alpha_i N_t(\omega|\underline{\mu}_i, \underline{\underline{\Sigma}}_i)$.

We also discuss the relation of these methods. In Section 10.6, we present results about the effect of parameter priors on classification in ovarian cancer using the informative prior from the domain expert described in Section 8.2.1.

10.2.1 Prior transformation between BNs

First we discuss the derivation of parameter priors for Bayesian network classifiers, such as for the Naive BN (used with exact model averaging) and for the TAN and BAN models (in each case the mean parameterization is used, see Section 3.2.1.2 for its discussion in the conditional approach). As these classification models are standard Bayesian networks, we can use the general results from Section 3.1.5.1 for parameter prior elicitation and transformation for Bayesian networks. Specifically, Th. 3.1.6 states that parameter independence and modularity with likelihood equivalence implies that the parameter prior is the product of Dirichlet priors with hyperparameters $N'p(X_i = k, \mathrm{pa}(X_i, G) = \mathrm{pa}_{ij}\,|\xi^+)$, where $p(\underline{V}|G_c, \xi^+)$ is a point parameter for a complete or for maximally detailed model $p(\underline{\theta}|G_c, \xi^+)$ and N' expresses confidence by specifying a prior sample size N' representing the complete cases underlying the point estimates. Then for any classification Bayesian network G we can compute its hyperparameters by the marginalization of the distribution $p(\underline{V}|G_c, \xi^+)$ into $p(\underline{V}|G, \xi^+)$ and multiplying the appropriate local probability models with N'. In the "counting" interpretation of the Dirichlet hyperparameters this means the if a prior virtual data set $D^+_{N'}$ is available then the prior distribution $p(\underline{V}|G_c, \xi^+)$ is set to the relative frequencies or vice versa the prior distribution $p(\underline{V}|G_c, \xi^+)$ and N' can be used to reconstruct a prior virtual data set $D^+_{N'}$ or we can define a distribution over prior virtual data sets as

$$p(D^+_{N'}|\xi^+) = \int \prod_{l=1}^{N'} p(D_l|G_c, \theta) p(\theta|G_c, \xi^+, N')\,\mathrm{d}\theta, \tag{10.1}$$

where $p(\theta|G_c, \xi^+, N')$ is the product of Dirichlets with the hyperparameters defined above ($N'p(X_i = k, \mathrm{pa}(X_i, G) = \mathrm{pa}_{ij}\,|\xi^+)$). Though this may seem an unnecessary complication in this context, the prior distribution over (prior) data

sets offers a more general method for expressing prior knowledge than a single prior data set as we discuss in Section 10.2.3.2.

10.2.2 Noninformative priors for LRs and MLPs

The parameters of the multilayer perceptron are the weights and biases of the composing neurons at layer k $(\omega_{ki}, \omega_{k0})$, which are hard to interpret in a multilayer model. This prohibits the incorporation of prior knowledge into this model by directly specifying a *prior* distribution over the parameters (versus the intuitive interpretation of the Bayesian network or logistic regression parameters). In practice, the prior is used only for controlling the complexity of the model class, because the implementable functions are equally dependent on the number of neurons and on the size of the parameters. The first factor is related to the overall irregularity of the function, such as disconnected regions with some $\tau < f(\underline{x}, \underline{\omega})$. The second factor is more related to abrupt local changes, discontinuities as the activation functions have a more and more stepwise form with increasing weights. In fact, large weights typically arises in MLP models overtrained to a particular data set in the frequentist optimization framework, which impairs the inductive capacity of the model for new cases. This lead to various weight-decay regularizers such as

$$\sum_i \omega_i^2 \text{ or } \sum_i |\omega_i|, \tag{10.2}$$

and to the corresponding Gaussian and Laplacian priors assuming independent parameters (e.g., see [36, 110])

$$p(\omega_i|N(0,\sigma_i)) \propto \exp(-\omega_i^2/\sigma_i^2) \text{ or } p(\omega_i|\lambda) \propto \exp(-\lambda|\omega_i|). \tag{10.3}$$

We used Gaussian priors, so we follow this terminology. Because of the assumed zero covariance terms, the goal is the specification of the variances σ_i^2. In the simplest approach the same variance is used for all parameters, leading to the Gaussian prior distribution $N(0, \sigma^2 \underline{\underline{I}})$. However, this prior does not recognize the different effect of the bias terms and the weights on each layers, which would lead to different variances for each such group [36]. Another refinement is to define a hyperprior for the variances for example using the Gamma distribution [194]. Because our goal was to investigate the effect of incorporating informative priors, we used the Gaussian prior with the same variance. For the usage of different variances and hyperpriors for MLPs in overian cancer and for the usage of the full covariance structure of the parameters through modeling their distribution with a continuous Gaussian Bayesian network, see [85].

10.2.3 Informative MLP prior from a Bayesian network

If the encoded prior knowledge in the Bayesian network comes from N previously seen complete cases, then this prior data set can be used as well as a real, potentially weighted data set in the Bayesian inference with the multilayer

perceptron. Therefore, we examine the possibility of generating a *prior data set* from the domain model. The generalization of this method is to allow a prior over prior data sets [195].

However, if the Bayesian network is hyperparametrized from heterogeneous sources (e.g., various parts of the model are quantified by different experts or studies), then such prior complete data sets are not appropriate. Even if a complete prior data set is available and reconstructed, then its direct usage as real data loses certain aspects of the prior uncertainty modeling (e.g., the Dirichlet assumption). So we introduce a second method that directly transforms the domain knowledge encoded in the prior distribution of the Bayesian network into an *informative prior (conditional) distribution* for the multilayer perceptron.

First we overview transformation methods, then we describe three supplementary methods used with all transformation methods: (1) a probabilistic smoothing schemes to convert the discrete values into continuous ones, (2) an estimation scheme to approximate the transformed informative priors with mixture of Gaussians, and (3) symmetry elimination schemes to cope with symmetries in the MLP parameter space.

10.2.3.1 Using a prior data set

In the context of eliciting and constructing parameter priors for Bayesian network the concept of prior or virtual data set has a widespread use because of the "counting" interpretation of the hyperparameters of the frequently used Dirichlet distributions. The elicitation of a prior data set to enhance the classification performance of a Bayesian network was reported in [179]. The usage of data sets from earlier studies to define an informative prior for logistic regression models was reported in [45]. In our approach we assumed that this prior data set is implicitly formalized in the elicited prior domain model formalized as a Bayesian network.

To illustrate the effect of the prior sample, let us first assume that the prior domain knowledge consists of N complete cases, called the prior sample $D^+_{N'}$, which will be used together with the real data D_N. Assuming a noninformative prior distribution $P(\omega|\xi^-)$ for the multilayer perceptron, the Bayesian update is defined as follows:

Definition 10.2.1.

$$P(\omega|D^+_{N'}, D_N, \xi^-) \propto P(D_N|\omega)P(\omega|D^+_{N'}, \xi^-) = P(D_N|\omega, \xi^-)P'(\omega|\xi^-). \quad (10.4)$$

This grouping of the terms illustrates that the prior sample $D^+_{N'}$ transforms the noninformative prior distribution $P(\omega|\xi^-)$ into an informative prior $P(\omega|\xi^+)$ ($\triangleq P(\omega|D^+_{N'}, \xi^-)$), so we call this distribution transformation a *prior sample transformation (PS-T)*. If the prior data set follows the real conditional distribution, then the effect of this Bayesian update by the prior sample is the same as by real data. A problematic issue is the selection of the prior sample size — particularly, if the prior Bayesian network is heterogeneously hyperparametrized. It

is difficult in general, because it means selecting an optimal complexity regularization, consequently it can be influenced by the real sample size, the problem, the general correctness of the prior domain model and in practice even by the inference scheme. In our domain the results of this method proved robust for a virtual sample size in the range of 15 to 50 (cf. Fig.8.2).

In our experiments, we use a stochastic scheme to generate a prior sample equivalent to N' samples. We fix the mean parameterization in the Bayesian network and use it to generate the i.i.d. complete samples. Instead of generating a sample of size N', we generate a larger number of prior samples that we rescale to an *effective sample size* in the update and inference process. This approach reduces the impact of stochastic effects.

10.2.3.2 Using a prior over data sets

This method is the generalization of the prior data set method by a distribution over prior virtual data sets $p(D^+_{N'})$ with a given size N' instead of a fixed data set $D^+_{N'}$, specified as

$$p(D^+_{N'}|\xi^+) = \int p(D^+_{N'}|\theta)p(\theta|\xi^+)\,\mathrm{d}\theta = \prod_{l=1}^{N'} p(D^+_l|\theta)p(\theta|\xi^+)\,\mathrm{d}\theta, \tag{10.5}$$

which suggests to call this distribution transformation a *prior over prior samples transformation (PPS-T)*. It was proposed by R.M.Neal in [195], assuming a somewhat different context than ours. First, his approach assumes that the "donor" model with $p(\theta)$ is more restrictive than the "recipient" model with $p(\omega)$, making the donor model more appropriate for parameter elicitation. This differs from our domain model and classification model assumption and that interpretability differentiates between the model spaces. Actually, the domain model can be more complex, even its conditionally relevant subpart. Second, he rightly identified the issue of fair model comparison as an equally important motivation for such prior transformation methods, besides our goals to derive informative priors for enhancing predictions and to investigate the probabilistic link between a domain model and a classification submodel. Third, he propose the usage of a sample size N' with an appropriately selected scaling to exploit the "fuzzifying" effect of finite data sets, which is different from our earlier reported method in [10, 18] corresponding to its asymptotic case.

The transformed informative prior is as follows

$$\begin{aligned} p(\omega|\xi^+) &\triangleq \sum_{D^+_{N'}} p(\omega|D^+_{N'},\xi^-)p(D^+_{N'}|\xi^+) && (10.6)\\ &= \sum_{D^+_{N'}} p(\omega|D^+_{N'},\xi^-)\int p(D^+_{N'}|\theta)p(\theta|\xi^+)\,\mathrm{d}\theta && (10.7)\\ &\propto \sum_{D^+_{N'}} p(\omega|\xi^-)p(D^+_{N'}|\omega)\int p(D^+_{N'}|\theta)p(\theta|\xi^+)\,\mathrm{d}\theta. && (10.8) \end{aligned}$$

That is we assume that all our prior information is transferred by the prior data sets $D^+_{N'}$with size N'.

The definition of the transformed informative prior $p(\omega)$ as the marginal of the joint $p(\omega, D^+_{N'})$ defined conditionally by $p(\omega|D^+_{N'})$ and $p(D^+_{N'})$ suggests the following scheme for sampling $p(\omega)$: sample the joint probability distribution by sampling $p(D^+_{N'})$, and sequentially sampling $p(\omega|D^+_{N'})$ (and discard the prior data set). In the general case if there is no closed form for $p(D^+_{N'})$ to sample it directly, then the prior is the marginal of $p(\omega, D^+_{N'}, \theta)$ defined as $p(\omega|D^+_{N'})p(D^+_{N'}|\theta)p(\theta)$ in Eq. 10.5, which allows sampling $p(D^+_{N'})$ again conditionally. The complete conditional sampling scheme consists of the following three steps:

Algorithm 2 Sampling MLP parameters from a BN through data sets.

1, Sample Bayesian network parametrizations $\{\theta_1, \ldots, \theta_l\}$.
2, Sample blocks of prior samples from the parametrizations $\{D^p_1, \ldots, D^p_l\}$.
3, Sample multilayer perceptron parametrizations from the posterior based on the blocks, resulting in a block of MLP parametrizations $\{\omega_1, \ldots, \omega_l\}$.

In the first and second phase we applied direct sampling, in the third phase the hybrid MCMC was used from Section 2.3.2 (for a discussion of various MC schemes, see [195]). For applications, see [15, 14])

As in the method using a single prior data set it is possible to reduce the impact of stochastic effects while keeping the effect of the prior sample on a prespecified level by generating a larger data set with size N'' and rescale it to an *effective sample size* N' in the update and inference process (e.g., by $p(D'_{N''}|\underline{\omega})^{\frac{N'}{N''}}$).

10.2.3.3 Using conditional distance minimization transformation

Let us consider the asymptotic behavior of the prior transformation method defined by Eq. 10.7 if N' goes to infinity. It means that for a θ drawn from $p(\theta)$, we draw ω from

$$\lim_{N' \to \infty} p(\omega|D^+_{N'}), \text{ where } p(\omega|D^+_{N'}) \propto p(D^+_{N'}|\omega)p(\omega). \tag{10.9}$$

If the parameter for the conditional model ω would correspond to a discrete domain model specifying a joint distribution, which domain model is nested in the set of distributions over this domain encoded by θ, then according to Sanov's theorem under the i.i.d. sampling from $p(D^+_{N'}|\theta)$ the probability that the empirical joint distribution is in this smaller set depends roughly exponentially from the distance of the best approximation of θ (see [59],p.292)

$$\inf_{\omega} \mathrm{KL}(p(V|\theta)\|p(V|\omega)). \tag{10.10}$$

To see the effect of increasing N' in our conditional context let us rewrite

the transformed prior as

$$p(\omega|\xi^+) \triangleq \int p(\theta|\xi^+) \underbrace{\sum_{D^+_{N'}} p(\omega|D^+_{N'},\xi^-)p(D^+_{N'}|\theta)}_{p_{N'}(\omega|\theta)} \,\mathrm{d}\theta. \tag{10.11}$$

By assuming that $V = X \cup Y$ and that they are discrete variables, the $p(\omega|D^+_{N'})$ term can be rewritten as

$$\begin{aligned}
p(\omega|D^+_{N'}) &= \frac{p(\omega)}{p(y'_{N'}|x'_{N'})} \prod_{l=1}^{N'} p(y'_l|x'_l,\omega) && (10.12)\\
&\propto p(\omega) \prod_X \prod_Y p(y|x,\omega)^{N'_{y,x}} && (10.13)\\
&= p(\omega)(\prod_X (\prod_Y p(y|x,\omega)^{\frac{N'_{y,x}}{N'_x}})^{\frac{N'_x}{N'}})^{N'} && (10.14)\\
&= p(\omega)(\prod_X (\prod_Y p(y|x,\omega)^{\hat{p}_{N'}(y|x)})^{\hat{p}_{N'}(x)})^{N'}, && (10.15)
\end{aligned}$$

as $\hat{p}_{N'}(y|x)$ and $\hat{p}_{N'}(x)$ converge to $p(y|x,\theta)$ and $p(x|\theta)$ by taking logarithm for large enough N':

$$\begin{aligned}
\log p(\omega|D^+_{N'}) &\approx N' \sum_X p(x|\theta) \sum_Y p(y|x,\theta) \log p(y|x,\omega) + c && (10.16)\\
&= N' \,\mathrm{E}_{p(X|\theta)}[\mathrm{E}_{p(Y|X,\theta)}[\log p(Y|X,\omega)]] + c && (10.17)\\
&= N' \,\mathrm{E}_{p(X|\theta)}[\mathrm{E}_{p(Y|X,\theta)}[\log \tfrac{p(Y|X,\omega)}{p(Y|X,\theta)}]] + c' && (10.18)\\
&= N' \,\mathrm{E}_{p(X|\theta)}[-\mathrm{KL}(p(Y|X,\theta)\|p(Y|X,\omega))] + c'. && (10.19)
\end{aligned}$$

This shows that this term is N' times the expected conditional log-likelihood (cross-entropy) of the conditional model parameterized with ω w.r.t. the domain model parameterized with θ or after an expansion with a constant independent of ω, it is N' times the expected KL distance of the conditional model w.r.t. θ:

$$\begin{aligned}
p_{N'}(\omega|\theta) &= \sum_{D^+_{N'}} p(\omega|D^+_{N'},\xi^-)p(D^+_{N'}|\theta) && (10.20)\\
&\approx \exp(N' \,\mathrm{E}_{p(X|\theta)}[\mathrm{E}_{p(Y|X,\theta)}[\log p(Y|X,\omega)]]) && (10.21)\\
&\approx \exp(N' - \mathrm{E}_{p(X|\theta)}[\mathrm{KL}(p(Y|X,\theta)\|p(Y|X,\omega))]). && (10.22)
\end{aligned}$$

Because the effect of a non-restrictive prior becomes negligible (i.e., N'-free) and the Bayes normalizing constant does not depend on ω, asymptotically this determines the transformation θ to ω, which is more and more concentrated around ω^* (assuming uniqueness — which is an important issue discussed later)

$$\omega^* = \arg\max_{\omega} \mathrm{E}_{p(X|\theta)}[\mathrm{E}_{p(Y|X,\theta)}[\log p(Y|X,\omega)]] \quad (10.23)$$

$$= \arg\min_{\omega} \mathrm{E}_{p(X|\theta)}[\mathrm{KL}(p(Y|X,\omega)\|p(Y|X,\theta))]. \quad (10.24)$$

This also shows that we can define directly an asymptotic mapping from θ to ω without sampling and averaging over prior data sets. Based on this concept of minimizing the difference between the predictions of a domain and a conditional model on average we proposed the following transformation [10] (for applications, see [18, 11]).

Definition 10.2.2 ([10, 18]). *Let θ and ω denote the parameters of a domain model and a conditional model. The direct transformation of an informative prior from a domain model into an informative prior over a parametric black box conditional model ($\mathcal{T} : \mathbf{\Theta} \to \mathbf{\Omega}$) is defined as*

$$\mathcal{T}_{\mathrm{KL}}(\theta) = \arg\min_{\omega'} \mathrm{E}_{p(X|\theta)}[\mathrm{KL}(p(Y|X,\omega')\|p(Y|X,\theta))] + c(\omega) \quad (10.25)$$

$$\mathcal{T}_{L_2}(\theta) = \arg\min_{\omega'} \mathrm{E}_{p(X|\theta)}[L_2(p(Y|X,\omega'), p(Y|X,\theta))] + c(\omega). \quad (10.26)$$

The term $c(\omega)$ is a weight regularization term (see Eq. 10.2), which ensures the existence of such parameters (i.e., the existence of parameters corresponding to $\inf_{\omega'} \mathrm{E}_{p(X|\theta)}[d(p(Y|X,\omega'), p(Y|X,\theta))]$, where $d(,)$ denotes the distances KL or L_2). For handling aliasing (underidentified parameters), see Section 10.2.3.5.

We call this distribution transformation a *Conditional Distance Minimization* Transformation (CDM-T). Note that it is not based on prior data sets with a given size N', so the direct method avoids the "fuzzifying" effect of a finite N' in PPS-T method of [195] and avoids the corresponding computational and statistical consequences.

This definition of informative priors for black-box conditional models was proposed by the author in [10], its first application in a real world domain was reported in [18], independently of the PPS-T method [195]. This method applies the same technique as the recently proposed model projection method used for model selection [214], p.370.

The main steps for the practical application of this technique in the case of multilayer perceptron are as follows:

Algorithm 3 Sampling MLP parameters from a BN by direct transformation.

1, Generate Bayesian network parameters $\{\theta_1, \ldots, \theta_l\}$
2, Generate block of prior samples from each parameter $\{D_1^p, \ldots, D_l^p\}$
3, Train a multilayer preceptron for each block of samples resulting in a block of perceptron parameters $\{\omega_1, \ldots, \omega_l\}$

The Bayesian network parameters are generated from the Dirichlet distribution by standard methods. The sample blocks are generated according to the drawn Bayesian network parameters. Note that the role of the size of the prior

data sets technically is to eliminate the uncertainty in the mapping by selecting it to be large enough and not to control the fuzzifying effect in the mapping. For the L_2 based transformation it is advantageous to use the prior probabilistic domain model to compute $P(c_1|\mathbf{x})$ for each sample instead of random generated class labels, to eliminate a stochastic element (the class labels). For training the perceptron model on a block of samples, we used the scaled conjugate gradient algorithm [190]. In the terminology of the prior data sets method, this defines a transformation of the prior probability measure p_θ over the parameter space of the *donor* model (the Bayesian network in our case denoted by BN-I) to a prior probability measure p_ω over the parameter space of the *recipient* model (the MLP model).

10.2.3.4 Discrete-continuous transformations

The prior knowledge is frequently formalized using various discretization schemes for the otherwise continuous domain variables. Consequently the prior domain model built with domain experts may contains one or more discrete variables for the same continuous variable. For example in ovarian cancer, two discretization schemes are in use for the variable $CA125$. Because of the availability of earlier studies for both schemes in certain models we tried to include both versions. However the conditional model is often more powerful for continuous variables than for their discretized versions. Those discretizations are usually influenced by nonstatistical factors and not tailored to a particular data set. This means that all three methods require an additional prior as a smoothing scheme for generating continuous samples, which are using prior data sets $D^+_{N'}$ either technically as the direct transformation method or semantically as the methods using a prior data set or a prior over prior data sets.

The probabilistic smoothing scheme is ideally an integral part of the domain model, allowing complex dependencies between the continuous variables and their discretized variables, but in practice exactly the presence of discrete variables indicates that such background knowledge is not available. In ovarian cancer, we used purely technical probabilistic smoothing schemes treating independently each continuous-discrete variable pairs X^c_i, X_i by specifying fixed separate conditionals $p(X^c_i|X_i)$ independent of the beliefs for the parameter prior θ. The conditional distribution of a smoothed prior data set is

$$p(D^c_{N'}|D^+_{N'}) = \prod_{l=1}^{N'} \prod_{i=1}^{n} p(x^c_{li}|x_{li}), \tag{10.27}$$

with the understanding that the non-discretized variables are left unchanged.

We used two approaches to define $p(X^c_i|X_i = x_i)$, where the interval of the discretization bin x_i is $[l_i, u_i]$. The random sub-bin method mainly for test purposes requires the number of sub-bins s and defines a uniform distribution over the discrete values $l_i + \frac{u_i - l_i}{s}(j + \frac{1}{2})$ for $j = 0, \ldots, s-1$. The uniform method simply defines a uniform distribution over the interval $[l_i, u_i]$.

10.2.3.5 Analytic approximation of the transformed informative prior

The informative parameter priors for conditional models defined by the PPS-T and CDM-T methods does not allow the application of advanced Bayesian predictive methods, because in general there is no closed formula for the (unnormalized) transformed prior. This suggests the analytic approximation of the transformed priors, which can be advantageous for the PS-T method as well, because it eliminates the additional computational cost of using the prior data set. Additional advantage of an analytic prior is the possibility of expressing the certainty measure in the prior knowledge by changing its hyperparameters, which offer richer possibility than selecting an appropriate prior sample size N' in the PS-T and PPS-T methods.

To estimate the transformed prior distribution $p_{\mathrm{MLP-Informative}}(\omega)$ over the black-box model parameter space ω from the trained perceptrons, we used a mixture of Gaussians

$$p(\omega) \approx \sum_{i=1}^{L} \alpha_i N(\omega|\underline{\mu}_i, \underline{\underline{\Sigma}}_i), \quad \text{where} \quad 0 \leq \alpha_i \leq 1, \ \sum_{i=1}^{L} \alpha_i = 1.$$

The use of a Gaussian mixture was motivated by the asymptotic normality of a parameter posterior under broad conditions (see Section 2.2.4.1) and by its analytic tractability in EM methods. However, the conditions for asymptotic normality are violated by the parameters for a given MLP structure, because of the existence of finite number of underidentified parameters that are equivalent w.r.t the likelihood function called *aliasing* (see [108], p.102)

$$\exists \underline{\omega}, \underline{\omega}', \underline{\omega} \neq \underline{\omega}' \text{ such that } \forall\, x: \ f(\underline{x}, \underline{\omega}) = f(\underline{x}, \underline{\omega}'). \tag{10.28}$$

The total number of equivalences (due to possible permutations and sign symmetries) in a multilayer perceptron with k hidden layers and L_i neurons in layer i is given by $\prod_{i=1}^{k} 2^{L_i} L_i!$ and this prior estimation requires proper management of symmetries in the parameter space [217]. We applied a heuristic clustering algorithm exploiting the symmetries to map the parameters into clusters with minimal within-cluster variance [18, 85]. This method avoids the discontinuity of restricting the parameter space to a canonic subspace (i.e., mapping the parameters to a subspace without aliasing).

10.3 Structure priors for Bayesian classifiers

An informative structure prior for general Bayesian networks $p(G|\xi^+)$ can be used directly for Bayesian network classifiers simply by restricting it to a given model class BNC

$$p'(G^{BNC}) \propto p(G) 1(G \in \mathrm{BNC}). \tag{10.29}$$

An informative structure prior can be defined by any standard method discussed in Section 3.1.5.2, and it can incorporate reference models and estimates

for the properties of domain models from the domain expert or the literature-based automatically derived reference models and feature posteriors.

The specification of informative priors for LR and MLP models $p(\omega, S|\xi^+)$ is similarly separated into the specification of the parameter prior $p(\omega|S, \xi^+)$ and into the specification of the structure prior $p(S|\xi^+)$. Because of the interpretability of the LR model, a direct specification is possible for both the parameter and structure priors (for the interpretation of its parameters and (parameter) structure, see Section 9.5.1). However despite its established semantics, the LR model is a conditional, feedforward representation without intermediate variables, so the Bayesian network representation still offers a much richer language for prior specification both parametrically and structurally. We thus consider the question of the specification structure priors both for the LR and the MLP models (denoted equally as $p(S^{\text{MLP}}|\xi^+)$) based on the assumption that an informative structure prior for BNs $p(G|\xi^+)$ is available.

Depending on the form of $p(G|\xi^+)$ the derivation can be as follows. First consider the simplest case when the prior knowledge ξ^+ can be formalized as a prior data set $D^+_{N'}$ (see Section 10.2.3.1 for the analogous the case of parameter priors). Assuming a noninformative parameter prior and structure prior for the parametric conditional model satisfying the conditional modeling requirements, the informative structure prior is defined as

$$P(S|\xi^+) \triangleq P(S|D^+_{N'}, \xi^-) \propto P(S|\xi^-) \underbrace{P(Y_{N'}|S, X_{N'}, \xi^-)}_{\text{(conditional) marginal model likelihood}} \tag{10.30}$$

It shows that the informative structure prior is the (conditional) marginal model likelihood or the evidence for this structure based on the prior sample. For the application of this approach for the LR model with an additional hyperlayer on an analog of N', see [45], p.270.

After the prior sample approach, we consider now the analog of the CDM-T method for directly inducing the informative structure prior from $p(G|\xi^+)$ with a transformation $\mathcal{T} : G \to S^{\text{MLP}}$ as $p(G : \mathcal{T}^{-1}(S^{\text{MLP}})|\xi^+)$. First, consider the case of LR models without interaction terms, then the transformation $\mathcal{T}_{MB\to LR}$: $\text{MB}(G, Y) \to LR$ induces an informative structure prior for the LR structure space as

$$p(G : G \sim \mathcal{T}^{-1}_{MB\to LR}(S^{LR})|\xi^+), \tag{10.31}$$

which defines the probability of the LR structure as the probability of the Markov blanket for the output variable Y containing exactly the inputs of the LR model. However, because the independence structure of the BN depends on the discretization of the variables (see Section 3.1.4), the possible smoothing into continuous variables described in Section 10.2.3.4 weakens the validity of this approach, which can be represented with an additional uncertainty factor. In Section 9.5.2, we show a transformation from the structure of binary BNs (from MBGs in fact) onto LR structures with interaction terms as parametrically sufficient to represent the embodied conditional in an MBG. Let us denote this

transformation with $\mathcal{T}_{MBG \to LR}$, which induces an informative structure prior for the LR structure space as

$$p(G : G \sim \mathcal{T}^{-1}_{MBG \to LR}(S^{LR})|\xi^+). \tag{10.32}$$

It can be used again as a basis for an analytic approximation as in Section 10.2.3.5. However this is not minimal, only a small fraction of the LR models is used, the smoothing weakens its validity and it is not enough for the MLP model class, which has potentially multiple hidden layers (not just directly the inputs and their interactions).

Finally, we discuss a third option for defining a probabilistic link between the structures of domain models and conditional models through common high-level properties. We propose to use conditional features $F_1(G), \ldots, F_k(G)$ related to the properties of an MLP structure $F'_1(MLP), \ldots, F'_k(MLP)$ (such as the later discussed Markov blanket set and Markov blanket memberhsip features). This allows the definition of a feature prior for the MLP structures by defining $p(S^{\text{MLP}}|F_1(G), \ldots, F_k(G))$ as

$$p(S^{\text{MLP}}|F_1(G), \ldots, F_k(G)) \propto p(S^{\text{MLP}}|\xi^-) \prod_i p(F'_i(S^{\text{MLP}})|F_1(G), \ldots, F_k(G)),$$

and the derivation of an informative structure prior $p(S^{\text{MLP}}|\xi^+)$ as the marginal

$$p(S^{\text{MLP}}|\xi^+) = \sum_{\substack{F_1(G),\ldots,F_k(G) \\ F'_1(S),\ldots,F'_k(S)}} p(S^{\text{MLP}}|F_1(G), \ldots, F_k(G)) p(F_1(G), \ldots, F_k(G)|\xi^+). \tag{10.33}$$

To apply Eq. 10.33, we need the following: (1) high-level features with related interpretation in the BN and MLP classes (e.g., the set of relevant variables for a target variable), (2) the joint distribution of the BN feaures (e.g., the posterior of the MB(Y) feature), and (3) the probabilistic formalization of the relation between the BN and MLP features in Eq. 10.33 (e.g., Th. 7.1.2). Note that the probabilistic link between the features has to bridge both the different model spaces and the possibly different subdomains (in our case the literature vs. clinical subdomains).

In the thesis we investigated the following conditional features in this context: the Markov Blanket feature MB(Y, G), the size of the Markov Blanket (i.e., the number of inputs $|\,\text{MB}(Y, G)|$), the number of edges between the children of Y in the BAN converted MBG $|\text{BANEdges}(Y, G)|$ denoted as defined in Section 9.5.2 and the number of free parameters $|\underline{\theta}_{\text{MBG}(Y,G)}|$. Additional features were the misclassification rate $MR(Y, G, D_N)$ and the AUC measure $AUC(Y, G, D_N)$ to support manual interpretation and exploration. The next subsection presents results about their joint distribution in the OC domain, which allows a manual exploration to evaluate or to construct and refine an MLP structure.

Combining this method with the CDM-T transformation for deriving informative parameter prior, we created a complete probabilistic link between the parameter and structure spaces of Bayesian networks as domain models and of conditional models (i.e., we defined a full informative prior for the conditional models as the marginal of $p(\omega, S^{\mathrm{MLP}}|\theta, G)p(\theta, G|\xi^{+})$).

10.4 Joint probabilities of conditional features

The earlier section discussed a method to induce informative structure priors through high-level common or related features of model spaces using the joint distribution of the feature in the *donor* model space and a probabilistic link between them. However the conditional features of a domain model has a more basic usage simply to support the manual classifier construction process in the phases of data exploration and model construction, evaluation, and refinement. In fact, this usage logically precedes the more advanced, automated application with the probabilistic link. This use of the conditional features based on a domain model in the Bayesian framework is similar to the univariate and confounder analysis of the data set before the LR and MLP analysis to clarify the prior background knowledge relevant for the LR analysis, such as input selection, interaction term construction and transformations. The estimation and search method reported in Section 10.1 allows the offline construction of a knowledge base with the high-scoring MBGs, which can be used for subsequent complex, first-order like queries incorporating textual domain knowledge as well, such as discussed in Section 5.2.

In this section, we will present results about conditional features for the Pathology variable in ovarian cancer using the fourteen variables selected in the standard LR analysis in Section 10.5. The ordering-based MCMC algorithm from Section 8.5.1 is used with the clinical data set, without restrictions on the structure and on the orderings and using the standard settings.

Fig. 10.1 reports the estimated distribution of the AUC performance measure based on the fourteen classification variables and the clinical data set using the selected MBGs in the set $\hat{S}^{*}_{K}$ with their estimated values $\hat{p}(\mathrm{mbg})$. The estimated mean and variance for the AUC variable are 0.9386 and 0.0242 (for the misclassification rate (MR) are 0.0756 and 0.0093). The estimated distribution of the ratio of the number of parameters and the number of inputs are reported as well (the estimated mean and variance are respectively 21.6916 and 5.1318). The histogram of the number of parameters and inputs for the MBGs in the set $\hat{S}^{*}_{K}$ are reported in Fig. 10.2 (the estimated mean and variance are 283.9048 and 70.7626, and 12.4651 and 0.6827 respectively). The estimated mean of the AUC variable conditioned on this ratio and its estimated conditional distribution is shown in Fig. 10.3.

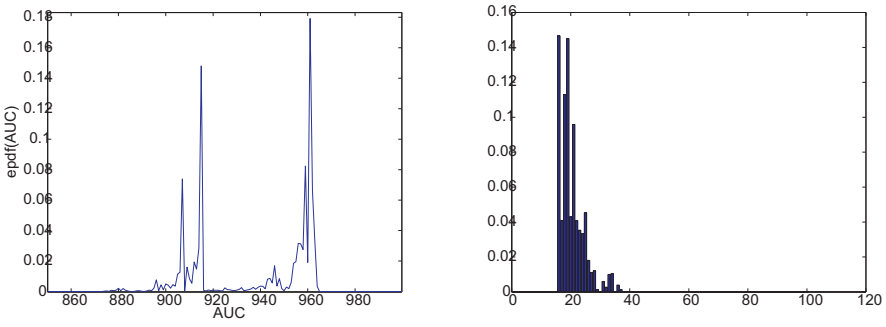


Figure 10.1: The estimated posterior distribution of the performance measure and the model complexity w.r.t. classification. The classification performance is measured by the AUC value, the model complexity is defined as the ratio of the number of parameters and the number of inputs. The models are the MBGs over the fourteen classification variables. The posterior corresponds to the IOTA-1.2 data set, which was approximated using the 10^4 MAP MBGs generated by Alg. 1. Interestingly, the domain model based posterior induces multiple, distinct modes, specifically the set of models with worse classification performance than $AUC = 0.92$ has a considerable posterior sum.

10.5 The frequentist LR modeling

After the discussion of methods to derive informative parameter and structure priors for BN classifiers in Sections 10.2 and 10.3, and before the experimental evaluation in Sections 10.6 and 10.7, this section summarize the frequentist LR analysis in ovarian cancer using standard statistical packages (SPSS 14.0 and STATA 8.2). The purpose is threefold and related to classification using the LR model and not to its possible role in causal modeling. First, we provide a baseline for classification performance. Second, we provide a baseline for selection of features (inputs) and interaction terms. Third, we provide a baseline for the preprocessing of the data for more complex classification models. The IOTA-1.2 data set will be used restricted to three sets of variables. The first "complete" set includes the thirty-five variables used in the BN-feature analysis and automated structure prior construction. The second "elicited" set includes the eleven variables used in parameter elicitation. The third set includes fourteen variables that were selected as relevant for the classification of Pathology.

Besides the discretization described in Section 4.1.1 we evaluated the univariate discriminative power of the variables using the AUC value and we determined the optimal cut-off value empirically and introduced corresponding binary variables.We evaluated also an approximation of the univariate conditional probability $p(y|X_i)$ to ensure that its logit transformation is approximately linear and applied a logarithmic transformation on Age, ReproductiveYears, Volume, PI, RI, PSV, TAMX, and CA125.

The following interaction terms are allowed in the LR models to express their

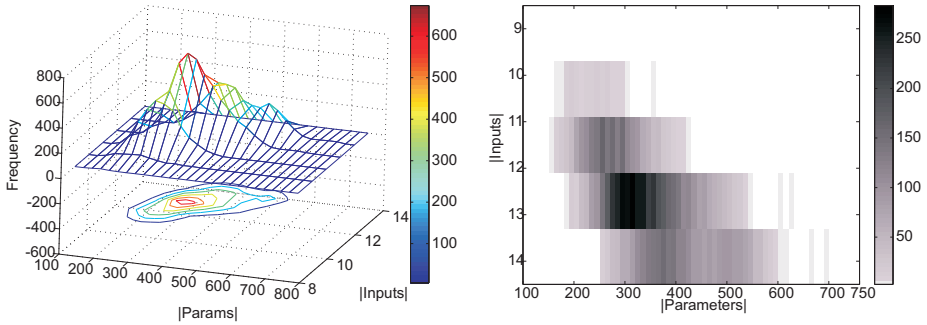


Figure 10.2: The estimated posterior of the number of parameters and the number of inputs for the MBGs based on the fourteen classification variables. The posterior diminishes outside the presented region, specifically below 10 inputs and 200 parameters (i.e., most of the variables are relevant and their effects are not independent). The posterior corresponds to the IOTA-1.2 data set, which was approximated using the 10^4 MAP MBGs generated by Alg. 1.

joint effect: PostMenoY-Meno, lnPI-cat2ColScore, lnRI-cat2ColScore, lnPSV-cat2ColScore, lnTAMX-cat2ColScore, and NrLocules-Multilocular (some of these are technical as the value of one of the pair is only conditionally interpretable). We used the default settings for the model construction in the SPSS 14.0 system (FSTEP (LR) /CRITERIA=BCON(0) LCON(0) ITERATE 50 PIN(0.05) POUT(0.1)). The selected variables in the final model with their corresponding coefficients and significance (using the Wald test) is shown in Table A.3. These variables compose the "medium" variable set, except NrofLocules, which was not included in the BN-based analysis using thirty-five variables. We also constructed a "large" variable set including the variables IncompleteSeptum and Echogenicity as well, because of their borderline significance levels (Bilateral and Shadows with similar significance were omitted finally to maintain simplicity, because of their negligible effect on classification performance).

10.6 Effect of parameter priors on classification

First, we investigated the performance of the prior Bayesian network described in Section 4.3.1. The misclassification rate is 12.0% on the data set and the mean of the Bayesian area under the ROC curve is 0.905. To get a more detailed understanding of the performance of the model we compared its predictions with those of medical experts. In a previous study six ultrasonographers (denoted by A to F in Table 10.1) have evaluated the 300 patients in the IDO data set based on the corresponding medical records and ultrasound images. Two of them were highly experienced (A and B), one moderately experienced (C) and three less experienced (D, E and F) [237]. Since these classifications were based on the original observations (such as images), in a recent experiment an

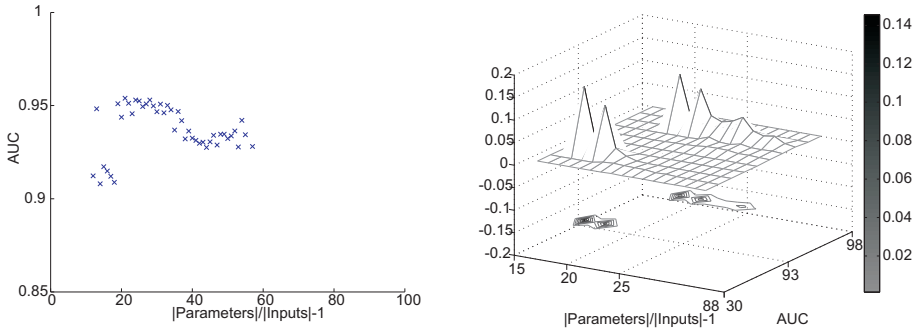


Figure 10.3: The estimated distribution of the ratio of the number of parameters and the number of inputs for the MBGs based on the fourteen classification variables. The high value of the ratio indicates multiple interactions (on average 3 variables). The posterior corresponds to the IOTA-1.2 data set, which was approximated using the 10^4 MAP MBGs generated by Alg. 1.

expert (G) has performed the classification using only the discrete values of the variables present in the prior Bayesian network. In this experiment the expert has also rated the cases as 1 = very certain benign, 2 = uncertain benign, 3 = uncertain, 4 = uncertain malignant and 5 = very certain malignant, which is an aggregate expression of the mixed nature of the adnexal mass ("fuzziness") and the probabilistic uncertainty.

Table 10.1 presents the number of previously performed examinations, the misclassification rate and the agreements with the prior model (Cohen's kappa) for the diagnosticians [30]. The correspondence with the prior model is the highest for expert A (indicated in bold) which is in line with our expectation because expert A participated in the construction of the prior model.

Table 10.1: Expert agreement with the prior domain model in discriminating benign and malignant adnexal masses [237] (number of examinations (#), misclassification rate (MR[%]), Cohen's kappa (κ)).

Expert	A	B	C	D	E	F	G	Prior-BN
#	≥4000	≥10000	≥1000	200	300	300	300	-
MR	8.3	8.3	11.0	17.7	7.7	13.3	18.0	12.0
κ	**0.713**	0.687	0.650	0.577	0.503	0.590	0.577	-

To evaluate the effect of the prior incorporation methods, we compare them in a retrospective setup with standard learning algorithms for Bayesian networks and multilayer perceptrons. For Bayesian networks, we present results for four models with a *noninformative* prior distribution (BN-Naive, BN-Fixed-Noninformative, BN-TAN and BN-General) and one with an *informative* prior

distribution (BN-Fixed-Informative). For the MLP model class, we describe one model with a *noninformative* prior (MLP-Noninformative), one with *prior samples* (MLP-Prior sample) and one with an *informative* prior (MLP-Informative) as explained in Sections 10.2.2, 10.2.3.1, and 10.2.3.3. The BN-Fixed-Informative and the BN-Fixed-Noninformative methods use the same fixed structure as the prior domain model shown in Fig. 4.1. In all the other learning methods two variables that are functions of the variable Locularity were removed, because these auxiliary variables were introduced to support knowledge elicitation.

The BN-Fixed-Noninformative, BN-Fixed-Informative and the BN-Naive methods perform only parameter learning (the BN-Naive method uses a Naive Bayes structure, a tree where all variables are direct childs of the predicted variable Pathology). The BN-TAN method searches in the space of generalized tree-augmented networks, which are extended Naive Bayesian network structures [91, 48]. Finally, the BN-General method searches the space of directed acyclic graphs (in each crossvalidation session, 10^5 random orderings of the variables are generated, for each ordering the parental sets are evaluated exhaustively up to three parents, and if necessary, by the greedy (not exhaustive) K2 algorithm [57]).

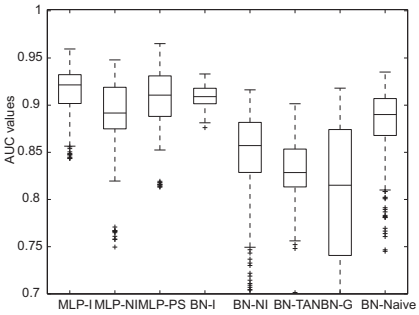
In the *prior sample* (MLP-Prior sample) method, 1,000 samples are generated from the prior Bayesian network rescaled to an effective sample size of 30 in the Bayesian inference scheme as explained in Section 10.2.3.1. In the *informative prior* (MLP-Informative) method, the *informative prior* was estimated on 5,000 multilayer perceptron parameterizations using the mixture of 3 Gaussian kernels. Each multilayer perceptron parameterization is computed from an independently drawn Bayesian network parameterization by training the multilayer perceptron on 1,000 random samples produced by the Bayesian network, as explained in Section 10.2.3.3.

For the Bayesian inference, direct sampling was used for the Bayesian network and hybrid Monte Carlo methods were used for multilayer perceptrons to draw 100 parameterizations from the a posteriori distribution, thus performing 100 inferences for the test set. This process was repeated for 100 cross-validation sessions (different partitions of the data set into test and training set). Fig. 1.7 shows the detailed effect of the prior incorporation for varying proportions of samples used in the training set.

Two characteristic points from this learning curve (the small and large sample region) are shown in Fig. 10.4 and 10.5 corresponding to the 5%–95% and 75%–25% training–test proportions.

10.7 Effect of structure priors on classification

To evaluate the value of the text-based prior distributions for Bayesian network structures, we report results for the classification of ovarian tumors. (For previous results about the application of Bayesian networks and multilayer perceptrons to classify ovarian tumors, see [15].) We report the classification performance of a Bayesian network using the Area Under the Receiver Operating



	$\widehat{E}$[AUC]	$\widehat{Std}$[AUC]
MLP-I	0.915	0.026
MLP-NI	0.889	0.039
MLP-PS	0.907	0.032
BN-I	0.909	0.011
BN-NI	0.851	0.040
BN-TAN	0.809	0.037
BN-G	0.738	0.113
BN-Naive	0.879	0.039

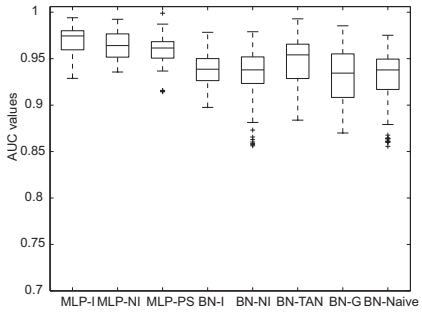
Figure 10.4: The effect of informative parameter prior on classification performance in case of small sample size. The a posteriori distribution of the area under the ROC curve at the 5%–95% training–test proportion for the MLP-Informative (MLP-I), MLP-Noninformative (MLP-NI), MLP-Prior sample (MLP-PS), BN-Fixed-Informative (BN-I), BN-Fixed-Noninformative (BN-NI), BN-TAN, BN-General (BN-G) and BN-Naive models. Besides the significantly better mean performance, the informative prior decreased the variance as well.

Curve (AUC). Since we work in the Bayesian framework, we have a posterior distribution $P(G|D_N)$ over the network structures and a conditional posterior $P(\underline{\Theta}|G, D_N)$ over its parameters, resulting in a posterior distribution of the AUC. We report the mean of this AUC using an informative text-based prior, the expert prior, or a noninformative uniform prior over the structure space:

$$\mathrm{E}[\mathrm{AUC}_{G,\underline{\Theta}}(D^{\mathrm{te}})|D^{\mathrm{tr}}] = \sum_{G} P(G|D^{tr}) \int_{\underline{\Theta}} \mathrm{AUC}_{G,\underline{\theta}}(D^{\mathrm{te}})\, \mathrm{d}P(\underline{\theta}|G, D^{\mathrm{tr}}).$$

where D^{tr} and D^{te} denote the training and test data. Because we want to focus on the usage of textual prior knowledge for learning Bayesian network *structures*, we used always the noninformative Bayesian Dirichlet prior BD_{eu} for the parameters [131].

We approximate the summation over the network structures with a Monte-Carlo approximation using 200 networks with a high posterior probability. We evaluate 200 randomly drawn orderings for the variables. Using a set of complete and discrete samples, we learn a Bayesian network structure by maximizing the R^{BD}_{Data} score for each variable for the given ordering [57]. For each fixed ordering, the parents are selected using an exhaustive search up to three parents. If this exhaustive search finds three parents, the greedy (not exhaustive) K2 algorithm continues the search [57]. The probabilities for the Bayesian network substructures are computed using both the training data and the edge probabilities according to Equation 3.19. The edge probabilities derived from the text were scaled with a ν value that results in prior networks with 3 parents for each node on average.



	$\widehat{E}$[AUC]	$\widehat{Std}$[AUC]
MLP-I	0.969	0.015
MLP-NI	0.964	0.015
MLP-PS	0.959	0.016
BN-I	0.937	0.021
BN-NI	0.933	0.024
BN-TAN	0.950	0.026
BN-G	0.947	0.027
BN-Naive	0.930	0.025

Figure 10.5: The a posteriori distribution of the area under the ROC curve at the 75%–25% training–test proportion for MLP-Informative (MLP-I), MLP-Noninformative (MLP-NI), MLP-Prior sample (MLP-PS), BN-Fixed-Informative (BN-I), BN-Fixed-Noninformative (BN-NI), BN-TAN, BN-General (BN-G) and BN-Naïve models. The classification performance is not improved significantly by the informative prior at this sample size (i.e., its initial advantage has diminished smoothly without delaying the long term improvement of the performance.

For these learned structures, the parameters are set to the maximum a posteriori value using the noninformative BD_{eu} prior for the parameters and the training set D^{tr}. Predictions for the test samples are generated using the probability propagation in tree of cliques (PPTC) algorithm and these predictions are used to compute the AUC value on the test set D^{te}. The AUC values reported in Figure 10.6 are the averages over 300 cross-validation sessions with random training–test partitioning of the data set D_N into $(D^{\mathrm{tr}}, D^{\mathrm{te}})$. The x axis indicates the number of samples in the training set, ranging up to 150 samples (out of a total of 604), the y axis contains the AUC averages for that specific training–test proportion.

The upper part of Figure 10.6 reports the learning curves for the co-occurrence- and corelevance-based text priors ($R_{\mathrm{COOC}}^{\mathrm{AND,ML}^R}$ and $R_{\mathrm{COREL}}^{\mathrm{AND,ML}^H}$), together with the kernel similarity prior R_{ASIM}, the expert prior R_{Expert}, and no prior, all scaled by $\nu(3)$. The bottom part shows the effect of scaling the best performing prior based on the kernel similarity score R_{ASIM} by $\nu(0.1)$, $\nu(0.5)$, $\nu(1)$, $\nu(2)$, and $\nu(3)$. The noninformative case is again reported for comparison.

10.8 Effect of model averaging on classification

In Section 10.6 and 10.7, we used fixed structures or MAP structures in prediction, and not Bayesian model averaging. In this section we report the effect of BMA on classification in ovarian cancer. In case of Naive-BNs we use the exact averaging [66]. We report the performance of the MAP MBG, which is already an aggregation of models and the approximation of the BMA for BNs

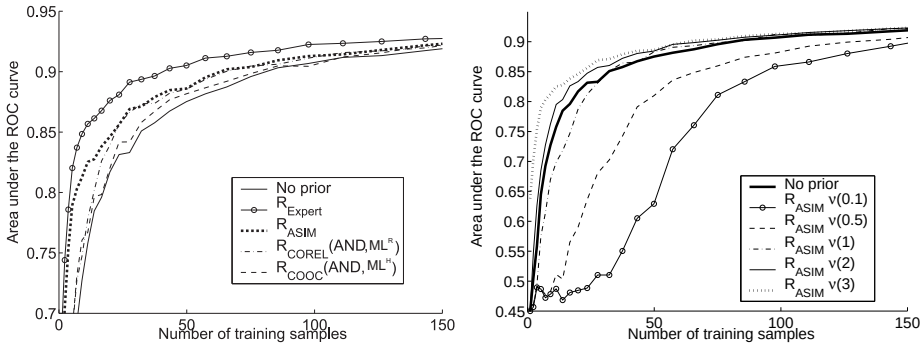


Figure 10.6: (Left) The AUC performance for BNs using different text-based priors ($R^{\mathrm{AND,ML}^R}_{\mathrm{COOC}}$, $R^{\mathrm{AND,ML}^H}_{\mathrm{COREL}}$), and R_{ASIM}, the expert prior R_{Expert}, and no prior. The priors are scaled to an average of 3 parents per variable. (Right) The AUC performance for BN using the R_{ASIM} prior for different ν scalings (average number of parents is scaled to 0.1, 0.5, 1, 2, and 3) together with the performance without any prior.

using MBGs. Note that earlier the parameters were exactly averaged in BN classifiers, which will be also used here for BN classifiers, though in conditional applications the mean parameters are not optimal (see Section 9.4.1). We report results at three levels of model complexity: (1) for the eleven variables used in parameter elicitation in Section 4.3.2 (the "small" set), (2) for the fourteen variables selected in conditional LR modeling in Section 10.5 (the "medium" set), and (3) for all the thirty-five variables used in general modeling (the "large" set).

As a reference, we include a non-informative LR (MLP) model, a non-informative MLP model with two hidden units and the TAN method as the best performing BNC. In the case of the medium set of variables, we use the equivalent set of input variables possibly with continuous variables and in the case of the thirty-five variables, the larger variable set as described in Section 10.5. In both cases, the MLP models include 2 hidden units with hyperbolic tangent transfer function before the output unit with logistic function. Note that the MLP models have fixed structure, conditionally optimized parameters and optionally use the original continuous values.

Additionally we report the effect of parameter priors from Section 4.3.2. To simplify the presentation the exact BMA is not reported for the Naive BN, because its effect was negligible and only the virtual sample size 150 is used in reporting the effect of parameter prior. The effect of BMA with non-informative and informative parameter priors for BN classifiers using the small, medium, and large set of variables are shown in Fig. 10.7, Fig. 10.8, and Fig. 10.9.

The standard deviations are between 0.01-0.04 for training proportions 0.1-0.9 in the cross-validations. The results in Fig. 10.7, Fig. 10.8, and Fig. 10.9 correspond to growing number of variables (i.e., increasing model complexity) and their left and right columns correspond to the non-informative and infor-

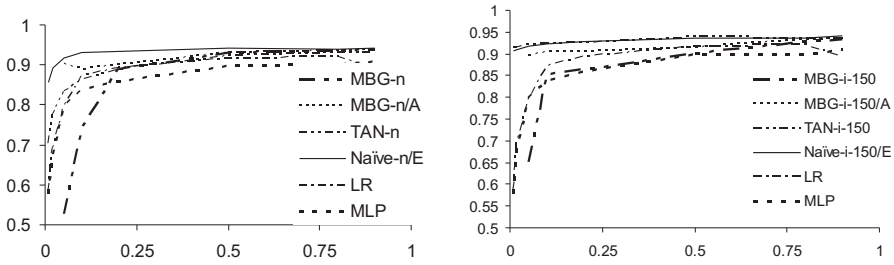


Figure 10.7: The effect of BMA using the learning curves of the AUC perfromance measure. The eleven variables from the elicited model were used. Naive, TAN, BN, MBG and LR, MLP denotes the appropriate structure learning method and model classes. The n/i postfix denotes the noninformative/informative case followed by the applied virtual sample size. The E/A postfix denotes the exact or approximate Bayesian model averaging.

mative cases in a tabular form. Their comparison again illustrates the effect of the complexity of the model class on the rate of its learning curve (model complexity is jointly defined by the growing number of variables and by the applied models, such as Naive-BN, lR, TAN, MLP, MBG). The comparison also shows the advantageous effect of the informative priors, particularly the effect of the prior w.r.t. the different number of variables (see the). It also illustrates the advantageous effect of Bayesian model averaging, because the MBG/A option (which is equivalent to the Bayesian averaging of BNs) has significantly better performance than its MAP approximation (MBG) in large regions. However, for the 90% training proportion, the differences of the MR and the AUC performances are not significant for the LR, MLP, MBG and TAN models using paired t-test ($0.05 < p$).

10.9 Discussion

As more and more domain knowledge becomes available beside statistical data, machine learning increasingly needs methods that integrate domain knowledge [8]. The first step in this prior incorporation is the acquisition, formalization, evaluation, and fusion of the heterogeneous a priori information. The second step is the incorporation of the formalized prior knowledge in a task specific model. The prior incorporation methods make it possible to integrate probabilistic domain knowledge into black-box models. This hybrid use of knowledge-oriented and data-driven methods can be particularly advantageous in constructing a classifier when the size of the sample is small or medium and a large amount of domain knowledge is available.

The efficiency and simplicity of the *prior sample* method makes it attractive. This method has the advantage that the generation of the prior data set from a Bayesian network is straightforward, computationally simple and that it can be

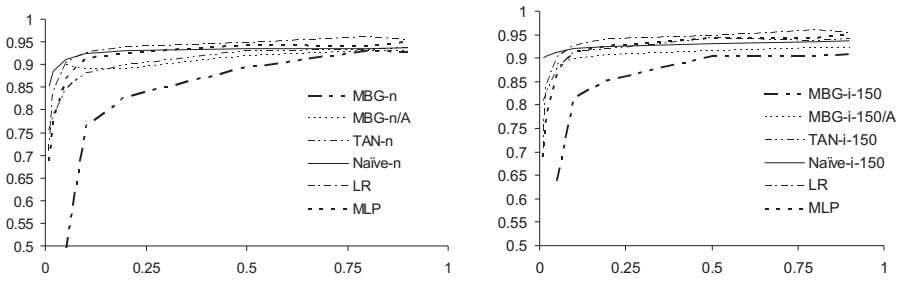


Figure 10.8: The effect of BMA using the learning curves of the AUC perfromance measure. The fourteen variables were used selected in the conditional LR modeling. Naive, TAN, BN, MBG and LR, MLP denotes the appropriate structure learning method and model classes. The n/i postfix denotes the noninformative/informative case followed by the applied virtual sample size. The /E/A postfix denotes the exact or approximate Bayesian model averaging.

applied to nonparametric models, such as support vector machines. Prior incorporation significantly enhances the performance for small sample sizes. More surprisingly, we also observed slight improvements when the size of the real data exceeds the effective prior sample size by a factor two to four (the 0.2-0.4 region on Fig. 1.7). The rescaled large *prior sample* block — which may contain infrequent samples as vital hints from the prior Bayesian network — is probably the source of this improvement.

Similarly, the *informative prior* method immediately achieves the same performance as the prior Bayesian network and gives better performance than the noninformative multilayer perceptron for any amount of real data available in the experiment. It means that the estimated prior is efficient in the small sample region and not restrictive in the large sample region; it has a balanced, lasting positive impact. Beside this statistical (machine learning) aspect, the related computational aspect of the inference similarly provides certain advantages. The high computational complexity of deriving the informative prior (when compared to simply generating blocks of samples in the *prior sample* method) is compensated by a lower complexity in the inference. Note that the additional blocks of prior samples slow down the likelihood and gradient computations in the Bayesian inference in MLP models. This precomputation property of the *informative prior* method is similarly relevant in the Bayesian network context. Remember that, for Bayesian networks (even using only a single structure with a point parameterization), the computational complexity of the exact inference or its approximation is NP hard [64]. Consequently, the transformation itself — the precomputation of the collapse of a complex general Bayesian model into a task specific, simpler classification model — can be advantageous, if the computational efficiency of the Bayesian inference is important — for example if a regular classification task in medical decision support involves only a relatively small, but fixed subpart of a complex a priori Bayesian network covering the

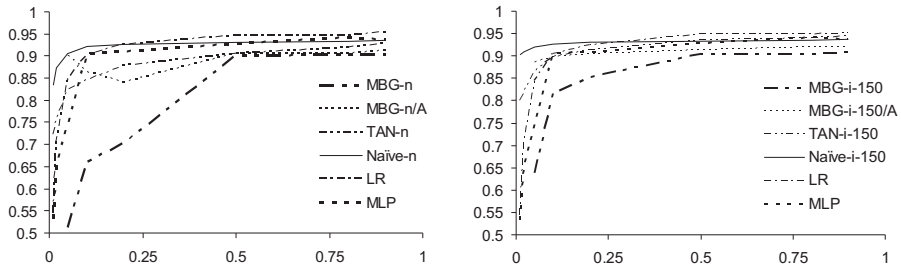


Figure 10.9: The effect of BMA using the learning curves of the AUC perfromance measure. All the thirty-five variables were used. Naive, TAN, BN, MBG and LR, MLP denotes the appropriate structure learning method and model classes. The n/i postfix denotes the noninformative/informative case followed by the applied virtual sample size. The /E/A postfix denotes the exact or approximate Bayesian model averaging.

overall domain.

Multilayer perceptrons perform generally better than Bayesian networks, but part of this difference comes from the refinement of the a priori discretization. Indeed, the performance of a corresponding noninformative multilayer perceptron with the original nominal inputs is similarly worse (for 70%–30% train–test ratio, the mean AUC is 0.945 and the misclassification rate is 9%).

Fig. 10.4 and 10.5 give a more detailed characterization of the different models, showing that in the small sample region (5%) the prior based methods have the best performance (BN-Fixed-Informative, MLP-Informative and MLP-Prior sample). From the *tabula rasa* methods (MLP-Noninformative, BN-General, BN-TAN and BN-Naive) the BN-Naive has the best performance. Another effect of the incorporation of the prior domain knowledge is that the performance of these models has smaller variance. In the large sample region (75%) the MLP methods have the best performance, specifically the MLP-Informative is still slightly better than the MLP-Noninformative.

In the classification task, the automatically constructed text-based prior for Bayesian network structures is beneficial in the small sample range, while it is not restrictive and vanishes in the middle and large sample range. That is, it provides advantages comparable to those of a manually constructed expert prior, which achieved classification performance as that of expert diagnosticians (see Table 10.1).

Chapter 11

Conclusion

The central motif of the thesis was the formalization, comparison and fusion of multiple models corresponding to different levels of human knowledge. The main goal was to investigate a probabilistic inference in this context, particularly the inductive inference incorporating prior domain knowledge, one of the most challenging problems in knowledge representation and machine learning. The investigated models included Bayesian networks as domain models based on personal belief, literature or publication domain models based on the statistical analysis of voluminous free-text corpus, classification subtheories as Markov Blanket subgraphs, and classification models used in clinical practice. These models embrace the causal-probabilistic and the conditional-domain aspects, which are heavily investigated in causal and statistical research. In the investigation, the Bayesian framework was adopted, especially because of the incorporation of prior knowledge. These allow the following views of the thesis. From the point of view of knowledge representation, how can we represent and formulate statements fusing together domain literature, expert priors and observational data. From the point of view of machine learning, how can we perform inductive probabilistic inference about knowledge-intensive, possibly causal statements.

11.1 Contributions of this dissertation

The contributions of the thesis are the following.

1. *Collection and formalization of the prior information related to the preoperative classification of ovarian masses.* We constructed parameterized Bayesian networks of the clinical aspect of the ovarian cancer domain [28]. We elicited logical, qualitative, probabilistic and causal characterizations of pairwise relations of the domain variables [16]. The models were also annotated by domain experts with free-text and links to the electronic literature. We constructed various collections of electronic domain literature

resources and compiled domain vocabularies [22]. Whereas the construction of a Bayesian network, even with attached informal knowledge is more and more common, our knowledge base is unique, because of its diversity. We discussed the consequences of Bayesianism and the increasing amount of electronic prior knowledge, statistical data, and computational resources for knowledge engineering of Bayesian networks and identified the challenges for this "Bayesian" knowledge engineering (see Section 5.1). Using the annotated Bayesian network knowledge base we developed a corresponding model-based information retrieval language and various profiling (explanation) methods to support information retrieval and knowledge engineering for complex Bayesian networks [23]. This method can be conceived as a work on ontology based information retrieval, which is unique in its use of keyword profiles and concepts related to Bayesian networks, and in its integration of knowledge engineering and information retrieval (i.e., that the applied and the expanded probabilistic "ontology" are the same).

2. *Knowledge extraction and discovery with Bayesian networks from domain literature.* We proposed various causal probabilistic models of biomedical publications and formalized their assumptions (see Def. 6.4.1). Furthermore, we introduced a related text mining method with Bayesian networks to support model-based knowledge extraction and exploration from free-text literature. We evaluated the inferred clinical models of ovarian cancer against gold standard priors and Bayesian networks based on clinical data at multiple levels of complexity such as pairwise, feature and full model level. The results showed that the proposed inferences based on literature are comparable to the gold standard priors w.r.t. the inferences based on clinical data (i.e., the clinical data ensures a reference, which is closer to each of them), and the proposed Bayesian network based text mining method gives more cautious (sparse) results than its currently prevailing bottom-up counterparts [19, 20, 13, 16, 114, 24, 25, 26]. In short, we demonstrated that the Bayesian network based Bayesian statistical analysis of the domain literature offers a feasible, complementary, model-based analysis besides non-model based co-occurrence and linguistic approaches (e.g., see [208, 135]).

3. *Learning Bayesian network features from heterogeneous sources.* We proposed a new structural Bayesian network feature called Markov Blanket (sub)Graph (MBG), a.k.a. classification subgraph, feature subgraph. We proved that it is a necessary and sufficient feature for conditional modeling (see Proposition 7.2.1 and Proposition 7.2.2). We gave an exact generalization of the feature subset selection (FSS) problem — which corresponds to the Markov Blanket set (MB) feature — by formulating its equivalent at the level of the MBG feature, as the feature (sub)Graph Selection (FGS) problem (see Def. 7.2.3). We discussed the probabilistic and causal interpretations of the MBG feature, and its relation to logistic regression (see Lemma 9.5.1 and Lemma 9.5.2). We discussed the application of the MBG

feature in ordering-based Monte Carlo methods (see Proposition 7.5.2 and Th. 7.2.1). The proposed Bayesian analysis of compact, but complex features (i.e., well-defined subtheories) led to the problem of joint estimation and selection with new statistical and computational challenges. Then we formalized the *Most Probable Features* problem (MPFs) (Def. 7.6.1) and we analyzed the effect of feature cardinality on estimating and selecting the most probable features. Here we proved that in the probably approximately correct (PAC) framework the sample complexity and the expectation of the average error of the empirically optimal feature values are similarly related to the logarithm of the cardinality of the feature set (Th. 7.6.1). We devised a new ordering-based estimation and search method for Markov Blanket sets and Markov Blanket subgraphs using the concept of truncated MBG space (see Def. 7.7.1, Lemma 7.7.2, and Alg. 1). With it we demonstrated that the normative Bayesian solution to the FSS problem, the FGS problem, and in general to the MPFs problem with the MB and MBG features are viable with recent computational resources in typical biomedical problems. In the Bayesian context we showed the relative independence of the MBG and MB feature levels (i.e., their autonomy) by evaluating them against complete models and simple features [25, 21]. With its options about the training proportion of the input data set (and averaging schemes), the Alg. 1 implements a "*Bayesian, four-level, sequential relevance analysis*" at the levels of Markov Blanket Memberships, Markov Blanket sets, Markov Blanket graphs, and complete Bayesian networks. The proposed Bayesian estimation and search method for the MBG feature can be seen as the integration of three research areas. These areas are the inference of posteriors of complex model properties (e.g., see [98]), the search for high-scoring complex model properties (e.g., see [203]), and the construction of offline, probabilistic knowledge bases (e.g., see [40, 181]). We formalized alternatives for Bayesian fusion of information in learning Bayesian network at the level of data sets, features, and models (see Th. 8.1.1 and Section 8.1.2). We proposed a direct fusion method for fusing clinical and literature data with expert priors (see Th. 8.1.1) and a two-step methodology based on the level of features, as an interpretable and computationally efficient way to fuse clinical and literature data with expert priors (i.e., to fuse heterogeneous information sources) [24, 26]. Within the "Bayesian knowledge engineering" framework, we realized that to ensure complex, knowledge rich queries is an equally important way of the incorporation of prior knowledge in statistical data analysis (besides influencing the posteriors). For such fusion of voluminous (electronically available) factual knowledge and complex, uncertain knowledge from data analysis we formulated the idea of Probabilistic Annotated Bayesian Network Knowledge Base, which allows the introduction and strict definition of a probabilistic truth value (i.e., probability) of first-order sentences (see Def. 5.2 and [25, 21, 189]). This hybrid approach embedding complex distributions specified by Bayesian networks into logical knowledge bases extended the research of probabilistic first-

order logic [121, 192, 145, 151, 187, 76]. We implemented a corresponding inference engine, which offers full Bayesian inference over complex statements including semantic conditions and the properties of the underlying model using wide range of expert priors and multiple, possibly temporal statistical data sets and electronic literature corpora [25, 26, 21, 189].

4. *Bayesian classification with informative priors.* We evaluated the classification performance of Bayesian classifiers including logistic regression, multilayer perceptrons and various Bayesian network models [28, 27, 11]. For Bayesian networks we analyzed the induced joint posterior over various structural features and performance measures [21]. We proposed a transformation based on "conditional distance minimization" for deriving structural and parametric priors for black-box classifiers (see Def.10.2.3.3 and [10]). We specialized it to derive informative priors for generalized logistic regression models, such as multilayer perceptrons (see Alg.3 and [18, 11]). We showed that it is an asymptotic version of the "prior sample" and "prior over sample" based methods (see Alg.2). This method avoids the problem of selecting prior virtual sample size of the other methods, because it allows the direct transformation of the priors (i.e., the transformation of the "prior over incomplete samples"). We evaluated the performance of the "prior sample", the "prior over sample", and the "conditional distance minimization" methods, which showed that classification performance is significantly improved in the small and medium sample size region [18, 15, 12, 14]. We formalized also a Bayesian decision problem including rejection besides classification and evaluated the effect of the rejection [17].

11.2 The developed software platform

The corresponding software platform to this research contains many modules for text-processing, knowledge engineering, learning Bayesian networks and multilayer perceptrons, and performing Bayesian inferences about Bayesian network features and about general ABN-KB statetements. The source of the implemented methods exceeds 10^5 lines of code written in C++ and MATLAB. It has a graphical user interface in the MS-Windows environment. Its command line version runs in a parallel computing grid environment. Its development started at the Technical University of Budapest in 1997 for performing inference and learning Bayesian networks. The main development happened at the Department of Electrical Engineering at the Katholieke Universiteit Leuven under the name of Software Environment for Bayesian and Neural Networks (SEBANN). It was expanded with knowledge engineering modules and methods for multilayer perceptrons with the help of Geert Fannes. It was used to compute the results in the papers written between 1999 and 2002. In 2002 on the one hand Geert Fannes started the development of a LINUX based version of the system and its expansion to continuous variables used in his doctoral thesis [85]. On the

other hand the development of the System for Probabilistic Annotated Networks (SPAN) had started to support the sequential analysis and the inference over complex Bayesian network features including the MB and the MBG features. The modules for the text-mining and for the *Bayesian, four-level, sequential relevance analysis* are currently integrated into a new software platform, which supports the design and analysis of "genomic" experiments including clinical observations, genotypic information about the patient, and genomic profiles of the related tissue.

11.3 Applicability in the postgenomic era

The original goal of the research was to develop new methods for knowledge rich statistical data analysis in biomedicine, specifically to develop methods, which can incorporate prior background knowledge. Such methods in turn support the development of new decision support models used in clinical practice. Due to its roots in both prior knowledge and statistical data, this research covered a broad spectrum of problems from biological knowledge discovery to clinical decision support. Another factor influencing the developed methodology was the timing: 1999-2002, the spread of high-throughput measurement methods and the "birth" of bioinformatics. The availability of massive amount of information about the underlying genomical and proteomical level had a major impact on the biomedical research. The availability of the genotypic information about the patient such as a profile of her/his single nucleotide polymorphisms (SNPs), and the availability of the genomic profiles of the related tissue, such as from comparative genome hybridization (CGH) analysis or gene expression (GE) analysis together with clinical information led to the concept of *personalized medicine* (i.e., to personalized prevention, diagnostics, and treatment).

This new situation motivated many elements in the developed methodology and in the corresponding software platform. We used in fact the clinical aspect of the ovarian cancer domain as the better explored aspect to develop and evaluate methods applicable in other biomedical domains with corresponding genomic data. Thus we enumerate in Section 11.3.3 some of the future applicability of the developed methods in this new context of biomedicine. Technically it does not change our goal of developing methods to engineer decision support models for clinical use incorporating prior information, but we assume that some of our input variables are "genomic" (e.g., denoting genomic information from SNP, CGH or GE analysis).

11.3.1 Main constructs and methods

First we summarize the relevant constructs, methods, and methodologies developed in thesis.

ABN The probabilistic annotated Bayesian network knowledge base (see Def. 5.2).

TM Textmining with Bayesian networks (see Def. 6.4.1 and Section 8.4.2).

MBG The Markov Blanket graphs for exploration and classification (see Def. 7.2.1, Proposition 7.2.2, Lemma 9.5.1, and Lemma 9.5.2).

MPFs Inferring the most probable Markov Blankets and Markov Blanket graphs (see Def. 7.2.3, Def. 7.6.1, Lemma 7.7.1, and Th. 7.6.1).

B-4S The "Bayesian, four-level, sequential relevance analysis" at the levels of Markov Blanket Memberships, Markov Blanket sets, Markov Blanket graphs, and complete Bayesian networks (Th. 7.2.1, Def. 7.7.1, and Alg. 1).

iMLP Informative parameter priors for multilayer perceptrons from Bayesian networks (see Section 10.2.3.3).

11.3.2 Main types of the prior knowledge

The following priors were constructed for these methods. Note that the purpose of these priors was also comparison (e.g., to evaluate methods using a reference). Furthermore, we conceive the formulation of knowledge-rich queries also as a kind of prior incorporation (besides the standard Bayesian interpretation of influencing the posterior). This referential and auxiliary use of priors is the consequence of an important lesson in bioinformatics that the interpretation of the results of the data analysis — i.e. the recognition of new and relevant knowledge — is a serious bottleneck. Table 11.1 shows the relation of the priors and methods (the MBG and the MPFs columns are not shown as they are included in the column of "Bayesian, four-level, sequential relevance analysis" (B-4S)).

Table 11.1: Main types of the elicited prior knowledge and their relation to constructs and methods.

prior	ABN	TM	B-4S	iMLP
Document collections	x	x	x	
Domain vocabulary	x	x	x	
Variables with annotations and references	x	x	x	
Groupings of the variables	x		x	
Partial and complete orderings of the variables	x	x	x	
Complete causal models	x		x	
Pairwise relevance	x		x	
Parameters of local probabilistic dependencies			x	x

11.3.3 From current results to proposed uses

The advantage of the *probabilistic annotated Bayesian network knowledge base* (pABN-KB) is that it retains the richness of the natural language publications, yet the uncertain raw results from statistical data analysis can be incorporated (i.e., the posterior over domain models). So this is a new kind of integration of knowledge engineering and statistics, which is essential in domains with huge

amount of factual knowledge like studies associating the clinical level and the biological level. However, there are many open issues, such as the design of the factual knowledge base and a corresponding scalable theorem prover, the user interface, the problem of the trivial sentences (e.g., tautologies with probability 1), and the design of biologically inspired schemes of queries [189].

The *Bayesian network based text-mining* (TM) method, using the document collections, domain vocabularies, and annotations of the variables, gave more cautious results than the co-occurrence methods. It could reconstruct structural prior information comparable to that of an expert. Because of its vector representation it is not sensitive to ill-formed relationships (i.e., it can be applied in a different phase than natural language processing (NLP) methods, see Section 6.2). Furthermore, because of its unique model-based foundation, Bayesian model averaging can be applied to derive posteriors about complex relationships such as Markov Blankets and Markov Blankets graphs, contrary to the typical pairwise usage of the co-occurrence and NLP methods. Due to the fact that the Bayesian network based text-mining method is multivariate, it can be a useful ingredient of domain exploration tools in genomic domains, specifically in designing genomic experiments or suggesting genomic test for a patient. This remains so, even if more and more curated databases are available, because it is model-based and multivariate, and because the document collections and the domain vocabularies used in the analysis can be selected arbitrarily.

The *feature subset selection* (FSS) problem is at the heart of the emerging personalized biomedicine, due to the multifactorial aspect of most of the diseases and the enormously big number of potential genomic variables (e.g., millions of SNPs, thousands of genes). The Feature (sub)Graph Selection problem (FGS) based on the concept of Markov Blanket subGraph is an exact generalization of the FSS problem, which helps to understand the conditional relevance of certain features, presumably common in diseases with many factors. The Most Probable Features problem over the MBG features is another kind of generalization of the FSS problem, as the typical data size in genomic experiments is not enough to ensure a feature value with dominant posterior, i.e. multiple selection is unavoidable. The developed algorithm provides a flexible, unique tool to solve the MPFs problem in case of Markov Blanket graphs and Markov Blanket sets, incorporating logical and numeric priors for the orderings, structural aspects, and parameters. It provides estimates for standard features, such as Markov Blanket Memberhsip and compelled edge relations as well. It allows a "Bayesian, four-level relevance analysis" at the levels of Markov Blanket Memberships, Markov Blanket sets, Markov Blanket graphs, and complete Bayesian networks. Additionally, it provides a sequential analysis as well, which proved to be an essential tool to explore the sufficiency of the data besides confidence measures. Because of the high computational complexity, the sequential analysis was greatly enhanced by the availability of priors for the ordering and grouping of the variables, which are optional prior constraints for the developed algorithm. Note that the multiple level analysis with growing model complexity together with the sequential option allows a broad vision to understand the power of the prior and the data. *In short, the proposed "Bayesian, four-level, sequential*

relevance analysis" capable for incorporating diverse priors is an up-to-date response for the challenges of the personalized medicine, and generally applicable in non-medical domains as well to facilitate knowledge-rich data analysis.

Finally the induced *informative parameter priors for MLPs* were efficient in the small sample region, although their effect on classification performance was not significant in the large sample region (for larger sample size than 200). This observation is consistent with posterior analysis of the hyperparameter expressing confidence in the prior estimates as a prior virtual sample size, which also peaked around 200 samples. Furthermore the effect of prior was similarly fading away above this limit in various prequential analysis as well. However, this small sample size region remains an important challenge in the postgenomic era, because of the subpopulations in personalized medicine and the more and more threatening problem of rapidly adapting (i.e., changing) diseases.

11.4 Challenges

First we summarize general challenges present in the fields of knowledge representation, intelligent data analysis, machine learning or data and text mining:

1. *Data heterogeneity*: the growing importance of interventional and structured data, besides heterogeneous observational, unstructured data sets,

2. *Electronic semi-formal priors*: the role of formal and semi-formal knowledge bases will increase, besides expert knowledge and free-text repositories,

3. *Complex models*: the learning of interpretable, hierarchially decomposed, relational and object-oriented models will be more and more important, besides the classical function and density estimation view of induction,

4. *Inference of complex properties of models*: the normative estimation and search of complex, semantic features formalized using a rich language will be more and more important, besides MAP model identification and learning simple properties.

From this perspective, we discuss possible future research related to the contributions of the thesis.

The constructed priors for ovarian cancer, including parametric and many cross-comparable structural elements could serve as a benchmark tool for the evaluation of other prior extraction, incorporation, and fusion methods. The annotated Bayesian network based information retrieval language can serve as a guide for the future development of biomedical information retrieval search engines and interfaces, because it illustrates the type of complex queries from the point of view of Bayesian knowledge engineering and Bayesian learning.

Text mining with Bayesian networks includes two open issues. On the one hand the current "publication" models for information extraction and knowledge discovery have to be expanded to handle explicitly the neutral omission and

the negation; and to cope with the references, the full-text, and the sequential-temporal nature of the publications. On the other hand, the purpose of learning generative publication models (i.e., information extraction and knowledge discovery) will probably be expanded to the analysis of the cognitive aspects of scientific explanations and understanding, and to the analysis of the social and the economic aspects of the collective behaviour of the research communities (e.g., for designing scientific policies and research programmes). It means that decision networks have to be learned as "publication" models, which means the incorporation of actions and utilities (e.g., scientific credits and financial costs) into the Bayesian network models of publication.

As for the estimation and search method for MBG feature, the task is more mundane. First of all note that our assumption of a single target variable was technical, and all the concepts and methods can be easily extended to a group of target variables (e.g., instead of Pathology we can define an aggregated variable representing the subtypes of the tumor and the follow-up of the patient). Second, the same MCMC scheme can be easily extended to continuous variables, because there exists a closed formula for the conditional posterior of DAG structures in case of Gaussian local models as for the discrete case with multinomial local models used in the thesis [106]. Third, we have to generalise the method to cope with incomplete data, using imputation, "Expectation-Maximization", or embedding this problem into the MCMC scheme (e.g., using an additional Gibbs sampler) [108]. Finally, we have to computationally scale up the capability of the module performing the Bayesian, sequential, four-level relevance analysis from hundred variables (enough currently for a restricted biomedical domain) to the range of thousand variables (usual in genomics). This means application of more advanced MCMC schemes (e.g., hierarchical-MCMC and coupled-MCMC) and more efficient parallel computing.

The use of informative priors for parametric black-box classifier, such as logistic regression and MLPs misses currently a natural step: the joint usage of informative structure and parameter priors. This can be done using MC methods over the space of model structures with varying dimensions and parameters [119].

One of the grand challenges besides the discussed scaling-up to genomic domains is the extension of the platform towards image analysis, i.e., to design an integrated Bayesian decision support system, which could utilize raw or crudely preprocessed images as the input, instead of the current manually constructed symbolic input features.

On the list of general challenges, currently we are focusing on the hierarchical, decomposed, annotated models, heterogeneous data sets, domain literature as electronic prior, and a rich language to formulate complex queries about the model posterior. Our goal is to provide a probabilistic framework for inference over linked, multiple, hierarchical models using heterogeneous data sets (including the domain literature), free-text annotations, and the power of first-order logic for formulating queries about the models.

Appendix A

Table A.1: The abbreviations and the short description of the domain variables taken from the IOTA protocol [240]. The ordering is the reference causal ordering $\prec_c$.

Variable	Description
(FamHistBrCa)	Number of first degree relatives with breast cancer.
(FamHistOvCa)	Number of first degree relatives with ovarian cancer.
(FamHist)	A derived nominal(6) variable for the genetic risk based on FamHistBrCa and FamHistOvCa.
(PMenoAge)	Derived from Age and PostMenoY.
(ReprYears)	Derived from Age, PostMenoY and the assumption of 12.0 years for the age at menarche.
(Meno)	Menopausal status.
(Age)	(years)
(PostMenoY)	Years after menopause.
(Hysterectomy)	
(CycleDay)	Day of cycle.
(PillUse)	Total years of oral contraceptive use.
(Parity)	Number of deliveries.
(HormTherapy)	Hormonal therapy.
(Pathology)	Manual peer-reviewed histopathology.
(PapFlow)	The presence of flow within a papillary projection.
(PapSmooth)	Solid papillary projections are described as being "smooth" or "irregular" (e.g., cauliflower-like)
(Papillation)	Solid papillary projection.
(Solid)	Solid means echogenicity suggesting the presence of tissue.
(WallRegularity)	The internal wall is also noted as being smooth or irregular.
(Septum)	A septum is defined as a thin strand of tissue running across the cyst cavity.
(IncomplSeptum)	An incomplete septum.
(Locularity)	All lesions are qualitatively classified as unilocular, unilocular cyst with solid component, multilocular, multilocular with solid component, solid and as "not classifiable as before".
(Echogenicity)	The dominant feature of the cystic contents is described as anechoic, low-level echogenic, "ground glass", hemorrhagic or mixed echogenic.
(Shadows)	The presence of acoustic shadows.
(TAMX)	Time-averaged maximum velocity.
(PSV)	Peak systolic velocity.
(PI)	Pulsatility index.
(RI)	Resistance index.
(ColScore)	A subjective semiquantitative assessment of the amount of blood flow.
(Volume)	The volume of the tumor is calculated from the three diameters in two perpendicular planes.
(Ascites)	The presence of ascites (i.e.fluid outside the pouch of Douglas).
(Fluid)	Fluid in the pouch of Douglas.
(Bilateral)	Patients with bilateral tumours are included in the study with both tumours.
(Pain)	Pelvic pain during the scan: "is the mass painful?"
(CA125)	CA 125 serum tumour marker.

Table A.2: Univariate statistics based on the IOTA-1.1 data set for the thirty-one variables containing 604 cases.

Variable	Value	%
Age	< 30	13.08
	$30 - 40$	18.21
	$40 - 50$	20.03
	$50 - 60$	22.19
	$60 - 70$	11.75
	$70 <$	14.74
Ascites (presence)	*yes*	16.56
Bilateral	*yes*	33.77
CA 125	< 35	61.09
	$35 - 65$	11.59
	$65 <$	27.32
ColScore	none	46.19
	minimal	27.98
	moderate	20.86
	strong	4.97
CycleDay	none	46.19
	$1 - 16$	27.98
	$16 - 40$	20.87
	$40 <$	4.97
Echogenicity	Anechoic	43.87
	Low-level	16.72
	'Ground glass'	20.20
	'Hemorrhagic'	9.90
	Mixed	11.09
	No cyst fluid	7.12
Familial history risk	normal	86.59
	increased	9.44
	significant	3.31
	high	0.170
	very high	0.50
Fluid	yes	29.64
Hormonal Therapy	yes	24.50
Hysterectomy	yes	8.11
Incomplete Septum	yes	5.30
Locularity	Unilocular	28.81
	Unilocular-solid	13.25
	Multilocular	15.23
	Multilocular-solid	29.30
	Solid	12.91
	Unclassified	0.50

Variable	Value	%
Menopause	pre	53.81
	post	38.08
	hysterectomy	8.11
Pain	yes	28.81
Papillation Flow	no papillation	74.67
	yes	14.40
Papillation	yes	74.67
Papillation smooth	no papillation	74.67
	yes	17.22
Parity	0	36.75
	1	20.20
	2	26.82
	3	10.10
	$4 <=$	6.13
Pathology (benign/malignant)	benign	72.35
	malignant	27.65
Pulsitality index (PI)	no flow	28.97
	< 1.0	44.37
PillUse (years)	< 0.5	52.32
	$0.5 - 5.0$	19.70
	$5.0 <$	27.98
PostMenopausal Years	< 10	14.74
	$10 - 20$	10.60
	$20 - 30$	6.79
	$30 - 405.96$	
	$40 <$	61.92
Peak Systolic Velocity (PSV)	no flow	28.97
	< 20.0	47.52
Resistance Index (RI)	no flow	28.97
	< 0.5	20.20
Septum	yes	43.71
Shadows	yes	8.28
Solid	yes	55.46
Time Averaged Velocity(TAMX)	no flow	28.97
	< 15.0	19.87
Volume(ml)	< 10	13.08
	$10 - 50$	27.31
	$50 - 40040.72$	
	$400 <$	19.20
WallRegularity	irregular	43.87

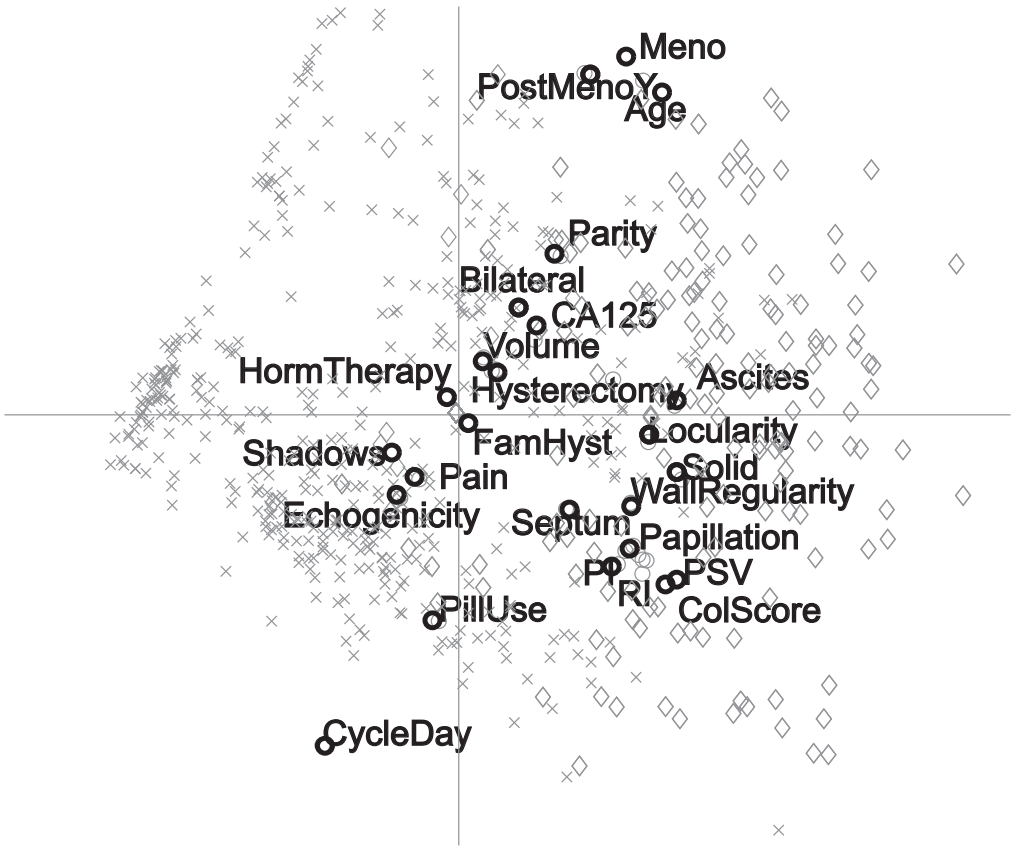


Figure A.1: The biplot of the domain variables and 604 cases used of the IOTA-1.1 data set (not all of the thirty-one variables are shown). The variables are denoted by '**o**', the malignant cases by '◇' and the benign cases by 'x'.

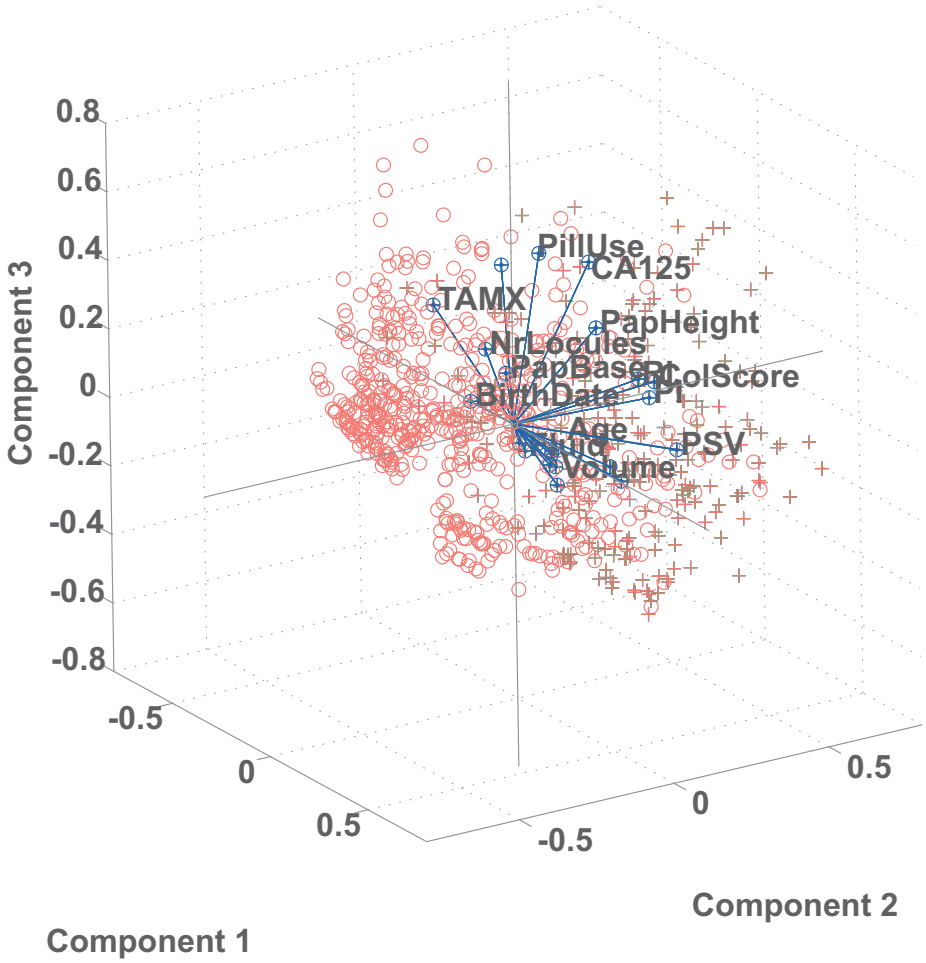


Figure A.2: The biplot of the domain variables and 782 cases of the IOTA-1.2 data set (not all of the thirty-five variables are labelled). The variables are denoted by '**o**', the malignant cases by '+' and the benign cases by 'o'.

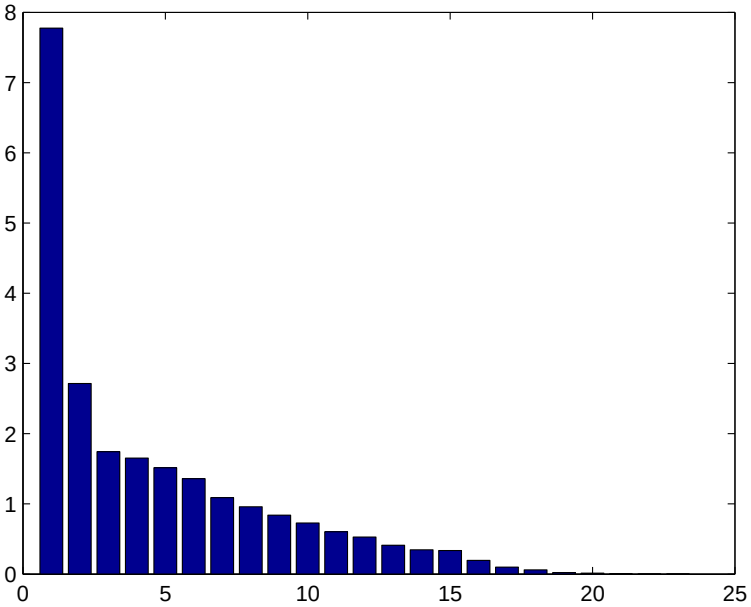


Figure A.3: The sorted eigenvalues of the covariance matrix of the IOTA-1.2 data set.

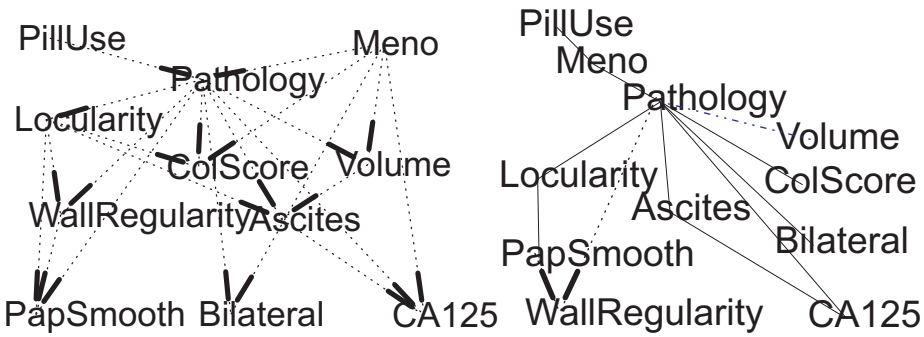


Figure A.4: (Left) The Bayesian network structure with eleven variables used in the parameter elicitation. (Right) The maximum a posteriori PDAG (Bayesian network equivalence class) over the eleven variables present in the parameter elicitation model (using the IOTA-1.2 data set, the $\mathrm{BD}_{\mathrm{eu}}$ parameter priors and noninformative structure priors in an exhaustive search.)

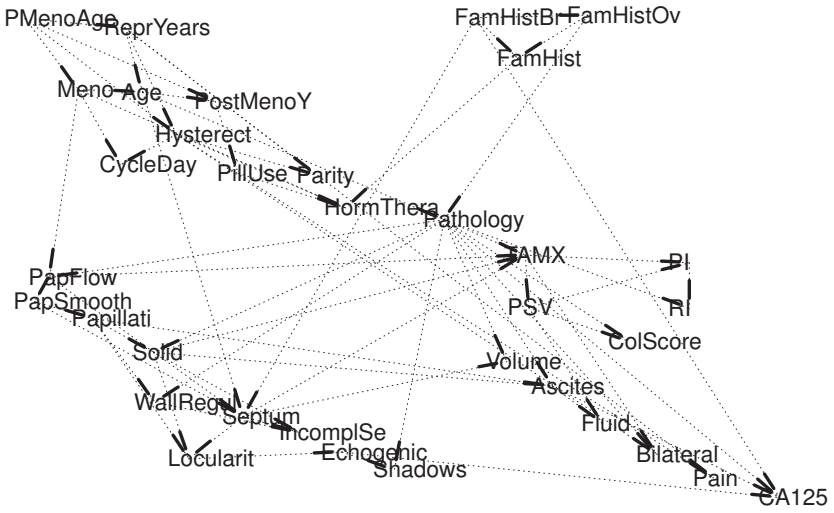


Figure A.5: The maximum a posteriori Bayesian network compatible with the expert's total ordering of the thirty-five variables using the IOTA-1.2 data set, the CH noninformative parameter priors and noninformative structure priors and exhaustive search to 3 parents with K2 greedy continuation.

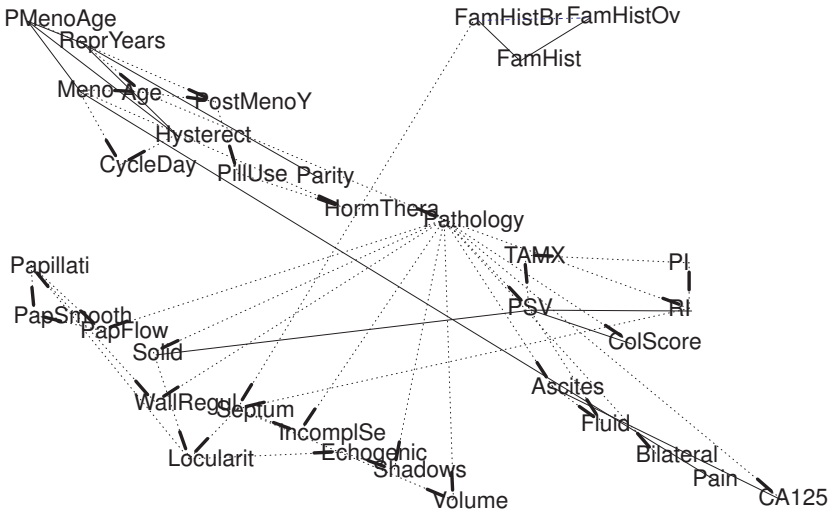


Figure A.6: The maximum a posteriori essential graph over the thirty-five variables using the IOTA-1.2 data set, the $\mathrm{BD}_{\mathrm{eu}}$ parameter priors and noninformative structure priors and exhaustive search to 3 parents with K2 greedy continuation over 10^6 random ordering.

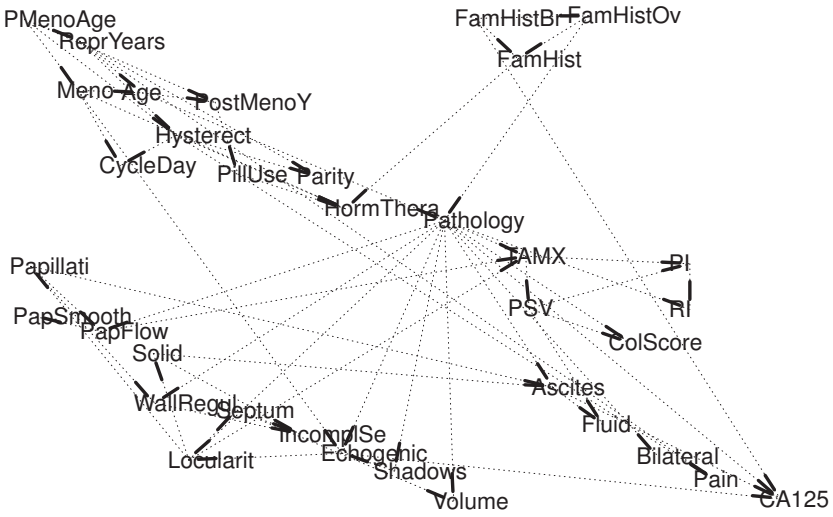


Figure A.7: The maximum a posteriori Bayesian network using the IOTA-1.2 data set, the CH noninformative parameter priors and noninformative structure priors and exhaustive search to 3 parents with K2 greedy continuation over 10^6 random ordering.

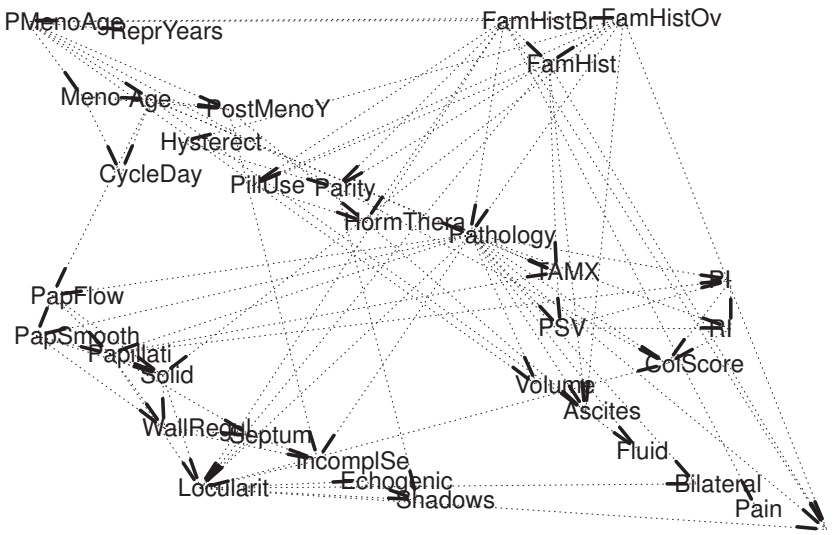


Figure A.8: The maximum a posteriori Bayesian network compatible with the expert's total ordering of the thirty-five variables using the $D^{PM^R_R}$ data set, the $\mathrm{BD}_{\mathrm{eu}}$ parameter priors and noninformative structure priors and exhaustive search to 3 parents with K2 greedy continuation.

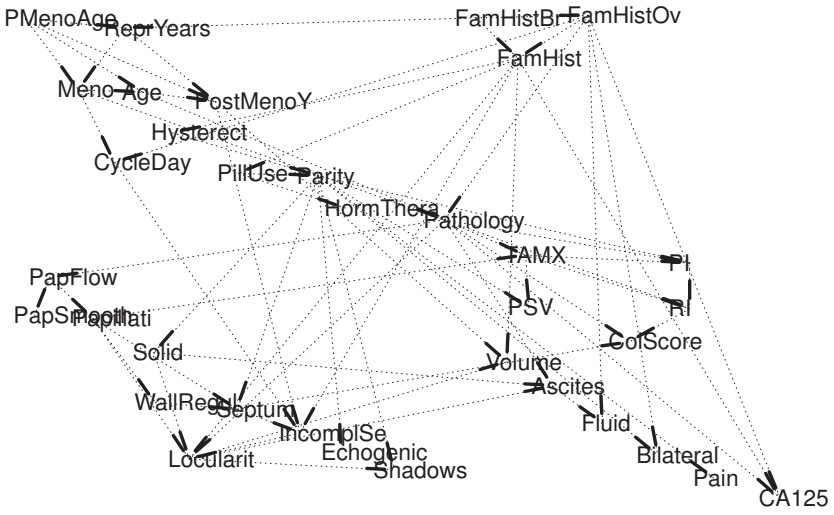


Figure A.9: The maximum a posteriori Bayesian network compatible with the expert's total ordering of the thirty-five variables using the $D^{PM_R^H}$ data set, the $\mathrm{BD_{eu}}$ parameter priors and noninformative structure priors and exhaustive search to 3 parents with K2 greedy continuation.

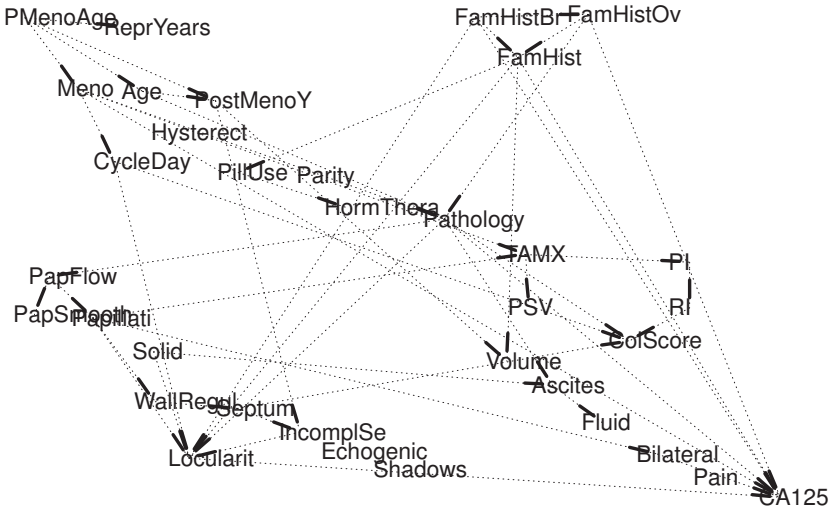


Figure A.10: The maximum a posteriori Bayesian network compatible with the expert's total ordering of the thirty-five variables using the $D^{PM_R^H}$ data set, the CH noninformative parameter priors and noninformative structure priors and exhaustive search to 3 parents with K2 greedy continuation.

Table A.3: The properties of the forward selected LR models over the elicited, medium and complete variable sets.

Variables	B	S.E.	Wald	df	Sig.
PersHistOvCa	2.88114	1.0702	7.2473	1	0.007100
dPostMenoY	9.9386			3	0.019095
dPostMenoY(1)	0.9766	0.3673	7.0692	1	0.007841
dPostMenoY(2)	0.8837	0.5098	3.0049	1	0.083008
dPostMenoY(3)	-0.4629	0.6658	0.4833	1	0.486931
lnVolume	0.2303	0.0894	6.6321	1	0.010015
Pain(1)	-0.7946	0.3480	5.2110	1	0.022443
Ascites(1)	1.2476	0.4194	8.8466	1	0.002936
Papillation(1)	-2.6262	0.6364	17.028	1	3.68E-05
PapNr	0.6078	0.1790	11.531	1	0.000684
PapFlow(1)	1.5077	0.4864	9.6081	1	0.001937
Solid(1)	3.0475	0.8185	13.862	1	0.000196
IrregularWall(1)	1.4707	0.3629	16.420	1	5.07E-05
lnCA125	0.7065	0.1154	37.462	1	9.31E-10
lnTAMX	0.7596	0.1509	25.315	1	4.86E-07
dPI(1)	-1.2664	0.3480	13.240	1	0.000273
NrLoc.byMLoc.(1)	0.4249	0.1747	5.9141	1	0.015020
Constant	-9.0173	0.9848	83.830	1	5.39E-20

Table A.4: The ordering conditional posteriors of the sets of parental sets in the expert's total ordering $\prec_0$. Only the logarithm of the eight most probable values are reported descending from left to right for each variable. The first 22 rows corresponds to the MBG case, the rest to the BN. The Pathology variable is not reported in the BN part as being identical. To easy comparison, the BN part does not report the aggregated parental sets (i.e., parental sets without the target variable), so this can be the only difference.

Pathology	-0.60	-1.00	-3.03	-4.39	-4.86	-5.12	-5.73	-5.96
PapFlow	0.00	-6.30	-6.50	-10.06	-12.99	-14.76	-19.59	-28.78
PapSmooth	0.00	-6.12	-445.99	-452.35	-452.54	-454.19	-458.67	-460.57
Papillation	-0.06	-3.48	-3.49	-435.62	-436.48	-437.86	-440.04	-442.35
Solid	-0.20	-2.39	-2.40	-65.30	-122.26	-122.66	-123.51	-125.01
WallReg.	-0.05	-3.74	-3.75	-25.43	-112.02	-170.96	-176.16	-178.47
Septum	0.00	-6.90	-12.58	-12.71	-12.92	-14.20	-15.43	-16.68
IncomplSep.	-0.28	-1.40	-12.56	-13.64	-14.31	-14.84	-14.91	-16.93
Locularity	0.00	-421.82	-424.46	-662.87	-674.67	-675.59	-780.30	-845.39
Echogenicity	0.00	-49.15	-83.43	-104.25	-105.84	-107.30	-113.34	-125.29
Shadows	0.00	-11.61	-42.24	-42.29	-42.56	-42.65	-43.89	-44.82
TAMX	0.00	-7.62	-13.01	-17.47	-20.81	-21.92	-22.03	-22.12
PSV	-0.01	-4.33	-607.86	-614.00	-620.51	-626.36	-628.41	-629.31
PI	0.00	-9.66	-14.56	-421.53	-432.88	-435.15	-438.35	-442.95
RI	0.00	-12.93	-92.14	-92.85	-500.80	-510.17	-513.65	-521.53
ColScore	-0.61	-0.79	-5.76	-33.16	-33.45	-439.34	-441.15	-445.63
Volume	0.00	-9.83	-14.29	-16.05	-18.34	-18.61	-20.27	-20.65
Ascites	-0.40	-1.87	-2.57	-3.67	-3.71	-3.88	-4.33	-5.35
Fluid	0.00	-5.79	-73.37	-74.30	-75.04	-75.26	-76.40	-76.43
Bilateral	-0.12	-3.19	-4.19	-4.28	-4.66	-4.82	-4.90	-5.11
Pain	-0.12	-2.50	-4.44	-5.23	-5.53	-5.54	-6.01	-6.97
CA125	0.00	-7.25	-14.75	-15.24	-16.78	-19.71	-20.41	-23.66
PapFlow	0.00	-6.30	-6.50	-10.06	-12.99	-14.76	-19.59	-28.78
PapSmooth	-6.12	-445.99	-452.35	-452.54	-454.19	-458.67	-460.57	-461.93
Papillation	-3.48	-3.49	-435.62	-436.48	-437.86	-440.04	-442.35	-443.21
Solid	-0.20	-2.39	-2.40	-122.26	-122.66	-123.51	-125.01	-125.70
WallReg.	-0.05	-3.74	-3.75	-112.02	-170.96	-176.16	-178.47	-180.01
Septum	-6.90	-12.58	-12.71	-12.92	-14.20	-15.43	-16.68	-16.77
IncomplSep.	-0.28	-12.56	-13.64	-14.31	-14.84	-14.91	-16.93	-17.02
Locularity	-421.82	-424.46	-662.87	-674.67	-675.59	-780.30	-845.39	-849.10
Echogenicity	-49.15	-83.43	-104.25	-105.84	-107.30	-113.34	-125.29	-127.75
Shadows	0.00	-42.24	-42.29	-42.56	-42.65	-43.89	-44.82	-45.15
TAMX	0.00	-7.62	-13.01	-17.47	-20.81	-21.92	-22.03	-22.12
PSV	-4.33	-607.86	-614.00	-620.51	-626.36	-628.41	-629.31	-629.71
PI	-9.66	-14.56	-421.53	-432.88	-435.15	-438.35	-442.95	-444.48
RI	0.00	-92.14	-92.85	-500.80	-510.17	-513.65	-521.53	-521.76
ColScore	-0.61	-0.79	-33.16	-33.45	-439.34	-441.15	-445.63	-453.21
Volume	0.00	-9.83	-14.29	-16.05	-18.34	-18.61	-20.27	-20.65
Ascites	-0.40	-1.87	-2.57	-3.67	-3.71	-3.88	-4.33	-5.35
Fluid	-5.79	-73.37	-74.30	-75.04	-75.26	-76.40	-76.43	-77.46
Bilateral	-0.12	-3.19	-4.28	-4.66	-4.82	-4.90	-5.11	-5.48
Pain	-2.50	-4.44	-5.23	-5.53	-5.54	-6.01	-6.97	-7.46
CA125	0.00	-7.25	-14.75	-15.24	-16.78	-19.71	-20.41	-23.66

Table A.5: The posteriors for the MBM(Pathology, X_i) features for the combinations of CH/BD_{eu} parameter priors and single/unconstrained orderings (Fix/MC) using the IOTA-1.2 data set and the maximum parental set size 4. The approximated posteriors based on the posteriors of the 100 most probable Markov blanket sets are reported respectively in the last 4 columns.

	Fix/CH	Fix/BD	MC/CH	MC/BD	MB-Fix/CH	MB-Fix/BD	MB-MC/CH	MB-MC/CH
FamH.	4.1E-01	7.5E-02	2.0E-01	1.5E-01	2.1E-01	4.4E-02	6.2E-02	1.9E-01
Age	9.9E-01	5.5E-01	1.1E-05	3.6E-07	9.3E-01	4.8E-01	7.9E-08	0.0E+00
Pari.	1.5E-03	6.5E-07	1.2E-03	2.9E-07	0.0E+00	0.0E+00	1.5E-06	0.0E+00
Meno	6.3E-01	4.6E-01	1.0E+00	9.8E-01	5.5E-01	3.8E-01	1.0E+00	9.8E-01
RYear.	1.7E-03	6.9E-06	1.7E-03	2.5E-06	0.0E+00	0.0E+00	1.5E-06	1.6E-06
CDay.	4.9E-03	6.5E-03	8.2E-04	9.6E-04	0.0E+00	0.0E+00	3.4E-13	2.7E-04
HTh.	5.5E-02	1.4E-01	3.0E-02	2.7E-02	0.0E+00	1.0E-01	6.5E-03	2.3E-02
PUse.	4.6E-04	2.5E-05	5.9E-03	3.7E-05	0.0E+00	0.0E+00	3.7E-04	0.0E+00
Bil.	9.1E-01	9.8E-01	9.4E-01	9.8E-01	8.5E-01	8.6E-01	9.5E-01	9.9E-01
Vol.	1.0E+00	1.0E+00	1.0E+00	1.0E+00	9.3E-01	8.6E-01	1.0E+00	1.0E+00
Pain	6.7E-02	1.1E-01	1.1E-01	1.4E-01	2.8E-03	5.9E-02	2.9E-02	1.1E-01
Asc.	1.0E+00	1.0E+00	1.0E+00	1.0E+00	9.3E-01	8.6E-01	1.0E+00	1.0E+00
Fluid	4.7E-02	1.2E-02	3.9E-02	9.5E-03	1.2E-03	0.0E+00	4.2E-03	4.9E-03
Sept.	9.9E-01	1.0E+00	9.8E-01	1.0E+00	9.3E-01	8.6E-01	1.0E+00	1.0E+00
ISept.	2.1E-01	7.8E-01	3.4E-01	8.8E-01	5.8E-02	6.9E-01	6.4E-01	8.4E-01
Pap.	1.0E+00	9.9E-01	7.5E-01	1.0E+00	9.3E-01	8.6E-01	8.9E-01	1.0E+00
PFl.	1.0E+00	1.0E+00	1.0E+00	1.0E+00	9.3E-01	8.6E-01	1.0E+00	1.0E+00
PSm.	1.5E-01	1.7E-01	3.1E-02	8.0E-03	4.1E-02	1.1E-01	1.5E-03	1.4E-03
Loc.	4.7E-01	8.7E-05	9.8E-01	7.9E-01	1.8E-01	0.0E+00	5.6E-01	8.3E-01
WReg.	1.0E+00	1.0E+00	1.0E+00	1.0E+00	9.3E-01	8.6E-01	1.0E+00	1.0E+00
Sh.	1.0E+00	1.0E+00	9.9E-01	1.0E+00	9.3E-01	8.6E-01	5.9E-01	1.0E+00
Egen.	1.0E+00	1.0E+00	1.0E+00	1.0E+00	9.3E-01	8.6E-01	1.0E+00	1.0E+00
CSc.	9.9E-01	9.9E-01	9.8E-01	9.9E-01	9.3E-01	8.6E-01	5.7E-01	9.6E-01
CA125	1.0E+00	1.0E+00	1.0E+00	1.0E+00	9.3E-01	8.6E-01	1.0E+00	1.0E+00
PI	9.8E-01	1.0E+00	5.7E-01	3.8E-01	9.2E-01	8.6E-01	4.8E-01	3.2E-01
RI	9.8E-01	1.0E+00	9.9E-01	1.0E+00	9.2E-01	8.6E-01	9.9E-01	1.0E+00
PSV	6.0E-01	5.4E-01	6.4E-01	8.5E-01	5.7E-01	4.7E-01	7.8E-01	8.9E-01
TAMX	1.0E+00	1.0E+00	8.6E-01	4.2E-01	9.3E-01	8.6E-01	6.0E-01	4.5E-01
Hyst.	8.2E-01	1.2E-01	7.7E-01	3.2E-02	8.4E-01	7.6E-02	9.0E-01	3.3E-02
Solid	1.0E+00	1.0E+00	9.3E-01	2.8E-01	9.3E-01	8.6E-01	9.7E-01	2.3E-01
PAge	1.0E-02	1.3E-04	5.8E-03	1.6E-05	0.0E+00	0.0E+00	9.7E-11	1.0E-05
FHOC	7.8E-01	6.7E-02	1.7E-01	5.5E-02	7.3E-01	2.7E-02	1.1E-01	3.2E-02
FHBC	9.6E-01	2.8E-01	8.7E-01	2.2E-01	9.1E-01	2.1E-01	9.4E-01	2.8E-01
PMY	3.4E-02	7.6E-06	5.0E-03	1.5E-06	0.0E+00	0.0E+00	2.4E-07	0.0E+00

Table A.6: The MCMC estimates of the posterior of the MBM(Pathology,.) features (w.r.t. one chain), their standard error, the maximum of the $|Z_G|$ single chain convergence test value Z_G per variable for all the chains, and the $\hat{R}$ multiple chain score per variable. The settings is the IOTA-1.2 data set, the BD_{eu} parameter prior and unconstrained ordering-MCMC simulation with 1000 burn-in sample, M=5000 used sample, four chains, and 10 batches for the std. error estimation.

Variable	$\hat{p}(MBM(Pathology, X_i)\|D_N)$	std.error	maxZ_G	$\hat{R}$
FamHist	0.1638	0.00754	0.2874	1.0028
Age	0.000001198	8.648E-07	0.1765	1.0044
Parity	2.365E-07	3.171E-08	0.0991	1.024
Meno	0.9884	0.0002863	0.6293	1.0044
ReprYears	0.000002625	1.488E-07	0.2253	1.0164
CycleDay	0.0008844	0.0002904	0.1021	1.0011
HormTherapy	0.02856	0.001065	0.2395	1.0036
PillUse	0.00004162	0.00000586	0.0997	1.1235
Bilateral	0.9886	0.0003922	0.1307	1.0008
Volume	1	0	0	0
Pain	0.1431	0.003487	0.1906	1.001
Ascites	1	0.000005155	0.1508	1.0026
Fluid	0.007042	0.0004665	0.1061	1.0004
Septum	1	7.108E-07	0.2921	1.0068
IncomplSeptum	0.877	0.005371	0.3086	1.0002
Papillation	1	1.938E-08	0.2806	1.0013
PapFlow	1	7.508E-07	0.4168	1.0031
PapSmooth	0.007686	0.0005047	0.1012	1.0002
Locularity	0.7928	0.05257	0.352	1.0124
WallRegularity	1	0	0	0
Shadows	0.9999	0.000002754	0.1652	1.0023
Echogenicity	1	0.000003014	0.7286	1.0085
ColScore	0.9771	0.01251	0.1206	1.0069
CA125	1	0	0	0
PI	0.3292	0.07948	0.3449	1.0157
RI	1	2.164E-07	0.4655	1.0094
PSV	0.9079	0.03741	0.5214	1.0126
TAMX	0.3632	0.05354	0.3141	1.0112
Hysterectomy	0.03078	0.002982	0.4107	1.0093
Solid	0.3022	0.05622	0.2821	1.0127
PMenoAge	0.00001654	7.112E-07	0.1754	1.0003
FamHistOvCa	0.06255	0.005788	0.2423	1.0012
FamHistBrCa	0.2088	0.006869	0.0394	1.0008
PostMenoY	0.000001546	1.656E-07	0.2115	1.0536

Table A.7: The most probable MB sets of the Pathology variable. The first ten column reports the Markov blanket sets of Pathology among the thirty-five variables using the IOTA-1.2 data set, the BD_{eu} parameter prior, a uniform structure prior for the fixed $\prec_0$ ordering, the second part of the unconstrained case using ordering-based MCMC simulation. The estimated posteriors with convergence and confidence values are reported in Table A.8.

FamHist	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
Age	1	1	0	0	1	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0
Parity	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Meno	0	0	1	1	0	0	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1
ReprYears	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CycleDay	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
HormTherapy	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PillUse	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Bilateral	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Volume	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Pain	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
Ascites	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Fluid	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Septum	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
IncomplSept.	1	1	1	1	1	0	1	1	0	1	1	1	1	0	1	1	0	1	1	1	1
Papillation	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
PapFlow	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
PapSmooth	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Locularity	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	0	1	
WallReg.	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Shadows	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Echogenicity	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
ColScore	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
CA125	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
PI	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	1	1	
RI	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
PSV	1	0	1	0	1	1	0	1	0	0	1	1	1	1	1	1	1	1	1	1	1
TAMX	1	1	1	1	1	1	1	1	1	1	0	1	0	0	1	0	1	0	1	1	
Hysterectomy	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Solid	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	1	0
PMenoAge	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
FamHistOvCa	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
FamHistBrCa	0	0	0	0	1	0	1	1	0	1	0	0	1	0	1	1	0	0	0	0	0
PostMenoY	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table A.8: The estimated posteriors with convergence and confidence values of the most probable MB sets of the Pathology variable reported in Table A.7 (in the unconstrained case). The columns report their estimated posterior $\hat{p}(\mathrm{mb}\,|D_N)$, the corresponding standard error, maximum of the $|Z_G|$ single chain convergence test value for all the chains, and the $\hat{R}$ multiple chain score. The IOTA-1.2 data set, the $\mathrm{BD}_{\mathrm{eu}}$ parameter prior were used in an unconstrained ordering-MCMC simulation with 10000 burn-in sample, M=4 chains, L=50000 used sample.

rank(Table A.7)	$\hat{p}(\mathrm{mb}\,\|D_N)$	std.error	max Z_G	$\hat{R}$
1.	0.1579	0.009062	0.2006	1.0022
2.	0.05419	0.004213	0.1396	1.0002
3.	0.04589	0.002821	0.0705	1.0004
4.	0.0237	0.001195	0.1204	1.0003
5.	0.01931	0.005172	0.3002	1.0073
6.	0.01892	0.007267	0.2766	1.009
7.	0.0162	0.001847	0.0986	1.0002

Table A.9: The most probable MBGs using the IOTA-1.2 data set, the $\mathrm{BD}_{\mathrm{eu}}$ parameter prior, and a uniform structure prior for the fixed $\prec_0$ ordering. The MBGs are reported in $child, parent^*$ form.

Pathology,Age Solid,Pathology,Papillation Age TAMX,Pathology,Septum WallRegularity,Pathology,Papillation Shadows,Pathology,Echogenicity Echogenicity ColScore,Pathology,TAMX CA125,Pathology,Ascites Bilateral,Pathology Volume,Pathology,Septum PI Ascites,Pathology RI,Pathology,PI Septum IncomplSeptum,Pathology,Septum Papillation PapFlow,Pathology
Pathology,Age Solid,Pathology,Papillation Age WallRegularity,Pathology,Papillation Shadows,Pathology,Echogenicity Echogenicity ColScore,Pathology,PSV CA125,Pathology,Ascites PI Bilateral,Pathology Volume,Pathology,Septum RI,Pathology,PI Ascites,Pathology PSV Septum IncomplSeptum,Pathology,Septum Papillation PapFlow,Pathology TAMX,Pathology,Septum
Pathology,Meno Solid,Pathology,Papillation TAMX,Pathology,Septum WallRegularity,Pathology,Papillation Meno Shadows,Pathology,Echogenicity Echogenicity ColScore,Pathology,TAMX CA125,Pathology,Ascites Bilateral,Pathology Volume,Pathology,Septum PI Ascites,Pathology RI,Pathology,PI Septum IncomplSeptum,Pathology,Septum Papillation PapFlow,Pathology
Pathology,Meno Solid,Pathology,Papillation WallRegularity,Pathology,Papillation Shadows,Pathology,Echogenicity Meno Echogenicity ColScore,Pathology,PSV CA125,Pathology,Ascites PI Bilateral,Pathology Volume,Pathology,Septum RI,Pathology,PI Ascites,Pathology PSV Septum IncomplSeptum,Pathology,Septum Papillation PapFlow,Pathology TAMX,Pathology,Septum
Pathology,Age TAMX,Pathology,Septum Age Solid,Pathology,Papillation WallRegularity,Pathology,Papillation Shadows,Pathology,Echogenicity Echogenicity ColScore,Pathology,PSV CA125,Pathology,Ascites Bilateral,Pathology Volume,Pathology,Septum PI Ascites,Pathology RI,Pathology,PI Septum PSV Papillation PapFlow,Pathology

Table A.10: The most probable MBGs of the Pathology variable in $child, parent^*$ form. The estimated posteriors with convergence and confidence values are reported in Table A.11. The IOTA-1.2 data set, the BD_{eu} parameter prior were used in an unconstrained ordering-MCMC simulation with 10000 burn-in sample

Pathology TAMX Locularity,Pathology WallRegularity,Pathology,Papillation Meno,Pathology,Echogenicity Shadows,Pathology,Echogenicity Echogenicity ColScore,Pathology,TAMX CA125,Pathology,Ascites Bilateral,Pathology Volume,Pathology,Septum PI Ascites,Pathology RI,Pathology,Septum Septum PSV,Pathology,PI Papillation PapFlow,Pathology,Papillation
Pathology,RI PapFlow,Pathology,Papillation RI Locularity,Pathology,Septum Meno,Pathology,Echogenicity WallRegularity,Pathology,Papillation Shadows,Pathology,Echogenicity Echogenicity ColScore,Pathology,PSV Bilateral,Pathology Volume,Pathology,Septum CA125,Pathology,Ascites Ascites,Pathology PSV,Pathology,RI Septum IncomplSeptum,Pathology,Septum Papillation
Pathology WallRegularity,Pathology,Papillation Shadows,Pathology,Echogenicity Echogenicity Meno,Pathology,Echogenicity ColScore,Pathology,TAMX CA125,Pathology,Ascites PI RI,Pathology,Septum Bilateral,Pathology Volume,Pathology,Septum PSV,Pathology,PI Ascites,Pathology TAMX Septum IncomplSeptum,Pathology,Septum,FamHistBrCa Papillation PapFlow,Pathology,Papillation FamHistBrCa Locularity,Pathology

Table A.11: The estimated posteriors with convergence and confidence values of the most probable MBGs of the Pathology variable reported in Table A.10. The columns report their estimated posterior $\hat{p}(\text{mbg}\,|D_N)$, the corresponding standard error, maximum of the $|Z_G|$ single chain convergence test value for all the chains, and the $\hat{R}$ multiple chain score. The IOTA-1.2 data set, the BD_{eu} parameter prior were used in an unconstrained ordering-MCMC simulation with 10000 burn-in sample, M=4 chains, L=50000 used sample.

rank(Table A.10)	$\hat{p}(\text{mbg}\,\|D_N)$	std.error	max Z_G	$\hat{R}$
1.	0.01109	0.001531	0.1729	1.0003
2.	0.00964	0.001339	0.0549	1
3.	0.009497	0.001073	0.1409	1.0003

Bibliography

[1] B. Abramson and K.-C. Ng. Towards an art and science of knowledge engineering. *IEEE Transactions on Knowledge and Data Engineering*, 5(4):705–711, 1993.

[2] Y. S. Abu-Mostafa. Hints and the VC dimension. *Neural Computation*, 5(2):278–288, 1993.

[3] S. Acid and L. M. de Campos. Searching for bayesian network structures in the space of restricted acyclic partially directed graphs. *Journal of Artificial Intelligence Research*, 18:445490, 2003.

[4] S. Acid, L. M. de Campos, and J. G. Castellano. Learning bayesian network classifiers: searching in a space of partially directed acyclic graphs. *Machine Learning*, 59:213–235, 2005.

[5] S. Aerts, P. Antal, B. De Moor, and Y. Moreau. Web-based data collection for ovarian cancer: a case study. In *Proc. of the 15th IEEE Symp. on Computer-Based Medical Sys. (CBMS-2002)*, pages 282–287, 2002.

[6] A. Agresti. *Categorical data analysis*. Wiley & Sons, 2002.

[7] M. E. Alfaro, S. Zoller, and F. Lutzon. Bayes or bootstrap? a simulation study comparin the performance of bayesian mcmc sampling and bootstrapping in assesing phylogenetic confidence. *Mol. Biol. Evol.*, 20(2):255–266, 2003.

[8] R. B. Altman. Challenges for intelligent systems in biology. *IEEE Intelligent Systems*, 16(6):14–18, 2002.

[9] C. Andrieu, A. Doucet, and C. P. Robert. Computational advances for and from bayesian analysis. *Statistical Science*, 19(1):118127, 2004.

[10] P. Antal. Applicability of prior domain knowledge formalised as Bayesian network in the process of construction of a classifier. In *Proc. of the 24th Annual Conf. of the IEEE Industrial Electronic Society (IECON '98)*, pages 2527–2531, 1998.

[11] P. Antal, G. Fannes, S. Van Huffel, B. De Moor, J. Vandewalle, and Dirk Timmerman. Bayesian predictive models for ovarian cancer classification: evaluation of logistic regression, multi-layer preceptron and belief network models in the Bayesian context. In *Proc. of the 10th Belgian-Dutch Conference on Machine Learning, BENELEARN 2000*, pages 125–132, 2000.

[12] P. Antal, G. Fannes, B. De Moor, J. Vandewalle, D. Timmerman, and Y. Moreau. From the acquisition of domain knowledge to its integration with data in Bayesian black-box classifiers: a comprehensive approach. Internal Report 01-11, ESAT-SISTA, K.U.Leuven (Leuven, Belgium), 2001.

[13] P. Antal, G. Fannes, Y. Moreau, and B. De Moor. Using domain literature and data to annotate and learn Bayesian networks. In *Proc. of 14th Belgian-Dutch Conference on Artificial Intelligence (BNAIC'02)*, pages 3–10, 2002.

[14] P. Antal, G. Fannes, Y. Moreau, and B. De Moor. Bayesian applications of belief networks and multilayer perceptrons for ovarian tumor classification with rejection. *Artificial Intelligence in Medicine*, 29:39–60, 2003.

[15] P. Antal, G. Fannes, Y. Moreau, B. De Moor, J Vandewalle, and D. Timmerman. Extended Bayesian regression models: a symbiotic application of belief networks and multilayer perceptrons for the classification of ovarian tumors. In *Lecture Notes in Artificial Intelligence (AIME 2001)*, pages 177–187, 2001.

[16] P. Antal, G. Fannes, Y. Moreau, D. Timmerman, and B. De Moor. Using literature and data to learn Bayesian networks as clinical models of ovarian tumors. *Artificial Intelligence in Medicine*, 30:257–281, 2004.

[17] P. Antal, G. Fannes, F. De Smet, and B. De Moor. Ovarian cancer classification with rejection by Bayesian belief networks. In *Workshop notes on Bayesian Models in Medicine, European Conference on Artificial Intelligence in Medicine (AIME'01)*, pages 23–27, 2001.

[18] P. Antal, G. Fannes, H. Verrelst, B. De Moor, and J. Vandewalle. Incorporation of prior knowledge in black-box models: Comparison of transformation methods from Bayesian network to multilayer perceptrons. In *Workshop on Fusion of Domain Knowledge with Data for Decision Support, 16th Uncertainty in Artificial Intelligence Conference*, pages 42–48, 2000.

[19] P. Antal, P. Glenisson, T. Boonefaes, P. Rottiers, and Y. Moreau. Towards an integrated usage of expression data and domain literature in gene clustering: representations and methods. Internal Report 01-69, ESAT-SISTA, K.U.Leuven (Leuven, Belgium), 2001.

[20] P. Antal, P. Glenisson, G. Fannes, J. Mathijs, Y. Moreau, and B. De Moor. On the potential of domain literature for clustering and Bayesian network

learning. In *Proc. of the 8th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (ACM-KDD-2002)*, pages 405–414, 2002.

[21] P. Antal, G. Hullám, A. Gézsi, and A. Millinghoffer. Learning complex bayesian network features for classification. In *Proc. of third European Workshop on Probabilistic Graphical Models*, pages 9–16, 2006.

[22] P. Antal, T. Meszaros, B. De Moor, and T. Dobrowiecki. Annotated Bayesian networks: a tool to integrate textual and probabilistic medical knowledge. In *Proc. of the 13th IEEE Symp. on Comp.-Based Med. Sys. (CBMS-2001)*, pages 177–182, 2001.

[23] P. Antal, T. Meszaros, B. De Moor, and T. Dobrowiecki. Domain knowledge based information retrieval language: an application of annotated Bayesian networks in ovarian cancer domain. In *Proc. of the 15th IEEE Symp. on Computer-Based Medical Sys. (CBMS-2002)*, pages 213–218, 2002.

[24] P. Antal and A. Millinghoffer. Learning causal bayesian networks from literature data. In *Proceedings of the 3rd International Conference on Global Research and Education, Inter-Academia'04*, pages 149–160, 2004.

[25] P. Antal and A. Millinghoffer. A probabilistic knowledge base using annotated bayesian network features. In *Proceedings of the 6th International Symposium of Hungarian Researchers on Computational Intelligence*, pages 1–12, 2005.

[26] P. Antal and A. Millinghoffer. Literature mining using bayesian networks. In *Proc. of third European Workshop on Probabilistic Graphical Models*, pages 17–24, 2006.

[27] P. Antal, H. Verrelst, D. Timmerman, S. Van Huffel, B. De Moor, and I. Vergote. How might we combine the information we know about a mass better? The use of mathematical models to handle medical data. Internal Report 00-145, ESAT-SISTA, K.U.Leuven (Leuven, Belgium), 1st Monte Carlo Conference on Updates in Gynaecology, Monaco, 2000.

[28] P. Antal, H. Verrelst, D. Timmerman, Y. Moreau, S. Van Huffel, B. De Moor, and I. Vergote. Bayesian networks in ovarian cancer diagnosis: Potential and limitations. In *Proc. of the 13th IEEE Symp. on Comp.-Based Med. Sys. (CBMS-2000)*, pages 103–109, 2000.

[29] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. ACM Press, New York, 1999.

[30] R. Bakerman and J. M. Gottman. *Observing interaction: An introduction to sequential analysis*. Cambridge University Press, 1986.

[31] J. O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag, 1980.

[32] J. O. Berger. Bayesian analysis: A look at today and thoughts of tomorrow. *Journal of the American Statistical Association*, 95(452):1269–1276, 2000.

[33] J. S. Berk and N. F. Hacker. *Practical Gynecologic Oncology*. Williams and Wilkins, 1995.

[34] J. M. Bernardo. *Bayesian Theory*. Wiley & Sons, Chichester, 1995.

[35] V. Berry and O. Gascuel. On the interpretation of bootstrap trees: appropriate threshold of clade selection and induced gain. *Mol. Biol. Evol.*, 13(7):999–1011, 1996.

[36] C. M. Bishop. *Neural Networks for Pattern Recognition*. Clarendon Press, Oxford, 1995.

[37] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *Journal of the ACM*, 36(4):929–965, 1989.

[38] R. C. Bodner and F. Song. Knowledge-based approaches to query expansion in information retrieval. In *Canadian Conference on AI*, pages 146–158, 1996.

[39] R. R. Bouckaert. *Bayesian Belief Networks: From construction to inference*. Ph.D. Thesis, Dept. of Comp. Sci., Utrecht University, Netherlands, 1995.

[40] W. L. Buntine. Theory refinement of Bayesian networks. In *Proc. of the 7th Conf. on Uncertainty in Artificial Intelligence (UAI-1991)*, pages 52–60. Morgan Kaufmann, 1991.

[41] B. Van Calster, D. Timmerman, C. Lu, J.A.K. Suykens, L. Valentin, C. Van Holsbeke, F. Amant, I. Vergote, and S. Van Huffel. Preoperative diagnosis of ovarian tumors using bayesian kernel-based methods. *Ultrasound Obstetrics Gynecology*, 29(5):496–504, 2007.

[42] R. Castelo and A. Siebes. Priors on network structures. biasing the search for Bayesian networks. *International Journal of Approximate Reasoning*, 24(1):39–57, 2000.

[43] P. Cheeseman. In defense of probability. In *Proceedings of the Ninth International Joint Conference on Artificial Intelligence (IJCAI-85)*, pages 1002–1009. Morgan Kaufmann, 1985.

[44] H. Chen, K. J. Lynch, K. Basu, and T. D. Ng. Generating, integrating, and activating thesauri for concept-based document retrieval. *IEEE Intelligent Systems*, 8(2):25–34, 1993.

[45] M. Chen, Q. Shao, and J. G. Ibrahim. *Monte Carlo Methods in Bayesian Computation*. Springer-Verlag, New York, 2000.

[46] J. Cheng, D. A. Bell, and W. Liu. Learning belief networks from data: an information theory based approach. In *Proc. of the 6th ACM International Conference on Information and Knowledge Management, CIKM'97*, pages 325–331, 1997.

[47] J. Cheng and R. Greiner. Comparing Bayesian network classifiers. In *Proc. of the 15th Conf. on Uncertainty in Artificial Intelligence (UAI'99)*, pages 101–107. Morgan Kaufmann, 1999.

[48] J. Cheng and R. Greiner. Learning Bayesian belief network classifiers: Algorithms and system. *Lecture Notes in Computer Science*, 2056:141–151, 2001.

[49] D. M. Chickering. A transformational characterization of equivalent Bayesian network structures. In *Proc. of 11th Conference on Uncertainty in Artificial Intelligence (UAI-1995)*, pages 87–98. Morgan Kaufmann, 1995.

[50] D. M. Chickering, D. Geiger, and D. Heckerman. Learning bayesian networks: Search methods and experimental results. In *Proceedings of Fifth Conference on Artificial Intelligence and Statistics*, pages 112–128, 1995.

[51] J. J. Cimino. Desiderata for controlled medical vocabularies in the twenty-first century. *Methods of Information in Medicine*, 37(4-5):394–403, 1998.

[52] E. B. Claus, N. Risch, and W. D. Thompson. Autosomal dominant inheritance of early-onset breast cancer. *Cancer*, 73(3):643–650, 1994.

[53] I. Cloete and J. M. Zurada. *Knowledge-Based Neurocomputing*. MIT Press, Cambridge, MA, 2000.

[54] G. Cooper. A simple constraint-based algorithm for efficiently mining observational databases for causal relationships. *Data Mining and Knowledge Discovery*, 2:203–224, 1997.

[55] G. F. Cooper. The computational complexity of probabilistic inference using bayesian belief network. *Artificial Intelligence*, 42:393–405, 1990.

[56] G. F. Cooper, C.F. Aliferis, R. Ambrosino, J. Aronis, B. G. Buchanan, R. Caruana, M. J. Fine, C. Glymour, G. Gordon, B. H. Hanusa, J. E. Janosky, C. Meek, T. Mitchell, T. Richardson, and P. Spirtes. An evaluation of machine-learning methods for predicting pneumonia mortality. *Artificial Intelligence in Medicine*, 9:107–138, 1997.

[57] G. F. Cooper and E. Herskovits. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9:309–347, 1992.

[58] G. F. Cooper and C. Yoo. Causal discovery from a mixture of experimental and observational data. In *Proc. of the 15th Conf. on Uncertainty in Artificial Intelligence (UAI-1999)*, pages 116–125. Morgan Kaufmann, 1999.

[59] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley & Sons, 2001.

[60] R. G. Cowell, A. P. Dawid, S. L. Lauritzen, and D. J. Spiegelhalter. *Probabilistic networks and expert systems*. Springer-Verlag, New York, 1999.

[61] W. Bruce Croft. Knowledge-based and statistical approaches to text retrieval. *IEEE Intelligent Systems*, 8(2):8–12, 1993.

[62] J. Cussens. *Statistical Relational Learning*, chapter Logic-based formalisms for statistical relational learning. MIT Press, Cambridge, MA, 2007.

[63] D. Timmerman D., H. Verrelst, and J. Vandewalle. Advanced statistical techniques: how to use your data more efficiently? *Ultrasound in Obstetrics and Gynecology*, 13(1):4–6, 1999.

[64] P. Dagum and M. Luby. Approximating probabilistic inference in Bayesian belief networks is NP-hard. *Artificial Intelligence*, 60:141–153, 1993.

[65] S. Dasgupta. The sample complexity of learning fixed-structure Bayesian networks. *Machine Learning*, 29:165–180, 1997.

[66] D. Dash and G. F. Cooper. Exact model averaging with naive bayesian classifiers. In *Proceedings of the Nineteenth International Conference on Machine Learning*, pages 91–98, 2002.

[67] R. Davis, H. Schrobe, and P. Szolovits. What is knowledge representation? *AI Magazine*, 14(1):17–33, 1993.

[68] A. P. Dawid. *Encylopedia of Statistical Sciences*, volume 1, chapter Prequential Analysis, pages 464–470. Wiley & Sons, 1997.

[69] A. P. Dawid. Probability, causality and the empirical world: A bayes-de finetti-pooper-borel synthesis. *Statistical Science*, 19(1):44–57, 2004.

[70] A. P. Dawid and V. G. Vovk. Prequential probability: Principles and properties. *Bernoulli*, 5:125–162, 1999.

[71] L. M. de Campos, J. M. Fernández, and J. F. Huete. Query expansion in information retrieval systems using a Bayesian network-based thesaurus. In *Proc. of the 14th Conf. on Uncertainty in Artificial Intelligence (UAI-1998)*, pages 53–60. Morgan Kaufmann, 1998.

[72] D. G. T. Denison, C. C. Holmes, B. K. Mallick, and A. F. M. Smith. *Bayesian Methods for Nonlinear Classification and Regression*. Wiley & Sons, 2002.

[73] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer-Verlag, Berlin, 1996.

[74] J. Ding, D. Berleant, D. Nettleton, and E. Wurtele. Mining medline: Abstracts, sentences, or phrases? In *Proc. of Pacific Symposium on Biocomputing (PSB 2002)*, pages 326–337, 2002.

[75] P. Domingos and M. Pazzani. On the optimality of the simple bayesian classifier under zero-one loss. *Machine Learning*, 29:103–130, 1997.

[76] P. Domingos and M. Richardson. Markov logic networks. *Machine Learning*, 62:107–136, 2006.

[77] M. J. Druzdzel, A. Oniśko, D. Schwartz, J. N. Dowling, and H. Wasyluk. Knowledge engineering for very large decision-analytic medical models. In *Proc. of the 1999 Annual Meeting of the American Medical Informatics Association (AMIA-99)*, page 1049, Washington, D.C., November 6-10 1999.

[78] M. J. Druzdzel and H. Simon. Causality in bayesian belief networks. In David Heckerman and Abe Mamdani, editors, *Proceedings of the 9th Conf. on Uncertainty in Artificial Intelligence (UAI-1993)*, pages 3–11. Morgan Kaufmann, 1993.

[79] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern classification*. Wiley & Sons, 2001.

[80] R. Durbin, S. R. Eddy, and A. Krogh anf G. Mitchison. *Biological Sequence Analysis : Probabilistic Models of Proteins and Nucleic Acids*. Chapman & Hall, London, 1995.

[81] D. F. Easton, D. Ford, and D. T. Bishop. Breast and ovarian cancer incidence in BRCA1-mutation. *American Journal of Human Genetics*, 56:265–271, 1995.

[82] B. Efron, E. Halloran, and S. Holmes. Bootstrap confidence level for phylogenetic trees. *Proc. Natl. Acad. Sci.*, 93:13429–34, 1996.

[83] B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall, London, 1993.

[84] J.W. Egar and M.A. Musen. Graph-grammar assistance for automated generation of influence diagrams. *IEEE Transactions on Systems Man and Cybernetics*, 24(11):1625–1642, 1994.

[85] G. Fannes. *Transforming informative priors between Bayesian networks and multilayer perceptrons*. Ph.D. dissertation, Leuven University Press, 2004.

[86] J. Felsenstein and H. Kishino. Is there something wrong with bootstrap on phylogenies? *Syst. Biol.*, 42(2):193–200, 1993.

[87] N. J. Finkler, B. Benaceraf, and P. T. Lavin. Comparison of serum CA 125, clinical impression, and ultrasound in the preoperative evaluation of ovarian masses. *Obstetrics & Gynecology*, 72(4):659–663, 1998.

[88] W. B. Frakes. *Information retrieval: Data Structures and Algorithms*, chapter Stemming Algorithms. Prentice Hall, 1992.

[89] J. H. Friedman. On bias, variance, 0/1-loss, and the curse of dimensionality. *Data Mining and Knowledge Discovery*, 1(1):55–77, 1997.

[90] N. Friedman. The bayesian structural EM algorithm. In *Proc. of the 14th Conf. on Uncertainty in Artificial Intelligence(UAI-1998)*, pages 129–138. Morgan Kaufmann, 1998.

[91] N. Friedman, D. Geiger, and M. Goldszmidt. Bayesian networks classifiers. *Machine Learning*, 29:131–163, 1997.

[92] N. Friedman, L. Getoor, D. Koller, and A. Pfeffer. Learning probabilistic relational models. In *Proc. of the 16th Intl. Joint Conf. on Artificial Intelligence*, pages 1300–1309, 1999.

[93] N. Friedman and M. Goldszmidt. Discretizing continuous attributes while learning Bayesian networks. In *Proc. 13th Int. Conf. on Machine Learning (ICML)*, pages 157–165, 1996.

[94] N. Friedman and M. Goldszmidt. Learning Bayesian networks with local structure. In Eric Horvitz and Finn V. Jensen, editors, *Proc. of the 20th Conf. on Uncertainty in Artificial Intelligence (UAI-1996)*, pages 252–262. Morgan Kaufmann, 1996.

[95] N. Friedman, M. Goldszmidt, and A. Wyner. Data analysis with bayesian networks: A Bootstrap approach. In *Proc. of the 15th Conf. on Uncertainty in Artificial Intelligence(UAI-1999)*, pages 196–205. Morgan Kaufmann, 1999.

[96] N. Friedman, M. Goldszmidt, and A. Wyner. On the application of the bootstrap for computing confidence measures on features of induced bayesian networks. In *AI&STAT VII.*, 1999.

[97] N. Friedman and D. Koller. Being Bayesian about network structure. In *Proc. of the 16th Conf. on Uncertainty in Artificial Intelligence(UAI-2000)*, pages 201–211. Morgan Kaufmann, 2000.

[98] N. Friedman and D. Koller. Being Bayesian about network structure. *Machine Learning*, 50:95–125, 2003.

[99] N. Friedman and Z. Yakhini. On the sample complexity of learning Bayesian networks. In *Proc. of the 12th Conf. on Uncertainty in Artificial Intelligence (UAI-1996)*, pages 274–282. Morgan Kaufmann, 1996.

[100] R. M. Fung, S. L. Crawford, L. A. Appelbaum, and R. M. tong. An architecture for probabilistic concept-based information retireval. In *Canadian Conference on AI*, pages 146–158, 1996.

[101] D. Galles and J. Pearl. Axioms of causal relevance. *Artificial Intelligence*, 97(1-2):9–43, 1997.

[102] D. Gamerman. *Markov Chain Monte Carlo*. Chapman & Hall, London, 1997.

[103] E. Garfield. *Essays of an Information Scientist*, chapter Towards the World Brain. ISI Press, Cambridge, MA, 1977.

[104] D. Geiger and D. Heckerman. A characterization of the Dirichlet distribution with application to learning Bayesian networks. In Philippe Besnard, Steve Hanks, Philippe Besnard, and Steve Hanks, editors, *Proc. of the 11th Conf. on Uncertainty in Artificial Intelligence (UAI-1995)*, pages 196–207. Morgan Kaufmann, 1995.

[105] D. Geiger and D. Heckerman. Knowledge representation and inference in similarity networks and Bayesian multinets. *Artificial Intelligence*, 82:45–74, 1996.

[106] D. Geiger and D. Heckerman. Parameter priors for directed acyclic graphical models and the characteriration of several probability distributions. *The Annals of Statistics*, 30(2):216–225, 2002.

[107] A. E. Gelfand and A. F. M. Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85(410):398–409, 1990.

[108] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Chapman & Hall, London, 1995.

[109] S. Geman, S. Bienenstock, and R. Doursat. Neural networks and the bias/variance dilemma. *Neural Computation*, 4:1–58, 1992.

[110] A. Genkin, D. D. Lewis, and D. Madigan. Bayesian logistic regression software, 2004.

[111] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter. *Markov Chain Monte Carlo in Practice*. Chapman & Hall, London, 1996.

[112] P. Giudici and R. Castelo. Improving Markov Chain Monte Carlo model search for data mining. *Machine Learning*, 50:127–158, 2003.

[113] P. Glenisson. *Integrating scientific literature with large scale gene expression analysis*. Ph.D. dissertation, Leuven University Press, 2004.

[114] P. Glenisson, P. Antal, J. Mathys, Y. Moreau, and B. De Moor. Evaluation of the vector space representation in text-based gene clustering. In *Proc. of the Pacific Symposium on Biocomputing (PSB03)*, pages 391–402, 2003.

[115] P. Glenisson, B. Coessens, S. Van Vooren, J. Mathijs, Y. Moreau, and B. De Moor. Txtgate: Profiling gene groups with text-based information. *Genome Biology*, 5(6), 2004.

[116] C. Glymour and G. F. Cooper. *Computation, Causation, and Discovery.* AAAI Press, 1999.

[117] P. Glynn and D. Ormoneit. Hoeffding's inequality for uniformly ergodic markov chains. *Statistics and Probability Letters*, 56:143–146, 2002.

[118] S. Granberg, M. Wikland, and I. Jansson. Macroscopic characterization of ovarian tumors and the relation to the histological diagnosis. *Gynecological Oncology*, 35:139–144, 1989.

[119] P. J. Green. Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrica*, 82:711–732, 1995.

[120] L. Győrfi. *T omegkiszolgálás informatikai rendszerekben.* Műegyetemi Kiadó, 1996. (in Hungarian).

[121] J. Y. Halpern. An analysis of first-order logics of probability. *Artificial Intelligence*, 46:311–350, 1990.

[122] D. J. Hand. *Construction and Assessment of Classification Rules.* Wiley & Sons, Chichester, 1997.

[123] J. A. Hanley and B. J. McNeil. The meaning and use of the area under receiver operating characteristic (ROC) curve. *Radiology*, 143:29–36, 1982.

[124] R. Harris, A. S. Whittemore, J. Itnyre, and the Collaborative Ovarian Cancer Group. Characteristics relating to ovarian cancer risk (i, ii, iii, iv). *American Journal of Epidemiology*, 136:1175–1220, 1992.

[125] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning: data minig inference and prediction.* Springer-Verlag, 2001.

[126] D. Haussler. Quantifying inductive bias: AI learning algorithms and Valiant's learning framework. *Artificial Intelligence*, 36:177–221, 1988.

[127] D. Haussler. Bounds on the sample complexity of Bayesian learning using information theory and the Vapnik-Chervonenkis dimension. *Machine Learning*, 14:83–113, 1994.

[128] D. Haussler. Bounds on the sample complexity of Bayesian learning using information theory and the vc dimension. *Machine Learning*, 14:83–113, 1994.

[129] S. Haykin. *Neural Networks, A Comprehensive Foundation*. MacMillen College Publishing Company, New York, 1995.

[130] D. Heckerman and J. S. Breese. Causal independence for probability assesment and inference using bayesian networks. *IEEE, Systems, Man, and Cybernetics*, 26:826–831, 1996.

[131] D. Heckerman, D. Geiger, and D. Chickering. Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20:197–243, 1995.

[132] D. Heckermann. A tutorial on learning with Bayesian networks., 1995. Technical Report, MSR-TR-95-06.

[133] D. Heckermann, C. Meek, and G. Cooper. A bayesian aproach to causal discovery. Technical Report, MSR-TR-97-05, 1997.

[134] D. M. Hillis and J. J. Bull. An empirical test of bootstrapping as a method for assessing confidence in phylogeneteci analysis. *Syst. Biol.*, 42(2):182–192, 1993.

[135] L. Hirschman, J. C. Park, J. Tsujii, L. Wong, and C. H. Wu. Accomplishments and challenges in literature data mining for biology. *Bioinformatics*, 18:1553–1561, 2002.

[136] J. A Hoeting, D. Madigan, A. E. Raftery, and C. T. Volinsky. Bayesian model averaging: A tutorial. *Statistical Science*, 14(4):382–417, 1999.

[137] C. Van Holsbeke, B. Van Calster B., L. Valentin L., A.C. Testa, E. Ferrazzi, I. Dimou, C. Lu, P. Moerman, S. Van Huffel, I. Vergote, and D. Timmerman. External validation of mathematical models to distinguish between benign and malignant adnexal tumors: a multicenter study by the international ovarian tumor analysis (iota) group. *Clinical Cancer Research*, 13(15):4440–7, 2007.

[138] D. W. Hosmer and S. Lemeshow. *Applied Logistic Regression*. Wiley & Sons, Chichester, 2000.

[139] http://thymoma.de/meddict.htm. The cancernet dictionary, 2000.

[140] http://www.graylab.ac.uk/omd/index.html. On-line medical dictionary, 2000.

[141] http://www.merck.com/pubs/mmanual/. The merck manual, 2000.

[142] http://www.ncbi.nlm.nih.gov/PubMed/. Pubmed/medline, 2000.

[143] C. Huang and A. Darwiche. Inference in belief networks: A procedural guide. In *International Journal of Approximate Reasoning*, volume 15, pages 225–263. Elsevier Science Inc., 1996.

[144] I. Inza, P. Larranaga, and B. Sierra. Bayesian networks for feature subset selection. In *Proceedings of the Workshop on Bayesian and Causal Networks (CaNew2000), ECAI2000*, pages 143–164, 1997.

[145] Manfred Jaeger. Relational bayesian networks. *Proc. of the 13th Conference on Uncertainty in Artificial Intelligence (UAI-1997)*, pages 266–273, 1997.

[146] W. H. Jefferys and J. O. Berger. Sharpening ockham's razor on a bayesian strop, 1991.

[147] T. K. Jenssen, A. Laegreid, J. Komorowski, and E. Hovig. A literature network of human genes for high-throughput analysis of gene expression. *Nature Genetics*, 28:21–28, 2001.

[148] G. H. John, R. Kohavi, and K. Pfleger. Irrelevant features and the feature subset selection problem. In *Proc. of the 11th International Conference on Machine Learning*, volume 97, pages 121–129. Morgan Kaufmann, 1994.

[149] R. E. Kass and A. E. Raftery. Bayes factors. Technical Report no. 254, Dept. of Statistics, University of Washington, 1994.

[150] E. Keogh and M. Pazzani. Learning the structure of augmented bayesian classifiers. *International Journal of Artificial Intelligence Tools*, 11(4):587–601, 2002.

[151] K. Kersting and L. De Raedt. Bayesian logic programs. In J. Cussens and A. Frisch, editors, *Proceedings of the Work-in-Progress Track at the 10th International Conference on Inductive Logic Programming*, pages 138–155, 2000.

[152] T. Kocka and R. Castelo. Improved learning of Bayesian networks. In Jack S. Breese and Daphne Koller, editors, *Proc. of the 17th Conference on Uncertainty in Artificial Intelligence (UAI-2001)*, pages 269–276. Morgan Kaufmann, 2001.

[153] R. Kohavi and G. H. John. Wrappers for feature subset selection. *Artificial Intelligence*, 97:273–324, 1997.

[154] M. Koivisto and K. Sood. Exact bayesian structure discovery in bayesian networks. *Journal of Machine Learning Research*, 5:549–573, 2004.

[155] D. Koller and A. Pfeffer. Object-oriented Bayesian networks. In Dan Geiger and Prakash P. Shenoy, editors, *Proc. of the 13th Conf. on Uncertainty in Artificial Intelligence (UAI-1997)*, pages 302–313. Morgan Kaufmann, 1997.

[156] D. Koller and A. Pfeffer. Probabilistic frame-based systems. In *Proc. of the 15th National Conference on Artificial Intelligence (AAAI), Madison, Wisconsin*, pages 580–587, 1998.

[157] D. Koller and M. Sahami. Toward optimal feature selection. In *International Conference on Machine Learning*, pages 284–292, 1996.

[158] P. Kontkanen, P. Myllymäki, T. Silander, and H. Tirri. On supervised selection of bayesian networks. In *Proc. of the 15th Conference on Uncertainty in Artificial Intelligence (UAI-1999)*, pages 334–342. Morgan Kaufmann, 1999.

[159] P. Kontkanen, P. Myllymäki, T. Silander, and H. Tirri. Urban legends in bayesian network research i: Model selection for supervised problems. *Arpakannus*, 1:8–14, 1999.

[160] P. Kontkanen, P. Myllymäki, and H. Tirri. Comparing prequential model selection criteria in supervised learning of mixture models. In *Proc. of the 18th Conference on Uncertainty in Artificial Intelligence (UAI-2001)*, pages 233–238. Morgan Kaufmann, 2001.

[161] R. Korfhage. *Information Storage and Retrieval*. Wiley & Sons, Chichester, 1997.

[162] M. Krauthammer, P. Kra, I. Iossifov, S. M. Gomez, G. Hripcsak, V. Hatzivassiloglou, C. Friedman, and A. Rzhetstky. Of truth and pathways: chasing bits of information through myriads of articles. *Bioinformatics*, 18:249–257, 2002.

[163] W. Lam and F. Bacchus. Using causal information and local measures to learn Bayesian networks. In David Heckerman and Abe Mamdani, editors, *Proc. of the 9th Conference on Uncertainty in Artificial Intelligence (UAI-1993)*, pages 243–250. Morgan Kaufmann, 1993.

[164] H. Langseth and T. D. Nielsen. Classification using hierarchical naive bayes models. *Machine Learning*, 63(2):135–159, 2006.

[165] A. A. Langston. Hereditary ovarian cancer. *Gynecological Oncology And Pathology*, 9:3–7, 1997.

[166] P. Larrañaga, C. M. H. Kuijpers, R. H. Murga, and Y. Yurramendi. Learning bayesian network structures by searching for the best ordering with genetic algorithms. *IEEE Transactions on Systems, Man, and Cybernetics*, 26(4):487–493, 1996.

[167] K. Laskey and S. Mahoney. Network fragments: Representing knowledge for constructing probabilistic models. In Dan Geiger and Prakash P. Shenoy, editors, *Proc. of the 13th Conf. on Uncertainty in Artificial Intelligence (UAI-1997)*, pages 334–341. Morgan Kaufmann, 1997.

[168] K.B. Laskey and S.M. Mahoney. Network engineering for agile belief network models. *IEEE Transactions on Knowledge and Data Engineering*, 12(4):487–498, 2000.

[169] S. L. Lauritzen. *Graphical Models*. Oxford, UK, Clarendon, 1996.

[170] T. Y. Leong. Representing context-sensitive knowledge in a network formalism: A preliminary report. In *Proc. of the 8th Conference on Uncertainty in Artificial Intelligence (UAI-1992)*, pages 166–173. Morgan Kaufmann, 1992.

[171] J. S. Liu. *Monte Carlo Strategies in Scientific Computing*. Springer-Verlag, 2004.

[172] C. Lu, T. Van Gestel, J.A.K. Suykens, S. Van Huffel, I. Vergote, and D.Timmerman. Classification of ovarian tumor using bayesian least squares support vector machines. In *Proc. of the 9th Conference on Artificial Intelligence in Medicine in Europe (AIME 2003)*, pages 219–228, 2003.

[173] C. Lu, T. Van Gestel, J.A.K. Suykens, S. Van Huffel, I. Vergote, and D.Timmerman. Preoperative prediction of malignancy of ovarium tumor using least squares support vector machines. *Artificial Intelligence in Medicine*, 28(3):281–306, 2003.

[174] P. Lucas. Restricted Bayesian network structure learning. In H. Blockeel and M. Denecker, editors, *Proc. of 14th Belgian-Dutch Conference on Artificial Intelligence (BNAIC'02)*, pages 211–218, 2002.

[175] D. J. C. MacKay. Probable networks and plausible predictions - a review of practical Bayesian methods for supervised neural networks. *Neural Computation*, pages 469–505, 1996.

[176] D. J. C. Mackay. *Learning in graphical models*, chapter Introduction to Monte Carlo Methods. MIT Press, Cambridge, MA, 1999.

[177] D. Madigan and R. Almond. Test selection strategies for belief networks. StatSci Research Report 20., 1993.

[178] D. Madigan, S. A. Andersson, M. Perlman, and C. T. Volinsky. Bayesian model averaging and model selection for Markov equivalence classes of acyclic digraphs. *Comm.Statist. Theory Methods*, 25:2493–2520, 1996.

[179] D. Madigan, J. Gavrin, and A. E. Raftery. Eliciting prior information to enhance the predictive performance of bayesian graphical models. *Comm.Statist. Theory Methods*, 24:22712292, 1995.

[180] D. Madigan and J.York. Bayesian graphical models for discrete data. *Internat. Statist. Rev.*, 63:215–232, 1995.

[181] D. Madigan and A. E. Raftery. Model selection and accounting for model uncertainty in graphical models using occams window. *J. Amer. Statist. Assoc.*, 89:15351546, 1994.

[182] S. Mahoney and K. B. Laskey. Network engineering for complex belief networks. In *Proc. 12th Conf. on Uncertainty in Artificial Intelligence*, pages 389–396, 1996.

[183] D. Malakoff. Bayes offers a 'new' way to make sense of numbers. *Science*, 286:1460–1464, 1999.

[184] S. Mani and G. F. Cooper. Causal discovery from medical textual data. In *AMIA Annual Symposium*, pages 542–6, 2000.

[185] C. D. Manning and H. Schutze. *Foundations of statistical natural language processing*. MIT Press, Cambridge, MA, 2002.

[186] C. Meek. Causal inference and causal explanation with background knowledge. In Philippe Besnard, Steve Hanks, Philippe Besnard, and Steve Hanks, editors, *Proc. of the 11th Conf. on Uncertainty in Artificial Intelligence (UAI-1995)*, pages 403–410. Morgan Kaufmann, 1995.

[187] B. Milch, B. Marthi, and S. Russell. Blog: Relational modeling with unknown objects. In *Proc. 21th Int. Conf. on Machine Learning (ICML)*, pages 157–165, 2004.

[188] A. Millinghoffer, G. Hullám, and P. Antal. Statisztikai adat- és sz ovegelemzés bayes-hálókkal: a valószínuségektol a f uggetlenségi és oksági viszonyokig. *Híradástechnika*, 60:40–49, 2005. (in Hungarian).

[189] A. Millinghoffer, G. Hullám, and P. Antal. On inferring the most probable sentences in bayesian logic. In *Workshop notes on Intelligent Data Analysis in bioMedicine And Pharmacology (IDAMAP-2007), 11th Conference on Artificial Intelligence in Medicine (AIME 07)*, pages 13–18, 2007.

[190] M. F. Moller. A scaled conjugate gradient algorithm for fast supervised learning. *Neural Networks*, 6:525–533, 1993.

[191] Y. Moreau, P. Antal, G. Fannes, and B. De Moor. Probabilistic graphical models for computational biomedicine. *Methods of Information in Medicine*, 42(4):161–168, 2002.

[192] S. Muggleton. Stochastic logic programs. In *Proceedings of the 5th International Workshop on Inductive Logic Programming*, page 29. Department of Computer Science, Katholieke Universiteit Leuven, 1995.

[193] P. Myllymaki. *Mapping Bayesian Networks to Stochastic Neural Networks: A Foundation for Hybrid Bayesian-Neural systems*. Ph.D. dissertation, University of Helsinki, No. A-1995-1, 1995.

[194] R. M. Neal. *Bayesian Learning for Neural Networks*. Springer, Berlin, 1996.

[195] R. M. Neal. Transferring prior information between models using imaginary data. Technical Report No. 0108, Dept. of Statistics, University of Toronto, 2001.

[196] M. Neil, N. E. Fenton, and L. Nielsen. Building large-scale Bayesian networks. *The Knowledge Engineering Review*, 15(3):257–284, 2000.

[197] S. J. Nelson, T. Powell, and B. L. Huhmpreys. The unified medical language system (umls) project, 2001. http://www.nlm.nih.gov.

[198] P. Niyogi, T. Poggio, and F. Girosi. Incorporating prior information in machine learning by creating virtual examples. *Proceedings of the IEEE*, 86(11):2196–2209, 1998.

[199] J. Forster P. Dellaportas and I. Ntzoufras. On bayesian model and variable selection using mcmc. *Statistics and Computing*, 12:27–36, 2002.

[200] J. Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, San Francisco, CA, 1988.

[201] J. Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669710, 1995.

[202] J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2000.

[203] D. Pe'er, A. Regev, G. Elidan, and N. Friedman. Inferring subnetworks from perturbed expression profiles. *Bioinformatics, Proceedings of ISMB 2001*, 17(Suppl. 1):215–224, 2001.

[204] E. Segal D. Peer, A. Regev, D. Koller, and N. Friedman. Learning module networks. In *Proc. of the 19th Conf. on Uncertainty in Artificial Intelligence (UAI-2003)*, pages 525–534. Morgan Kaufmann, 2003.

[205] A. Pfeffer, D. Koller, B. Milch, and K. T. Takusagawa. Spook: A system for probabilistic object-oriented knowledge representation. In *Proc. of the 15th Conf. on Uncertainty in Artificial Intelligence(UAI-1999)*, pages 541–550. Morgan Kaufmann, 1999.

[206] K. R. Popper. *The Logic of Scientific Discovery*. Hutchinson, London, 1959.

[207] M. Pradhan, G. Provan, B. Middleton, and M. Henrion. Knowledge engineering for large belief networks. In *Proc. of the 10th Conf. on Uncertainty in Artificial Intelligence*, pages 484–490, 1994.

[208] D. Proux, F. Rechenmann, and L. Julliard. A pragmatic information extraction strategy for gathering data on genetic interactions. In *Proc. of the 8th International Conference on Intelligent Systems for Molecular Biology (ISMB'2000), LaJolla, California*, pages 279–285, 2000.

[209] G. M. Provan. Tradeoffs in knowledge-based construction of probabilistic models. *IEEE Transactions on Systems Man and Cybernetics*, 24(11):1580–1592, 1994.

[210] G. M. Provan and M. Singh. Learning bayesian networks using feature selection. In *Proc. of the Fifth International Workshop on Artificial Intelligence and Statistics*, pages 450–456, 1995.

[211] S. Renooij, S. Renooij, and L. van der Gaag. Context-specific sign-propagation in qualitative probabilistic networks. In *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*, pages 667–672, 2000.

[212] A. Rényi. *Probability Theory*. Akadémiai Kiad, Budapest, 1970.

[213] C. P. Robert. *Markov Chain Monte Carlo in Practice*, chapter phy. Chapman & Hall, London, 1996.

[214] C. P. Robert. *The Bayesian Choice*. Springer-Verlag, New York, 2001.

[215] I. Ruczinski, C. Kooperberg, and M. LeBlanc. Logic regression. *Journal of Computational and Graphical Statistics*, 12(3):475–511, 2003.

[216] I. Ruczinski, C. Kooperberg, and M. LeBlanc. Exploring interactions in genomic data using logic regression, 2004. ISMB'04 poster.

[217] S. M. Rüger and A. Ossen. Clustering in weight space of feedforward nets. In C. von der Malsburg, editor, *ICANN 96, Lecture Notes in Computer Science*, pages 83–88. Springer-Verlag, Berlin, 1996.

[218] S. Russel and P. Norvig. *Artificial Intelligence*. Prentice Hall, 2001.

[219] H. Schriger and G. F. Cooper. Understanding your data: Graphical and non-traditional approaches to data analysis. http://www.saem.org/download/02schriger.pdf, 2002.

[220] E. Segal, M. Schapira, A. Regev, D. Peer, D. Botstein, D. Koller, and N. Friedman. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nature Genetics*, 34(2):166–76, 2003.

[221] H. Shatkay, S. Edwards, and M. Boguski. Information retrieval meets gene analysis. *IEEE Intelligent Systems*, 17(2):45–53, 2002.

[222] J. W. Shavlik. An overview of research at Wisconsin on knowledge-based neural networks. In *Proc. of the Int. Conf. on Neural Networks*, pages 65–69, 1996.

[223] E. S. Shtatland, E. Cain, and M. B. Barton. The perils of stepwise logistic regression, 2001. SAS SUGI proceedings, poster.

[224] C. Silverstein, S. Brin, R. Motwani, and J. D. Ullman. Scalable techniques for mining causal structures. *Data Mining and Knowledge Discovery*, 4(2/3):163–192, 2000.

[225] R. Sowmya and R. J. Mooney. Theory refinement for Bayesian networks with hidden variables. In *Proc. 15th International Conf. on Machine Learning*, pages 454–462. Morgan Kaufmann, San Francisco, CA, 1998.

[226] D. J. Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society, Series B*, 50(2):157–224, 1988.

[227] D. J. Spiegelhalter, A. Dawid, S. Lauritzen, and R. Cowell. Bayesian analysis in expert systems. *Statistical Science*, 8(3):219–283, 1993.

[228] D. J. Spiegelhalter and S. L. Lauritzen. Sequential updating of conditional probabilities on directed acyclic graphical structures. *Networks*, 20(.):579–605, 1990.

[229] P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. MIT Press, 2001.

[230] B. Stapley and G. Benoit. Biobibliometrics: Information retrieval and visualization from co-occurrences of gene names in medline asbtracts. In *Proc. of Pacific Symposium on Biocomputing (PSB00)*, volume 5, pages 529–540, 2000.

[231] D. Subramanian, R. Greiner, and J. Pearl. The relevance of relevance. *Artificial Intelligence*, 97:l–5, 1997.

[232] D. R. Swanson and N. R. Smalheiser. An interactive system for finding complementary literatures: a stimulus to scientific discovery. *Artificial Intelligence*, 91:183–203, 1997.

[233] A. Tekay and P. Jouppila. Validity of pulsatility and resistance indices in classification of adnexal tumors with transvaginal color Doppler ultrasound. *Ultrasound Obstetrics Gynecology*, 2:338–344, 1992.

[234] P. Thagard. Explaining disease: Correlations, causes, and mechanisms. *Minds and Machines*, 8:61–78, 1998.

[235] L. Tierney. *Markov Chain Monte Carlo in Practice*, chapter Introduction to general state-space Markov chain theory, pages 59–75. Chapman & Hall, 1996.

[236] J. B. Tilbury, P. W. J. Van Eetvelt, J. M. Garibaldi, J. S. H. Curnow, and E. C. Ifeachor. Receiver operating characeristic analysis for intelligent medical systems - a new approach for finding confidence intervals. *IEEE Transactions on Biomedical Engineering*, 47(7):952–963, 2000.

[237] D. Timmerman. *Ultrasonography in the assessment of ovarian and tamoxifen-associated endometrial pathology.* Ph.D. dissertation, Leuven University Press, D/1997/1869/70, 1997.

[238] D. Timmerman. Artificial neural network models for the pre-operative discrimination between malignant and benign adnexal masses. *Ultrasound Obstetrics Gynecology*, 13:17–25, 1999.

[239] D. Timmerman, A.C. Testa, T. Bourne, E. Ferrazzi, L. Ameye, M.L. Konstantinovic, B. Van Calster, W.P. Collins, I. Vergote, S. Van Huffel, and L. Valentin. Logistic regression model to distinguish between the benign and malignant adnexal mass before surgery: a multicenter study by the international ovarian tumor analysis group. *J Clin Oncol*, 23(34):8794–801, 2005.

[240] D. Timmerman, L. Valentin, T. H. Bourne, W. P. Collins, H. Verrelst, and I. Vergote. Terms, definitions and measurements to describe the sonographic features of adnexal tumors: a consensus opinion from the international ovarian tumor analysis (IOTA) group. *Ultrasound Obstetrics Gynecology*, 16(5):500–505, 2000.

[241] G. Towell and J. Shavlik. Knowledge-based artificial neural networks. *Artificial Intelligence*, 70:119–165, 1994.

[242] I. Tsamardinos and C. Aliferis. Towards principled feature selection: Relevancy,filters, and wrappers. In *Proc. of the Artificial Intelligence and Statistics*, pages 334–342, 2003.

[243] I. Tsamardinos, C.F. Aliferis, and A. Statnikov. Algorithms for large-scale local causal discovery and feature selection in the presence of limited sample or large causal neighbourhoods. In *The 16th International FLAIRS Conference*, 2003.

[244] National Cancer Institute (US). SEER cancer data, 1998.

[245] L. Valentin. Gray scale sonography, subjective evaluation of the color Doppler image and measurement of blood flow velocity for distinguishing benign and malignant tumors of suspected adnexal origin. *European Journal of Obstetrics & Gynecology and the Reproductive Biology*, 72:63–72, 1997.

[246] L. Valiant. A theory of the learnable. *Comm. of the ACM*, 27:1134–1142, 1984.

[247] V. N. Vapnik. *The Nature of Statistical Learning Theory.* Springer, Berlin, 1995.

[248] T. Verma and J. Pearl. *Equivalence and synthesis of causal models*, volume 6, pages 255–68. Elsevier, 1990.

[249] H. Verrelst, Y. Moreau, J. Vandewalle, and D. Timmerman. Use of a multi-layer perceptron to predict malignancy in ovarian tumors. In *Proceedings of the 1997 Conference of Advances in Neural Information Processing Systems (NIPS 97)*, pages 978–984, 1997.

[250] H. Verrelst, J. Vandewalle, B. De Moor, and D. Timmerman. Bayesian input selection for neural network classifiers. In *Proc. of the Third International Conference on Neural Networks and Expert Systems in Medicine and Healthcare (NNESMED'98)*, pages 125–132, 1998.

[251] G. Vita'nyi and K. Li. *An introduction to Kolmogorov complexity and it applicationes*. Springer-Verlag, New York, 1990.

[252] J. Vomlel. Noisy-or classifiers. In *Proceedings of the 6th Workshop on Uncertainty Processing (WUPES 2003)*, pages 291–302, 2003.

[253] S. Waterhouse, D. MacKay, and T. Robinson. Bayesian methods for mixtures of experts. In *Neural Inf. Proc. Systems*, volume 8, pages 351–357, 1995.

[254] M. P. Wellman. Fundamental concepts of qualitative probabilistic networks. *Artificial Intelligence*, 44:257–303, 1990.

[255] M. P. Wellman, J. S. Breese, and R. P. Goldman. From knowledge bases to decision models. *The Knowledge Engineering Review*, 7(1):35–53, 1992.

[256] A. S. Whittemore, G. Gong, and J. Itnyre. Prevalence and contribution of BRCA1 mutations in breast cancer and ovarian cancer. *American Journal of Human Genetics*, 60:496–504, 1997.

[257] J. Williamson. *Foundations for Bayesian networks*, pages 11–140. Kluwer Academic Publ., 2001.

[258] X. Wu, P. Lucas, S. Kerr, and R. Dijkhuizen. Learning bayesian-network topologies in realistic medical domains. In *In Medical Data Analysis: Second International Symposium, ISMDA*, pages 302–308. Springer-Verlag, Berlin, 2001.

[259] Lei Yu and Huan Liu. Efficient feature selection via analysis of relevance and redundancy. *Journal of Machine Learning Research*, 5:1205–1224, 2004.

Name: Péter Antal Date of birth: 07-08-1971

1995: M.Sc. in Computer Science (Informatics Engineer), 1995, Faculty of Electrical Engineering and Informatics, Technical University of Budapest

1995-1998: PH.D. studies on the informatics Ph.D. programme of Faculty of Electrical Engineering and Informatics at the Department of Measurement and Information Systems, Technical University of Budapest.

1998 - 1999: International scholar, Department of Electrical Engineering, Katholieke Universiteit Leuven.

2000-2002: Studying for Ph.D. at the Department of Electrical Engineering, Katholieke Universiteit Leuven, Belgium. Research topic: Combination of prior domain knowledge and data in statistical learning methods.