KU Leuven
Biomedical Sciences Group
Faculty of Medicine
Department of Public Health and Primary Care
Leuven Biostatistics and Statistical Bioinformatics

KU LEUVEN

DOCTORAL SCHOOL
BIOMEDICAL SCIENCES

Bayesian model selection for longitudinal random-effects models

Samuel Oludare ARIYO

Jury:

Promoter:	Prof. Dr. Emmanuel Lesaffre	
Co-promoter:	Prof. Dr. Geert Verbeke	
Chair:	Prof. Dr. Geert Molenberghs	
Secretary:	Prof. Dr. Iven Van Mechelen	Dissertation presented
Jury members:	Prof. Dr. Iven Van Mechelen	in partial fulfilment
	Prof. Dr. An Carbonez	of the requirements
	Prof. Dr. Dimitris Rizopoulos	for the degree of Doctor
	Prof. Dr. Christel Faes	in Biomedical Sciences(Biostatistics)

June 30, 2020

KU Leuven
Biomedical Sciences Group
Faculty of Medicine
Department of Public Health and Primary Care
Leuven Biostatistics and Statistical Bioinformatics

KU LEUVEN
DOCTORAL SCHOOL
BIOMEDICAL SCIENCES

Bayesian model selection for longitudinal random-effects models

Samuel Oludare ARIYO

Jury:

Promoter:	Prof. Dr. Emmanuel Lesaffre
Co-promoter:	Prof. Dr. Geert Verbeke
Chair:	Prof. Dr. Geert Molenberghs
Secretary:	Prof. Dr. Iven Van Mechelen
Jury members:	Prof. Dr. Iven Van Mechelen
	Prof. Dr. An Carbonez
	Prof. Dr. Dimitris Rizopoulos
	Prof. Dr. Christel Faes

Dissertation presented
in partial fulfilment
of the requirements
for the degree of Doctor
in Biomedical Sciences(Biostatistics)

June 30, 2020

Thesis Submitted in Partial Fulfilment of the Requirements for the Degree of
Doctor in Biomedical Sciences (Biostatistics)

Department of Public Health and Primary Care
Group Biomedical Sciences
Faculty of Medicine
KU Leuven

Bayesian model selection for longitudinal random-effects models

Oludare Samuel Ariyo

Supervisors:

Emmanuel Lesaffre
Geert Verbeke

Academic year 2019–2020

Acknowledgements

First and foremost, I give to the Lord the glory He deserves (Psalm 96:8) and I say what shall I return to the Lord for all his goodness to me? (Psalm 116:12). I return all the glory to the Almighty God, who has guided me through the good and bad times of my life. *E mi 'ban'egberunahon, fun 'yin Olugbala, Ogo Olorun Oba mi, Isegun Ore Re*. (O for a thousand tongues to sing My dear redeemer's praise! The Glory of my God and King, the triumph of His grace.)

I would like to express my unreserved gratitude to my promoter Emeritus Prof Emmanuel Lesaffre for his support, inspiration, and guidance throughout these years—It has been a great privilege working with you. Furthermore, I would like to thank my co-promotor, Prof.Greet Verbeke for the understanding, encouragement and advice given towards this research. My KU Leuven jury members who have continued to support me since the very first year have been a great influence on the direction of my research. Prof. Ivan Van Mechelen and Prof. An Carbonez who contributed to the broadening of this research with their insightful comments. I would also like to thank the external jury members Prof. Dimitris Rizopoulos and Prof. Christel Faes, who kindly accepted our request to review and evaluate this thesis. I am immensely grateful to Prof. Geert Molenberghs for graciously agreeing to chair the PhD defence.

Special thanks to the sponsor of my research, the Tertiary Education Trust Fund (TETFund) of the Federal Republic of Nigeria and the management of the Federal University of Agriculture, Abeokuta Nigeria. Most notably, I would like to give my thanks to the past and current Vice-chancellors: Prof. Olusola Oyewole, Prof. Ololade Enikuomehin, Prof. Felix Salako and other principal officers whose assistance was beneficial. Permit me to mention a few of my senior colleagues as well who encouraged me to come to Belgium and assisted me with my move (not in any particular order): Prof. C.O. Eromosele, Prof. O. E. Asiribo, Prof. O. Isah, Prof. O.J Adeniran and Dr P. O. Akintokun. To my colleagues at the Department of Statistics, the Federal University of Agriculture Abeokuta, and most especially the past and current department heads: Dr Agwuegbo, S.O.N, Dr Dawodu, G.A, Dr Apantaku, F.S and Dr Olayiwola, O.M. I say thank you for your understanding and support.

There have also been wonderful people in I-BioStat (past and present) whose expertise saved me a great deal of time and taught me to solve my problems more efficiently. Some of them were my "mini-teachers." I would like to thank these colleagues, in particular, Drs. Cristian Villegas, Naranjo Lizbeth, Adrian Quintero, Ann Ivanova, Vahid Nassiri, and Anikó Lovik for their support during my first year of this PhD. I acknowledge the gift of friendship package in the form of Dung Trang Tran, Jeroen Sichen, Akalu Bantata, Bijit Roy, Isaac Fwemba, Martial Luyts, Daniel Olusoji, Olajumoke Owokotomo, Maarten Coemans and Jinfang Sun. Thanks especially to Kristen Verhaegen and Dr. Kris Bogaerts for their administrative and technical support. I would also like not to forget my special friend Johanna Muñoz.

I thank the great people that I met during these years I have spent living in Leuven, most notably the Deeper Life Bible Church family—you are family indeed. My relationship with you has been very inspiring and a blessing. As well, the past and current Deeper life Campus Fellowship Leuven brethren, most especially my dear brother and friend Jonathan Adelakun.

I will always be grateful for the love, sacrifice and prayers of my dear mother, Florence Ariyo, who supported by educational purses from elementary schools especially for the past twenty-eight years when my father passed away. I love you, "Iya Ibeji". To my siblings: Oluwaseun, Olusola and Oluseyi thank you for your love.

To my children, Jeremiah and Elizabeth, I am sorry for leaving you with your grandparents during these years. Most of the time, you have to talk to your Daddy through video calls and technical issues sometimes prevent you from expressing yourselves. The PhD is now over! Special thanks to Dr. and Mrs Osungade for their support, especially looking after Jeremiah and Elizabeth while myself and my wife were away for our PhD in Belgium. I appreciate the support of my brother Segun Osungade who always help me with special assignments in Nigeria.

Finally, I would like to thank the person dear to my heart, my wife, Esther. I owe her a special thanks for her patience, encouragement, sacrifice and love. Esther, there are no words to express my appreciation for your unconditional love during the first ten years of our marriage, most especially during this PhD years.

Thank you very much. Everyone!

Oludare Ariyo
Leuven, 30 June 2020

List of Papers

The contents of this thesis is based on the following research articles:

Chapter 3: Ariyo,O., Quintero, A., Muñoz, J., Verbeke, G., & Lesaffre, E. (2019). Bayesian model selection in linear mixed models for longitudinal data. *Journal of Applied Statistics* 47(5), 890-913.

Chapter 4: Ariyo, O., Lesaffre, E., Verbeke, G., & Quintero, A. (2019).Model selection for Bayesian linear mixed models with longitudinal data: Sensitivity to the choice of priors. *Communications in Statistics: Simulation and Computation* 0 (0), 1-25.

Chapter 5: Ariyo, O., Lesaffre, E., Verbeke, G., & Quintero, A.(2020). Bayesian model selection for longitudinal count data. Submitted to *Communications in Statistical Applications and Methods*

Chapter 6: Ariyo, O., Lesaffre, E., Verbeke, G., Huisman,M., Rijnhart,J., Heymans, M., & Twisk, J. (2020). Bayesian model selection for multilevel mediation models. Submitted to *Journal of Statistical Computation and Simulation*

The author has also been involved in the following research article:

Dawodu, G., Akintunde, A., & **Ariyo, O**. (2017). Stochastic prediction of monthly inflation rates through Kalma filtering. Journal of Natural Sciences, Engineering and Technology, 16(2),11-21.

Table of Contents

Summary

The popularity of the mixed model can be explained by its flexibility in modelling complex hierarchical data. Since the introduction of the basic mixed models—the linear mixed model (LMM), generalized linear mixed model (GLMM) and non-linear mixed model (NLMM)—a great variety of extensions have been suggested. For longitudinal studies, the mixed model consists of a fixed part expressing the effect of covariates on the mean evolution over time and a random part representing the variation of the individual curves around the mean curve. Selecting the appropriate fixed and random effect parts is an essential modelling exercise when choosing the best mixed model. As such, we considered three Bayesian model selection criteria: the Pseudo-Bayes factor (PSBF), Deviance Information Criteria (DIC) and Watanabe-Akaike Information Criterion (WAIC). Since there is little agreement in the statistical literature on the most suitable choice among these model selection criteria, it a useful exercise to evaluate their performance.

The above criteria can be classified according to the way the random effects are handled: given the random effects (conditional likelihood) resulting in conditional model selection criteria; or integrating out the random effects (marginal likelihood), resulting in the marginal model selection criteria. Although the conditional criteria have been criticized in the statistical literature, applied Bayesians still use the conditional criteria since they are built-in standard Bayesian software.

The marginal model selection criteria have been advocated, but less attention has been paid to the LMM. Besides, it is surprising that almost no mention is made of the suboptimality of the conditional criteria. Therefore, we compared via extensive simulations the performance of the conditional and marginal versions of DIC, WAIC and PSBF on the classical LMM, to create also awareness of the problem. We also considered extensions of the classical LMM. The results confirm the superiority of the marginal criteria in all settings and scenarios of our simulations. To promote the usage of the marginal criteria among practitioners, we provided an R function capable of computing the three marginal and condi-

tional criteria with a minimal computational cost.

We further determined the effect of vague priors on the performance of the above Bayesian model selection criteria. More specifically, we evaluated the impact of vague priors for the covariance matrix of the random effects on selecting the correct LMM. For two or more random effects, we considered five different specifications of the Inverse-Wishart (IW) prior, four different separation techniques, one hierarchical prior and a joint prior. We showed that the choice of a vague prior has a relatively low to minimal impact on the marginal criteria, in contrast to the conditional criteria. But if the conditional criteria are to be used, then it is best to use a separation or a joint prior.

Furthermore, we extended our exploration to check the performance of the model selection criteria to GLMMs for longitudinal count data. Since a GLMM does not have a closed-form likelihood like LMM, we, therefore, searched for efficient ways to compute the marginal criteria. Our computational procedure is based on the replication sampling approach in combination with importance sampling. We also provided an R function that computes the marginal model selection criteria for longitudinal Poisson models and their extensions.

Finally, we explored the performance of the Bayesian model selection criteria in the context of multilevel mediation models, namely for the 1-1-1 mediation model. We again demonstrated the superiority of the marginal selection criteria over their conditional counterparts in the mediated longitudinal settings through simulation. We also showed that the above R function for LMM needs only little modification for multilevel mediation models.

In total, we used four longitudinal clinical data sets and one longitudinal non-clinical data set to further illustrate the performance of the marginal and conditional criteria. These are Jimma infant survival study, Potthoff & Roy dental study and Nigeria indigenous datasets for both LMM and priors sensitivities papers. We used Epilepsy seizure data set for GLMM count data paper. For the mediation paper; we made use of data from the LASA (Longitudinal Aging Study of the Amsterdam) study. We made use of the Bayesian software package JAGS in combination with R. The relevant code can found on https://ibiostat.be/online-resources/bayesian and in the supplementary material of the papers.

Samenvatting

Het gemengd model is populair omdat het op een flexibele manier complexe hierarchische data kan modelleren. Vanaf de introductie van de klassieke gemenged modellen – het lineair gemengd model (LMM), het veralgemeend lineair gemengd model (GLMM) en het niet-lineaire gemengd model (NLMM) – werden een grote varieteit van uitbreidingen voorgesteld. Toegepast op longitudinale data, bestaat het gemengd model uit een populatiegedeelte, genaamd 'fixed effecten', wat het effect van covariaten op de gemiddelde evolutie over de tijd weergeeft en een stochastisch gedeelte dat de variatie van de individuele kurven rond de gemiddelde kurve beschrijft, genaamd 'random effecten'. De keuze van de geschikte fixed en random effecten is een essentieel onderdeel van het statistisch modelleren van het gemengd model. In die context hebben we drie Bayesiaanse model selectiecriteria uitgekozen, namelijk de Pseudo Bayes Factor (PSBF), het Deviance Informatie Criterium (DIC) en het Watanabe-Akaike Informatie Criterium (WAIC). Aangezien er onenigheid is in de statistische literatuur omtrent het meest geschikte criterium, leek het ons nuttig na te gaan welke van deze criteria meest performant is voor het selecteren van een gemengd model.

De bovenvermeldde criteria kunnen onderverdeeld worden aan de wijze waarop de random effecten behandeld worden. Namelijk men kan de random effecten in het statistisch model houden, dan spreekt men van een conditionele likelihood, om het met een Engelse term te benoemen. Deze resulteren dan in conditionele model selectiecriteria. Een andere manier is de random effecten uit de likelihood uit te integreren, men krijgt dan een marginale likelihood en dit resulteert dan in marginale model selectiecriteria. Hoewel dat de conditionele criteria bekritiseerd werden in de statistische literatuur, worden zij nog steeds het meest gebruikt door de toegepaste Bayesiaanse statistici vooral omdat deze criteria in de Bayesiaanse software ingebouwd werden.

De marginale model selectiecriteria werden in de statistische literatuur aanbevolen, maar hierbij is er tot nu weinig aandacht gegeven aan het lineair gemengd

model. Bovendien is het verrassend te zien dat praktisch nooit de suboptimaliteit van de conditionele criteria wordt vermeld. We hebben daarom de performantie van de conditionele en marginale versies van DIC, WAIC en PSBF vergeleken met behulp van extensieve simulaties op het klassieke LMM, gedeeltelijk om ook de statistici te wijzen dat men best afstapt van de conditionele criteria. Verder hebben we ook extensies bekeken van het klassieke LMM.

Onze resultaten bevestigen de eerder bekomen superioriteit van de marginal criteria in alle simulatiescenarios. Om het gebruik van de marginale criteria te promoten hebben we verder een R functie ontwikkeld die de drie conditionele en marginale criteria zonder veel extra inspanning berekent. Voorts hebben we het effect van vage prior verdelingen nagegaan op de effectiviteit van bovenvermeldde Bayesiaanse model selectiecriteria. Meer specifiek hebben we het impact van vage priors op de covariantie matrix van de random effecten geevalueerd op het selecteren van het correcte LMM. Wanneer 2 or meer random effecten in het LMM aanwezig waren, hebben we vijf verschillende specificaties van de Inverse Wishart (IW) prior bekeken, en verder vier verschillende separatie technieken, een hierarchische prior en een gezamenlijke prior. We hebben dan aangetoond dat de keuze van een vage prior weinig of geen impact had op de marginale criteria, in tegenstelling tot de conditionele criteria. Maar als we noodgedwonden een conditioneel criterium moeten gebruiken, dan bleek uit de simulatieresultaten dat een separatie of een gezamenlijke prior te verkiezen is.

In een volgende stap, hebben we ons onderzoek naar de performantie van de model selectiecriteria uitgebreid naar GLMMs for longitudinale data. Omdat een GLMM geen analytische resultaten toelaat zoals bij het LMM, hebben we naar efficiente methodes gezocht voor de berekening van de marginale criteria. Onze computationele procedure is gebaseerd op de replicatiemethode in combinatie met importance sampling. We hebben dan ook weer een R programma geschreven dat voor deze modellen de conditionele en marginale criteria berekent. We hebben tot slot ook de performantie bekeken van de Bayesiaanse model selectiecriteria voor een multilevel mediatie model, namelijk voor het 1-1-1 mediatie model. En weer heb-ben we de superioriteit van de marginale criteria aangetoond opnieuw gebruik makende van extensieve simulatiestudies. Een kleine aanpassing van het bestaande R programma voor een LMM was nodig om de criteria te berekenen voor dit mediatie model. We maak-ten in deze thesis gebruik van vier longitudinale data sets, namelijk: Jimma Infant Survival study, Potthoff & Roy en Nigeria indigenous Chicken data sets voor klassieke LMM en prior gevoeligheid artikel. We gebruikten epilepsie-aanvalsgegevensset voor GLMM-telgegevens artikel. Voor het mediatie artikel maakten we gebruik van data verzameld in de LASA (Longitudinal Aging Study of the Amsterdam) studie. Tot slot, alle progammas were ge-schreven in R in combinatie met JAGS. De code van de programma's kan men vinden op de website: https://ibiostat.be/online-resources/bayesian en in supplementair materiaal van de artikels.

Abbreviations

Here, we give a list of the most often used abbreviations in this thesis.

AED	Anti-epileptic drug
AIC	Akaike's Information Criterion
BF	Bayes factor
BGR	Brooks-Gelman-Rubin diagnostic
BIC	Bayesian information criterion
CPOs	Conditional predictive ordinates
cDIC	Conditional deviance information criterion
cPSBF	Conditional pseudo-Bayes factor
cWAIC	Conditional Watanabe-Akaike information criterion
DA	Data augmentation
DIC	Deviance information criterion
ELPPD	Expected log-pointwise predictive density
GLMM	Generalized Linear Mixed Model
GMVLG	Generalized multivariate log-gamma prior
HIW prior	Hierarchical inverse Wishart prior
JAGS	Just another Gibbs sampler
IW	Inverse-Wishart
LASA	Longitudinal Ageing Study of Amsterdam
LMM	Linear Mixed Model
LPPD	Log pointwise predictive distribution
LPML	log-pseudo likelihood
MCMC	Markov chain Monte Carlo methods
mDIC	Marginalized deviance information criterion
mPSBF	Marginalized pseudo-Bayes factor
mWAIC	Marginalized Watanabe-Akaike information criterion
mLPML	Marginalized log-pseudo likelihood
MLM	Multilevel-mediation

MSEM	Multilevel structural equation model
NIC	Nigerian Indigenous Chicken
NLMM	Non-Linear Mixed Model
PGMM	Poisson-gamma mixed effects model
PLMM	Poisson-lognormal mixed effects model
PSBF	Pseudo-Bayes factor
SNLMM	Skew-normal Linear Mixed Model
STLMM	Skew-t Linear Mixed Model
ZINB	Zero-inflated negative binomial mixed effects model

List of Figures

List of Tables

Chapter 1

General Introduction

Medical and epidemiological research applications often involve longitudinal studies which employ repeated measures to monitor individuals over time. The response of interest in a longitudinal study may be continuous, count, binary, categorical or a combination of two or more outcomes. Regardless of the nature of the response, mixed models are one of the most popular tools for analysing longitudinal data.

Mixed models combine fixed effects with random effects. Since the introduction of the basic mixed models: the linear mixed model (LMM), generalised linear mixed model (GLMM), and non-linear mixed model (NLMM), a great variety of extensions have been suggested. For example, the classic mixed model with the normality assumption for either the random and/or measurement error parts of the model has been extended to encompass a variety of distributions such as skew-normal and skew-t-distributions (Sahu et al., 2003; Arellano-Valle and Genton, 2005; Arellano-Valle et al., 2007; Arellano-Valle and Genton, 2010; Huang and Dagne, 2012). Others have utilised a joint model where the repeated outcome is analysed simultaneously with the time-to-dropout (Ivanova et al., 2016; Rizopoulos, 2012, 2011), with the repeated outcome being analysed jointly with the missing data process (Molenberghs et al., 2004; Ivanova et al., 2017). Mixed models have also been suggested for dealing with the presence of over-dispersion or under-dispersion in the longitudinal outcome (Ivanova et al., 2014; Aregay et al., 2013; Molenberghs et al., 2010, 2007). These extensions were made under both the frequentist and the Bayesian paradigms. Given a data set, a variety of mixed-effect models can be considered—hence, there is a need for model selection.

Three popular Bayesian model selection criteria often allow comparing Bayesian mixed models in practice. One of the earlier selection criteria suggested in the literature is Bayes' factor (Kass and Raftery, 1995). This criterion would be a logical choice; however, it is severely influenced by choice of prior as well as having severe computational issues. Some alternatives have been proposed (see for example De Santis and Spezzaferri, 1997) and the most popular is the Pseudo-Bayes factor (PSBF). The PSBF first updates the (improper) prior to a proper posterior and then computes the Bayes' factor using the generated posterior as the prior. This alternative criterion, although relatively easy to compute, is not yet commonly used. Indeed, the popular choice is the Deviance Information Criterion (DIC), although it has been demonstrated that DIC might be problematic in practice (see Spiegelhalter et al., 2014, and the discussions therein). For instance, when the effective degrees of freedom are estimated to be negative, DIC cannot be used. A promising alternative to DIC is the Widely Applicable Information Criteria (WAIC) (Watanabe, 2013), which has been singled out as a worthy successor to DIC (Spiegelhalter et al., 2014). This criterion estimates the predictive accuracy of the model. It includes a bias correction for the data to be able to be used twice, i.e., to estimate the model and evaluate its accuracy.

The computation of the criteria mentioned above (i.e. DIC, WAIC and PSBF) is based on the models' likelihood, which can either be the conditional likelihood or the marginal likelihood. The distinction between these likelihoods has been made previously in the literature. In their original paper, Spiegelhalter et al. (2002) discussed the conditional/marginal DIC where they refer to the issue as "model focus" (see also Celeux et al., 2006; Millar, 2009). For the conditional likelihood, the model is based on fixed and random effects, while for marginal likelihood, the random effects have been integrated out. Gelman et al. (2014) point out that there is a choice to be made when defining the likelihood for Bayesian information criteria between the conditional and marginal versions since both versions give different results under the same or similar settings. However, most researchers are often not aware of this distinction and only rely on the default software (Merkle et al., 2018).

The choice between the conditional and the marginal model selection criteria is based on whether we wish to measure the predictive ability of the model on new units from either the same or a new cluster respectively. That is to say, if we want our model to make predictions for new units from the same clusters as in our original data, then the conditional likelihood is appropriate, and we must then deal with conditional DIC (cDIC), conditional PSBF (cPSBF), and conditional WAIC (cWAIC). Conversely, the marginal likelihood will be an appropriate choice if we measure the predictive ability of our model for new units from clusters not in the original data, and in which case we would then deal with marginal DIC (mDIC), marginal PSBF (mPSFB), and marginal WAIC (mWAIC). The choice should be motivated by the research question (Vaida and Blanchard, 2005).

Studies have advocated the use of marginal criteria (Chan and Grant, 2016a; Quintero and Lesaffre, 2018; Merkle et al., 2018; Millar, 2018; Li et al., 2016) for different models; however, less attention has been given to the LMM. Therefore, our study opts to compare the performance of the conditional and marginal versions of DIC, WAIC, and PSBF in LMM but also to its extensions to skew-normal and skew-t distributions for either (or both) random effects and measurement error. To promote the usage of marginal criteria among the practitioners, we provided an easy-to-use R function which computes both the marginal and conditional criteria with only a minimal computational cost. Building upon previous results, we evaluated the effect of vague priors on the Bayesian model selection criteria. More specifically, we assessed the impact of vague priors for the covariance matrix of the random effects on the selection of the correct LMM. For two or more random effects, we considered five different specifications of the conjugate Inverse-Wishart (IW) prior, four different separation techniques (separation priors), and one joint prior.

Furthermore, we conducted the research utilising a GLMM for longitudinal count data. Since a GLMM does not have a closed-form likelihood, we considered sampling procedures (replication and importance sampling methods) to compute the marginal criteria for a GLMM. We also provided an R function that renders these computations flexible. Additionally, we illustrated the performance of the conditional and marginal criteria in multilevel mediation models. These topics have been introduced and explained using four longitudinal medical data sets alongside one non-medical longitudinal data set.

1.1 Medical Introduction

The ability of any patient to respond to treatments (vaccine/drug) depends on several factors. For example, a Covid-19 patient's recovery depends on multiple factors like age, underlying health issues, the efficacy of their immune system, and the viral load of the patient. Regardless of the health care received by the patient, these patient-specific factors are essential in the recovery process. As such, an individual-specific random effects model is essential in modelling data sets generated in these situations.

In trying to understand the nature of a disease and developing a new drug/ vaccine, researchers often employ a longitudinal study by repeatedly monitoring an individual over time. Still on the Covid-19 example, the research may track some vital signs of the patient at intervals over a specific period (i.e. the virus incubation period). These vital signs could take different forms, such as, count (patients' hearth beat), continues (body temperature) and binary (the presence/absence of a sign). Mixed model effects (individual-specific mixed models) are the most popular tools for analyzing this repeated data, see Section 1.0 for more information.

When the model under consideration contains the random effects (individual-specific effects), the definition of likelihood is not straightforward. The questions are, what type of likelihood should be used? Should the random effects be counted as a parameter or not? Indeed, the answer to these questions depends on the focus of the research. With the Covid-19 example, the question is, should we ignore or consider the effect of an individual patient in forecasting the efficacy of the new drug? The answer indeed depends on the researcher and the aim of the research.

As efforts are ongoing to find a drug/vaccine for this pandemic, the distinction between the choice of the conditional and marginal criteria should be made. When the focus is to measure the effect on a new drug on patients in a treatment centre (hospital), then the conditional criteria may be used. Conversely, if the focus is upon the overall treatment effect of the new drug/vaccine leaving aside the individual-specific factors (which is often the case in most medical research), we argued that the use of conditional criteria should be discouraged. Preferably the marginal criteria should be used. In the marginal criteria, the main objective is to measure the efficacy of the new drug/vaccine on patients in a new or similar treatment centre (hospital) rather than only within the treatment centres in the data set. As such, we are motivated to make a clear distinction between the conditional and the marginal criteria. We have also shown the superiority of the marginal criteria in most settings and scenarios. We are aware that some researchers may be reluctant in the use of conditional criteria. Consequently, we provided R functions that compute both the conditional and marginal criteria for researchers' judgement.

1.2 The linear mixed model with extensions

Linear mixed models (LMMs) are popular to analyze repeated measurements with a Gaussian response. These models represent an extension of the linear regression framework, in which model coefficients for predictive variables are permitted to vary randomly between individuals or groups. Since its introduction by Laird and Ware (1982), several extensions have been proposed in the literature, and thorough reviews of this topic are given by Verbeke and Molenberghs (2000) and

McCulloch et al. (2008) among others.
Let assume that there are n subjects (independent) and each is repeatedly measured m_i times. Let Y_{ij} be the observation of the j^{th} response of the i^{th} subject for $j = 1, \dots, m_i$ and $i = 1, \dots, n$ and let $x_{ij} = (1, x_{ij,1}, x_{ij,2}, \dots, x_{ij,p})^T$ and $z_{i,j} = (1, z_{ij,1}, z_{ij,2}, \dots, z_{ij,q})^T$ be the corresponding $(p+1) \times 1$ and $(q+1) \times 1$ predictor vectors. Let us assume that the data satisfy the LMMs
$Y_{ij} = \beta_0 + \beta_1 x_{ij,1} + \cdots + \beta_p x_{ij,p} + b_{i0} + b_{i,1} z_{ij,1} + \cdots + b_{i,q} z_{ij,q} + \epsilon_{i,j}$, where $\boldsymbol{\beta} = (\beta_0, \dots, \beta_p)^T$ is a $(p+1)-$dimensional vector of unknown fixed-effects coefficients, $\mathbf{b}_i = (b_{i,0}, \dots, b_{i,q})^T$ is a $(q+1)-$dimensional random effect belonging to the i^{th} subject with $z_{ij,1}$ instead of $x_{ij,1}$: $\mathbf{b}_i \sim N_{(q+1)}(\mathbf{0}, \mathbf{D})$ and $\epsilon_{i,j}$ are the random errors for $j = 1, \dots, m_i$, $i = 1, \dots, n$. Alternatively, the classical LMM can be expressed in matrix notation as

$$\mathbf{Y}_i = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \mathbf{b}_i + \boldsymbol{\epsilon}_i, \tag{1.1}$$

where $\mathbf{Y}_i$ is an m_i-dimensional response vector of measurements for the i-th subject, $(i = 1, \dots, n)$. $\mathbf{X}_i$ and $\mathbf{Z}_i$ are $m_i \times (p+1)$ and $m_i \times q$-dimensional covariate matrices, respectively, and $\boldsymbol{\beta}$ is a $(p+1)$-dimensional vector of fixed effects. The residual component vector $\boldsymbol{\epsilon}_i$ is distributed as $N_{m_i}(\mathbf{0}, \boldsymbol{\Sigma}_i)$, where $\boldsymbol{\Sigma}_i$ is an $m_i \times m_i$ positive-definite covariance matrix. It is usually assumed that $\boldsymbol{\Sigma}_i = \sigma_\epsilon^2 \mathbf{I}_{m_i}$, where $\mathbf{I}_{m_i}$ denotes the identity matrix of dimension m_i. The $(q+1)$-dimensional random-effects vectors $\mathbf{b}_i$ are assumed independent from the residuals and distributed as $N_q(\mathbf{0}, \mathbf{D})$, where $\mathbf{D}$ is a $(q+1) \times (q+1)$ positive-definite covariance matrix.
A great number of extensions have been made to the LMMs. For example, in longitudinal studies on HIV patients, both viral load response and CD4 cell counts are highly skewed, it might be more realistic to assume a multivariate skew-normal for both random effects and measurement error (Huang and Dagne, 2011, 2012). Hence, Arellano-Valle and Genton (2005) extended LMMs by assuming that both the random effects and measurement error follow a skew-normal linear mixed model (SNLMM) (see also Sahu et al., 2003; Arellano-Valle et al., 2007; Huang and Dagne, 2012; Lachos et al., 2013, 2010). Arellano-Valle et al. (2007) also suggested a skew-t distribution whereby the t-distribution replaces the classical Gaussian distribution.

1.3 Prior sensitivity

An essential step in the Bayesian modelling is the choice of priors for the model parameters. When prior knowledge is available, an informative prior is typically chosen while in the absence of prior information, a vague prior must be used. In many situations, vague prior distributions are selected with the intention that they should have little or no impact on inference. However, a naive choice of a vague prior may have unwillingly a significant effect on posterior inference especially when the data set is small (Lambert et al., 2005).

For a mixed model (e.g. LMM), a vague prior is mostly taken for the fixed effects and the variance components. In this thesis, we have taken vague normal priors for fixed effects while will focus on the effects of different vague priors for the covariance matrix of the random effects. We consider the univariate case of a random intercept and the multivariate case of several random effects. For the univariate case, various vague priors have been suggested for the level-2 variance (variance of random intercept) of the Gaussian hierarchical model. This model is a special case of the LMM with only a random intercept. In that case, (1.1) can be written as

$$\boldsymbol{Y}_i \sim N(X_i\boldsymbol{\beta} + \mathbf{1}_{m_i}b_i, \sigma_\epsilon^2) \quad i = 1, \ldots, n. \tag{1.2}$$

with $\mathbf{1}_{m_i}$ is a $m_i \times 1$ vector of ones, and where the random intercept $b_i \sim N(0, \sigma_b^2)$. The improper prior $p(\sigma_b^2) \propto 1/\sigma_b^2$, suggested by Jeffreys for the simple case of $N(\mu, \sigma^2)$ yields an improper posterior for model (1.2) if applied to σ_b^2. This was recognised a long time ago, see e.g Lesaffre and Lawson (2012). Hence, we considered different vague priors for σ_b and evaluate their impact on the performance of the conditional and marginal criteria in identifying the appropriate model.

Specifying an appropriate prior for a covariance matrix has been the topic of intensive research in the last two decades. The IW distribution gives the mathematically convenient prior for a covariance matrix. This prior is often used in Bayesian modelling for an unknown covariance matrix due to its conditional conjugacy and its implementation in most of the Bayesian statistical software. Still, there are practical problems with this prior. In Chapter 4.5.2, we review the issues involved with the IW prior, then we discuss some generalizations of this prior to improve convergence properties and its ability to represent (absence of) prior knowledge in an appropriate manner.

To address the problems with the IW prior, different separation strategies (Barnard et al., 2000) have been proposed including: Cholesky decomposition (Wei and Higgins, 2013), spherical decomposition (Pinheiro and Bates, 1996), Fisher's z-transformation (Daniels and Kass, 1999), partial prior (Barnard et al., 2000) among others. Often, priors for error variance and variance-covariance matrix of the random effects are independently modelled. However, it has been shown by Demirhan and Kalaylioglu (2015) and by Kalaylioglu and Demirhan (2017) that a joint prior for these variance terms is more appropriate. Hence, we compare the performance of IW, separation priors and a joint prior on the performance of the conditional and marginal criteria in selection true data-generating models.

1.4 The generalized linear mixed model

Generalized linear mixed models (GLMMs) have become popular in analyzing longitudinal data with a non-Gaussian longitudinal response. The fitting of GLMMs to a longitudinal data structure, has been the subject of a great deal of research over the past decades see also Breslow and Clayton (1993), Engel and Keen (1994), Wolfinger and O'Connell (1993) ,Verbeke and Molenberghs (2000), Dean and Nielsen (2007), Tuerlinckx et al. (2006) and Molenberghs et al. (2002). Let Y_{ij} be the j^{th} outcome for subject $i = 1, \ldots, n$, $j = 1, \ldots, m_i$ where m_i measurements into vector $\mathbf{Y}_i$, given $(q+1)$-dimensional random effects $\mathbf{b}_i$, then y_{ij}'s are independent with model

$$\begin{aligned} f_i(y_{ij}|\mathbf{b}_i, \boldsymbol{\beta}, \phi) &= \exp\left\{\phi^{-1}[y_{ij}\lambda_{ij} - \zeta(\lambda_{ij})] + c(y_{ij}, \phi)\right\}, \\ \eta[\zeta^{'}(\lambda_{ij})] &= \eta[E(y_{ij}|\mathbf{b}_i, \boldsymbol{\lambda})] = \mathbf{x}_{ij}^T\boldsymbol{\beta} + \mathbf{z}_{ij}^T\mathbf{b}_i \end{aligned} \tag{1.3}$$

where $\mathbf{x}_{ij}$ and $\mathbf{z}_{ij}$ are $(p+1)$-dimension and $(q+1)$-dimension covariate, vectors $\boldsymbol{\beta}$ is a $(p+1)$-dimension vector of unknown fixed effects parameters, ϕ is a scale (overdispersion) parameter and $\eta(\cdot)$ is a known link function. Assume that the density of the random effects $\mathbf{b}_i$ is given as $f(\mathbf{b}_i|\mathbf{D})$, then the marginal likelihood function is:

$$L(\boldsymbol{\Theta}, \mathbf{D}) = \prod_{i=1}^{n} \int \prod_{j=1}^{m_i} f_{ij}(y_{ij}|\boldsymbol{\Theta}, \mathbf{b}_i) f(\mathbf{b}_i|\mathbf{D}) d\mathbf{b}_i. \tag{1.4}$$

Here, $\boldsymbol{\Theta}$ represents all model parameters for $\mathbf{Y}_i$ given $\mathbf{b}_i$. The integral (1.4) most often cannot be computed analytically. Hence the need for an alternative approach to compute the marginal likelihood.

1.5 Mediation Analysis

Mediation analysis enables researchers to investigate how exposure affects an outcome in the presence of a mediator. The concept of mediation has broad applications in both biomedical and social science research. For example, in the Longitudinal Ageing Study of Amsterdam (LASA), processing speed has been investigated as a mediator of the relationship between age and cognitive function (Robitaile et al., 2013).

Mediation analysis may present complex-mediated relationship where a mediator influences the outcome. For example, in a study of the relationship between hepatitis C progression, alcohol consumption may affect adherence to antiretroviral therapy, which, in turn, affects the viral load. However, heavy alcohol consumption may affect liver function. If the study aims to evaluate the total effect of an exposure (main independent variable) on an outcome (dependent variable), a linear mixed model could be fit to the data (Blood et al., 2010; Blood and Cheng, 2011).

Let i be the index of a first-level unit and j be index of a second level unit. A two-level mediation model can be expressed as

$$\begin{aligned} M_{ij} &= \beta_{1j} + \alpha_j X_{ij} + e_{M_{ij}} \\ Y_{ij} &= \beta_{2j} + \beta_j M_{ij} + \tau_j^{'} X_{ij} + e_{Y_{ij}}, \end{aligned} \tag{1.5}$$

and level 2 is given as

$$\begin{aligned} \beta_{1j} &= \beta_3 + u_{1j} \\ a_j &= \alpha + u_{2j} \\ \beta_{2j} &= \beta_4 + u_{3j} \\ \beta_j &= \beta + u_{4j} \\ \tau_j^{'} &= \tau^{'} + u_{5j}, \end{aligned}$$

where $e_{M_{ij}}$ and $e_{Y_{ij}}$ are level 1 error terms for M and Y respectively; the parameters β_{1j} and β_{2j} are random intercepts, and α_j, β_j and $\tau_j^{'}$ are random slopes. The parameters β_2 and β_3 are population (or average) effects. For multilevel modelling, the first-level residuals $e_{M_{ij}}$ and $e_{Y_{ij}}$ are assumed to be independent and follow normal distribution, that is $e_{M_{ij}} \sim N(0, \sigma^2_{e_{M_{ij}}})$ and $e_{Y_{ij}} \sim N(0, \sigma^2_{e_{Y_{ij}}})$ and the second-level residuals $\mathbf{u}_j = (u_{1j}, u_{2j}, u_{3j}, u_{4j}, u_{5j})^T$ follow a multivariate normal distribution $\mathbf{u}_j \sim N(\mathbf{0}, \mathbf{D})$ where $\mathbf{D}$ is 5×5 covariance matrix.

In multilevel mediation, the average indirect effects in the population are often of primary interest. Yuan and MacKinnon (2009) gave the average indirect effects (applies to models with only random slopes) formula to be

$$ab = E(\alpha_j \beta_j) = \alpha\beta + \sigma_{\alpha_j \beta_j}, \tag{1.6}$$

where $\sigma_{\alpha_j \beta_j}$ denotes the covariance between α_j and β_j.

MacKinnon (2008) and Kenny et al. (2003) also showed that the total effect in a fully random, lower mediated multilevel model is

$$c = \tau^{'} + \alpha\beta + \sigma_{\alpha_j\beta_j} \tag{1.7}$$

and the relative average indirect effect can be expressed as

$$ab/c = \frac{\alpha\beta + \sigma_{\alpha_j\beta_j}}{\tau' + \alpha\beta + \sigma_{\alpha_j\beta_j}}. \tag{1.8}$$

The above statistic is often called the proportion mediated in the literature see Ditlevsen et al. (2005); Ananth (2019). Its main setback is that it cannot be used when the direct and indirect effects have a different sign (i.e. the mediation model is inconsistent). This setback implies that the proportion mediated model can exceed 1 and can be negative. This renders its interpretation meaningless.

1.6 Motivating data sets

1.6.1 Jimma Infant Survival Study

In this data set, 495 newborns from the Jimma town in Ethiopia born in the period: 11-9-1992 until 10-9-1993 were examined from birth and approximately every two months for their height, weight and arm circumference. This epidemiological cohort study aims at relating demographic and other variables on the child's first year's weight and height evolution. The measured covariates are the gender of the child, age of the mother in the first year, cultural practices applied to the child, etc. There are many missing values due to children that die or get lost-to-follow-up. Besides, the visits are only approximately taken at equal time lags. Furthermore, when weight is regressed on height and other covariates, one has to take into account that height is measured with error. For a reference, see Lesaffre et al. (1999).

1.6.2 Potthoff & Roy dental data set

We considered a well known balanced longitudinal dental study analyzed by Potthoff and Roy (1964). Dental measurements on eleven girls and sixteen boys at four different ages (t1 = 8, t2 =10, t3 = 12, and t4 = 14) were taken. Each measurement is the distance, in millimetres, from the centre of the pituitary to pteryo-maxillary fissure. Other authors have considered this data in the literature (Baey et al., 2017; Al-Rawwash and Pourahmadi, 2013).

1.6.3 A Clinical Trial in Epileptic Patients

Epilepsy data set is a public data set of 89 patients who have epilepsy. The data is from a randomized, double-blind, parallel-group, multi-centre study for the comparison of placebo with a new anti-epileptic drug (AED), in combination with one or two other AED's (Faught et al., 1996). Patients were randomized after a 12-week stabilization period for the use of AED's, and during which the number of seizures was counted. After that run-in period, 45 patients were assigned to the placebo group, 44 to the new treatment. Patients were measured weekly and followed (double-blind) during 16 weeks; thereafter, they entered a long-term open-extension study. Some patients were followed for up to 27 weeks.
The experiment compares the number of seizures experienced by the patients among the groups. Booth et al. (2003) used this data set as an illustrating example when modelling longitudinal counts data with overdispersion and correlation. Other authors Aregay et al. (2013); Rakhmawati et al. (2016); Iddi and Doku-Amponsah (2016) have used this data and for more elaborate discussions, refer to Faught et al. (1996) and Molenberghs et al. (2007). The outcome of interest is the number of epileptic seizures experienced during the last week, i.e., since the last time the outcome was measured.

1.6.4 The Longitudinal Aging Study Amsterdam (LASA)

LASA data set is a longitudinal study that started in 1992 to determine the predictors and consequences of ageing. It consists of a national representative sample of 3,805 older adults stratified according to the year of birth, sex and geographical locations of people born between 1908 and 1937. Data relating to physical, emotional, cognitive and social functioning in late life, the connections between these components, the changes in these components that occur with time within and between respondents, and the consequences of these changes were measured every 11 months by trained medical and psychology interviewers.

Attrition in LASA can be attributed for the most significant part to mortality, and a lesser extent to refusal, or other reasons (Huisman et al., 2011). For a further description of the data set, see Huisman et al. (2011).

1.6.5 The Nigerian Indigenous Chicken (NIC) data set

The Nigerian Indigenous Chicken data set describes the longitudinal evolution of the body weight (BW) of chickens of different breeds raised in a Federal University of Agriculture, Abeokuta Nigeria experimental farm. Four hundred and sixteen chickens were measured every week from hatching up to 20 weeks. While some chickens live up to the completion of the study, some die before 20 weeks, hence creating an unbalanced data set. The study aimed to evaluate the growth of different chicken breeds. Here we considered two classes of progenies. Two hundred and seventy chickens were produced from the same parent stock (pure breed), while 146 chickens have different parents (cross breed). The rationale for the study and the experimental design can be found in Adeleke et al. (2011).

1.7 Thesis contribution

The main aim of this thesis is to provide practical guidelines for practitioners, especially for clinical researchers, on the use of appropriate Bayesian model selection criteria for the analysis of longitudinal data. More specifically, this thesis evaluates the operational performance of the conditional and the marginal versions of DIC, PSBF and WAIC in selecting the correct data-generating model. We were triggered to conduct this research based on the fact that (1) researchers are reluctant to use the marginal criteria despite their advantages shown in the literature and (2) the criteria have not yet been extensively checked for their performance in commonly used LMMs. To this end, we explored: (i) the properties of Bayesian model selection criteria in LMMs with some extension, and (ii) measured the impact of the prior on both versions of the criteria, (iii) evaluated the performance of these criteria when the likelihood is not analytically available in GLMMs and (iv) illustrated the usefulness of these exercises in multilevel mediation models. Five longitudinal data sets were used as motivating data sets for the simulation setups. For this purpose, we addressed the following specific objectives:

1. Study the performance of the conditional and marginal versions of the Deviance Information Criteria, Pseudo-Bayes factor and Widely Applicable Information Criteria in identifying the correct data-generating model in LMM.

2. Extend objective (1) in the skew-normal LMM (SNLMM) and the skew-t LMM (STLMM) for either or both random effects and measurement error.

3. Evaluate the impact of vague priors for the variance of the univariate and multivariate random effects on the performance of the marginal and conditional model selection criteria.

4. Extend the objectives above to a generalized linear mixed model especially with over-dispersed count data.

5. Evaluate the performance of two sampling techniques (replication and importance sampling methods) in computing the marginal criteria.

6. Evaluate the selection criteria on multilevel mediation models.

1.8 Overview of the subsequent chapters

This thesis is structure as follows: Chapter 2 presents a general introduction to Bayesian inference discussing the central methodology used in the thesis. The four subsequent chapters correspond to published or submitted manuscripts that address the specific objectives of this thesis. In the last chapter, some general conclusions are presented as well as it exhibits a discussion of the limitations of our research and it discusses a possible future research topics.

Chapter 2

Bayesian Inference

2.1 Bayesian Inference

Statistical inference is the ensemble of activities of using data to extract information on the probability distribution that generated the data. Statistical inference is made via at least two mainstream paradigms: (i) the frequentist approach and (ii) the Bayesian approach.

The frequentist approach to inference is based on exploring the repeated sampling properties of test statistics given an unknown but fixed (actual) value of the model parameters $\mathbf{\Theta}$. Conversely, Bayesian inference considers parameters $\mathbf{\Theta}$ as stochastic, and the probability statements are conditional on the observed value of $\mathbf{y}$. The probability statements about these parameters ($\mathbf{\Theta}$ given $\mathbf{y}$) are derived from the model providing a joint probability distribution for $\mathbf{\Theta}$ and $\mathbf{y}$, given by

$$p(\mathbf{\Theta}, \mathbf{y}) = p(\mathbf{\Theta})p(\mathbf{y}|\mathbf{\Theta}),$$

where $p(\mathbf{\Theta})$ is referred to as the prior distribution and $p(\mathbf{y}|\mathbf{\Theta}) = L(\mathbf{\Theta}|\mathbf{y})$ the sampling distribution evaluated in $\mathbf{\Theta}$, also called the likelihood function evaluated in $\mathbf{\Theta}$ given $\mathbf{y}$. Together with $p(\mathbf{\Theta}, \mathbf{y}) = p(\mathbf{\Theta}|y)p(\mathbf{y})$ yields Bayes' rule and the posterior density:

$$p(\mathbf{\Theta}|\mathbf{y}) = \frac{p(\mathbf{\Theta})p(\mathbf{y}|\mathbf{\Theta})}{p(\mathbf{y})}, \tag{2.1}$$

where $p(\mathbf{y}) = \int p(\mathbf{\Theta})p(\mathbf{y}|\mathbf{\Theta})d\mathbf{\Theta}$ in the case of continuous $\mathbf{\Theta}$. The denominator $p(\mathbf{y})$ does not depend on $\mathbf{\Theta}$ and is regarded as a normalizing constant. The prior distribution $p(\mathbf{\Theta})$ can be described based on past evidence, previous studies or expert opinion. When such information is not available, the use of vague prior

may be an option.
The posterior distribution (2.1) combines the prior belief $p(\mathbf{\Theta})$ with the likelihood $L(\mathbf{\Theta}|\mathbf{y})$ to carry out the inference. The main interest in Bayesian inference is to summarise $p(\mathbf{\Theta}|\mathbf{y})$ and draw inference (conclusions) about the model parameters. To achieve this, summary statistics such as the posterior mean and the standard deviation are obtained. The credible interval (also called Bayesian confidence interval) contains the most plausible values of the model parameters. There are two commonly used credible interval: the equal tail credible interval and the highest posterior density credible interval. The posterior distribution (2.1) parameter has a closed-form in case of a conjugate prior distribution of the model. For example, when counts follow a Poisson distribution and the mean rate (λ) is assigned a Gamma prior, the posterior distribution of (λ) has a Gamma density. This ensures a more straightforward computation of the posterior summary measures. Conversely, when the analytic determination of the posterior is not possible, i.e. when $p(\mathbf{y}) = \int p(\mathbf{y}|\mathbf{\Theta})p(\mathbf{\Theta})d\mathbf{\Theta}$ is not analytically available, then one needs numerical techniques such as numerical integration and Monte Carlo sampling to compute the posterior probability distribution.

2.2 Prior Distribution

The prior distribution of $\mathbf{\Theta}$ represents the distribution of possible parameter values imperative of the observed data. There are mainly two kinds of prior distributions: non-informative prior and informative prior distributions. With non-informative priors the data mainly determine the posterior distribution. Some examples of non-informative priors include: (i) locally uniform prior (Hartigan et al., 1996) (ii) Jeffreys invariant priors (Jeffreys, 1961), (iii) reference priors (Berchialla et al., 2009), (iv) probability matching priors (Mukerjee and Ghosh, 1997), other examples are described in Kass and Raftery (1995).
It is important to note that no prior is completely non-informative (see for example Lambert et al., 2005; Ariyo et al., 2019a) as the name is a misnomer. When prior knowledge is available, either from previously analyzed data or from an expert's opinion, an informative prior distribution may be used. An informative prior dominates the likelihood, that is when using an informative prior, data are allowed to "speak" for itself and that has an impact on the posterior distribution. Gelman and Hill (2007) and Gelman et al. (2004b) advocated the use of informative priors for final analysis after initial model development using non-informative priors.

2.2.1 Conjugate priors

Raiffa and Schlaifer (1961) proposed the formal definition and concept of conjugate prior distributions. The formal definition is given as follows:
If $\mathcal{F}$ is a class of sampling distribution $p(\mathbf{y}|\boldsymbol{\Theta})$, and $\mathcal{P}$ is a class of prior distribution for $\boldsymbol{\Theta}$, then the class $\mathcal{P}$ is conjugate for $\mathcal{F}$ if

$$p(\boldsymbol{\Theta} \mid \mathbf{y}) \in \mathcal{P} \quad \text{for all} \quad p(\cdot \mid \Theta) \in \mathcal{P} \quad \text{and} \quad p(\cdot) \in \mathcal{P}.$$

The conjugate prior has computational advantages and easily interpretable as additional data.

2.3 Markov chain Monte Carlo sampling

In many applications, the analytical calculation of the posterior distribution (and its summary measures) is often not feasible due to the difficulty in determining the integration constant $p(\mathbf{y})$. In this case, numerical integration methods may be used but they fall short for dimensions above 10. An often more attractive approach is to sample from the posterior. The most popular class of sampling algorithms is called Markov chain Monte Carlo (MCMC) methods.
Markov Chain Monte Carlo sampling generates a sequence $\boldsymbol{\Theta}^1, \boldsymbol{\Theta}^2, \ldots, \boldsymbol{\Theta}^K$ satisfying the Markov property, i.e. $p(\boldsymbol{\Theta}^{k+1}|\boldsymbol{\Theta}^k, \boldsymbol{\Theta}^{k-1}, \ldots, \mathbf{y}) = p(\boldsymbol{\Theta}^{k+1}|\boldsymbol{\Theta}^k, \mathbf{y})$. This property means that given the current $\boldsymbol{\Theta}^k$, the value $\boldsymbol{\Theta}^{k+1}$ is independent from the previous elements in the sequence.
MCMC techniques have been used in the simulation of stochastic systems. They are also useful for estimating integrals in situations where a closed-form analytical solution cannot be attained. This is often the case in Bayesian methods, where the problem is multidimensional, and integration is not possible. Integrals are also to determine the marginal distribution of a random variable or taking its expected value. The two most important MCMC procedures are (a) the Gibbs sampler and (b) the Metropolis(-Hastings) algorithm.

2.3.1 The Gibbs sampler

Gibbs sampling (Geman and Geman, 1984) has become very popular in statistics when Gelfand and Smith (1990) showed its applicability in the Bayesian framework. It is a special case of the Metropolis-Hastings algorithm and based on the property that a multivariate distribution is uniquely determined by its conditional distributions. Gibbs sampling is often easy to implement. Given a starting position $\boldsymbol{\Theta}^0 = (\Theta_1^0, \Theta_2^0, \ldots, \Theta_d^0)$, the method samples from the full conditional distributions at iteration $(k+1)$ with the following d steps:

- Sample $\Theta_1^{(k+1)}$ from $p(\Theta_1 | \Theta_2^k, \Theta_3^k, \ldots, \Theta_d^k, \mathbf{y})$
- Sample $\Theta_2^{(k+1)}$ from $p(\Theta_2 | \Theta_1^{(k+1)}, \Theta_3^k, \ldots, \Theta_d^k, \mathbf{y})$
  $\vdots$
- Sample $\Theta_d^{(k+1)}$ from $p(\Theta_d | \Theta_1^{(k+1)}, \Theta_2^{(k+1)}, \ldots, \Theta_{(d-1)}^{(k+1)}, \mathbf{y})$.

It has been shown (see for details Lesaffre and Lawson, 2012) that under mild regularity conditions and from iteration k_0, the Gibbs sampler generates a sequence $\boldsymbol{\Theta}^{k_0+1}, \boldsymbol{\Theta}^{k_0+2}, \ldots, \boldsymbol{\Theta}^{k_0+K}$ which can be regarded as observations from the posterior distribution $p(\boldsymbol{\Theta}|\mathbf{y})$.

2.3.2 Metropolis-Hastings (MH) algorithm

The Metropolis-Hastings (MH) algorithm is an MCMC method for obtaining a sequence of random samples from a probability distribution which is difficult to obtained directly. Its main difference from the Gibbs sampler is that it does not require the full conditionals. The idea is to obtain the posterior distributions through a proposal density q which is straightforward to sample from. Let suppose that when exploring $p(\boldsymbol{\Theta}|\mathbf{y})$, the Markov chain is at position $\boldsymbol{\Theta}^k$ at the kth iteration. Then, we sample $\tilde{\boldsymbol{\Theta}}$ from a proposal density $q(\tilde{\boldsymbol{\Theta}}|\boldsymbol{\Theta}^k)$ and compute the acceptance probability as

$$\alpha(\boldsymbol{\Theta}^k, \tilde{\boldsymbol{\Theta}}) = \text{min}\left(\frac{p(\tilde{\boldsymbol{\Theta}}|\mathbf{y})q(\boldsymbol{\Theta}^k|\tilde{\boldsymbol{\Theta}})}{p(\boldsymbol{\Theta}^k|\mathbf{y})q(\tilde{\boldsymbol{\Theta}}|\boldsymbol{\Theta}^k)}, 1\right),$$

with the next value $\boldsymbol{\Theta}^{(k+1)}$ equal to either the sampled $\tilde{\boldsymbol{\Theta}}$ with probability $\alpha(\boldsymbol{\Theta}^k, \tilde{\boldsymbol{\Theta}})$ or equal to $\boldsymbol{\Theta}^k$ with probability $1 - \alpha(\boldsymbol{\Theta}^k, \tilde{\boldsymbol{\Theta}})$. Therefore, $\boldsymbol{\Theta}^{(k+1)}$

is likely to stay at the same position as $\mathbf{\Theta}^k$ when $\tilde{\mathbf{\Theta}}$, presents a low value for $p(\tilde{\mathbf{\Theta}}|\mathbf{y})$. The multivariate normal distribution with equal mean to the current position $\mathbf{\Theta}^k$ and a well-chosen covariance matrix $\mathbf{\Sigma}$ is a popular choice, but also a multivariate t-distribution is often used. For a good performance of the MH algorithm, the proposal covariance matrix $\mathbf{\Sigma}$ should be chosen such that the acceptance probability is $\approx 45\%$ for $d = 1$, and $\approx 24\%$ for $d > 1$.

2.4 Convergence of the MCMC Algorithms

The main idea of the MCMC algorithm is to create a Markov chain that has a stationary distribution equal to the posterior distribution. However, it not straightforward to determine whether convergence has been realized in practice. Convergence is evaluated using dedicated tests often by checking stationarity of the chain (e.g.,Geweke test) or by checking mixing of the multiple chains,(Brooks-Gelman-Rubin diagnostic).
To access the stationary, a simple graphical tool is the traceplot: a time series plot with the iteration number on the x-axis and the sampled value on the y-axis. When convergence is concluded, the traceplot appears as a horizontal strip where individual movements across iterations are hardly discernible.

2.5 Software

The development of MCMC algorithms has popularized the use of Bayesian statistics. Nowadays, many applied statisticians make use of BUGS software, after its release in 1989. The realization of different Bayesian R packages after that has facilitated the coding of algorithms to the sample from the posterior distributions.
Just Another Gibbs Sampler (JAGS); an alternative Bayesian package based on (an extended version of) the BUGS language has become popular in recent years. The motive for JAGS development can be summarised as: (i) to have a cross-platform engine for the BUGS language, (ii) to be extensible, allowing users to write their functions, distributions and samplers and (iii) to be a platform for experimentation with ideas in Bayesian modelling.
All the models considered in this thesis were implemented in JAGS running inter-

actively from R using the rjags package. The relevant code can be found online as supplementary material of the published papers and in the Appendix of this thesis.

Chapter 3

Bayesian model selection in linear mixed models for longitudinal data

This chapter has been published as:

Ariyo, O., Quintero, A., Muñoz, J., Verbeke, G., and Lesaffre, E. (2019). **Bayesian model selection in linear mixed models for longitudinal data**. *Journal of Applied Statistics* 47 (5), 890-913.

Abstract

Linear mixed models (LMMs) are popular to analyze repeated measurements with a Gaussian response. For longitudinal studies, the LMMs consist of a fixed part expressing the effect of covariates on the mean evolution in time and a random part expressing the variation of the individual curves around the mean curve. Selecting the appropriate fixed and random effect parts is an important modeling exercise. In a Bayesian framework, there is little agreement on the appropriate selection criteria. This paper compares the performance of the deviance information criterion (DIC), the pseudo-Bayes factor and the widely applicable information criterion (WAIC) in LMMs, with an extension to LMMs with skew-normal distributions. We focus on the comparison between the conditional criteria (given random effects) versus the marginal criteria (averaged over random effects).In spite of theoretical arguments, there is not much enthusiasm among applied statisticians to make use of the marginal criteria. We show in an extensive simulation study that the three marginal criteria are superior in choosing the appropriate longitudinal model. In addition, the marginal criteria selected most appropriate model for growth curves of Nigerian chicken. A self-written R function can be combined with standard Bayesian software packages to obtain the marginal selection criteria.

3.1 Introduction

Longitudinal studies have become central in a great variety of research areas. The longitudinal study design is the only study design that allows to relate determinants measured at the start of the study to changes in the subjects' condition over time. Numerous books have recently appeared on longitudinal study designs, e.g, Anderson (2018); Funatogawa (2017); McArdle and Nesselroade (2014); Gayle and Lambert (2018); Hoffman (2015). When the response is Gaussian, linear mixed-effects models (LMMs) are one of the most popular tools to analyze longitudinal data. Since its introduction by Laird and Ware (1982), the LMM has been applied in a great variety of research areas and extended in many ways, e.g. to generalized linear mixed-effects models and non-linear mixed-effects models. Its popularity has much to do with its ability to describe both the impact of covariates on the mean longitudinal evolution as well as how individual profiles differ over time from the mean curve. The impact on the mean longitudinal curve is evaluated by their regression coefficients, which are referred to as the fixed effects. The subject-specific profiles are expressed as latent variables, called random effects. In this way, the LMM fits subject-specific profiles and accounts for correlation among responses from the same subject. Another important feature is that the LMM allows for unbalanced data, i.e., when the number and timing of the observations per subject differ between subjects.

The LMM parameters may be estimated using a frequentist approach. The properties of the estimated model parameters are then based on (restricted) maximum likelihood theory (Verbeke and Molenberghs, 2000). Alternatively, one could use the Bayesian framework. In the Bayesian approach prior information on the model parameters is combined with information coming from the data. Using Bayes' theorem, an updated idea on the model parameters is obtained from the posterior distribution. The posterior distribution provides all information that is needed, and hence there is no need to refer to asymptotic normality properties for inference on the model parameters. This is especially useful in longitudinal studies with a small number of subjects and when the data are unbalanced (Raudenbush and Bryk, 2002). Since most posterior distributions are analytically intractable, they need to be determined in a numerical way. Most popular numerical techniques are based on sampling from the posterior distribution. The Markov chain Monte Carlo (MCMC) techniques provide an important class of such methods. In this paper we focus on fitting Bayesian LMMs to longitudinal data and compare the performance of different selection criteria. While in a Bayesian model all parameters are stochastic (and thus random), we will (as many others) still use the standard terminology of fixed and random effects.

A variety of LMMs can be fitted to the data at hand depending on several aspects such as: (i) the covariates that are considered in the fixed part of the model, (ii) the random effects structure to be included, e.g., random intercepts and/or random slopes, and (iii) possible transformations of the response. When considering several LMMs, it is important to select a parsimonious model that fits adequately the current and also future data. Unfortunately, there is little agreement on what criterion to choose for Bayesian model selection.

One of the first model selection criteria suggested in the literature is the Bayes factor (Kass and Raftery, 1995), which is defined as the ratio of the marginal likelihood of two competing models. Although this criterion has a natural interpretation, its computation remains difficult in practice and the results can be sensitive to the choice of the prior distributions, presenting difficulties especially with improper priors. Gelfand and Dey (1994) proposed the pseudo-Bayes factor (PSBF), which updates the (improper) prior to a proper posterior and calculates the Bayes factor using the generated posterior as prior. This alternative criterion, although relatively easy to compute, is not yet commonly used.

The most popular Bayesian model selection criterion is the deviance information criterion (DIC) (Spiegelhalter et al., 2002). The DIC is similar to the AIC often used in the frequentist framework, i.e., it represents a trade-off between model fit and model complexity. The aim of DIC is to estimate the predictive ability of the fitted model to future samples from the same population. More recently, the widely applicable information criterion (WAIC) was proposed (Watanabe, 2010) for model selection in the Bayesian framework. This criterion estimates the predictive accuracy of the model and includes a bias correction for using the data twice, i.e., to estimate the model and to evaluate model's accuracy. It has also been argued that WAIC is a more fully Bayesian approach (compared to DIC) and is suitable for singular models, such as LMMs for longitudinal data when the random effects are considered as parameters in the model (Gelman et al., 2014). Apart from the above three model selection criteria, a wide variety of (Bayesian) statistical approaches have been suggested to select the most appropriate LMM. While it is not the aim of this paper to give a comprehensive overview, the reader should be aware of the large number of alternative approaches proposed in the literature. For instance, a popular alternative approach is to use Bayesian variable selection techniques, often based on the SSVS approach of George and McCulloch (1993). Examples of this approach can be found in Chen and Dunson (2003), Cai and Dunson (2006) and Gong et al. (2015). Bayesian software for hierarchical models most often makes use of the data augmentation (DA) algorithm. For the LMM, this implies that the random effects are estimated jointly with the other parameters. Hereby, the DA algorithm avoids to take the integral over the distribution of the random effects, which is the classical approach in the frequentist framework. Thus, in the frequentist approach classically the marginal version of the LMM is fitted to the data, while in the Bayesian approach the hierarchical or conditional version of the LMM is usually fitted.

Whether the marginal or the conditional version of the LMM is fitted to the data, it has an impact on the performance of the model selection criteria even when the conditional and marginal LMM essentially lead to the same model. The model selection criteria applied to the hierarchical specification of the LMM is referred to as the conditional criterion. Hence, one has the conditional DIC (cDIC), and similarly the conditional PSBF (cPSBF) and the conditional WAIC (cWAIC). On the other hand when the model selection criterion is applied to the marginal specification of the LMM, one speaks of the marginal DIC (mDIC), marginal PSBF (mPSFB) and marginal WAIC (mWAIC). As will be shown in Section 3.5, these two versions of the model selection criteria are associated with different aims: cDIC (and similarly for cPSBF and cWAIC) considers the random effects as parameters of focus in the model whereas for mDIC (also mPSBF and mWAIC) the population of random effects represents the focus. In practice, this implies for mixed effects models that the conditional selection criteria evaluate the performance of the model when the population consists of all (future) measurements of the subjects included in the current study, while the marginal version of the criteria measures the performance of the model for all (future measurements of all) future subjects from the same population.

The problem is that in practice, model selection is most often based on cDIC (cPSBF, cWAIC) because of computational convenience. Indeed, cDIC can be immediately calculated using the conditional likelihood and it is automatically reported by WinBUGS (Spiegelhalter et al., 2003) and other Bayesian software. However, most researchers are interested in knowing how well the model performs in the future. That is why one argues that conditional model selection criteria have the wrong focus, see e.g.Vaida and Blanchard (2005). Apart from not having the correct focus, model selection based on cDIC is questionable because the properties of DIC are based on the log-concavity of the likelihood, a condition that is violated in hierarchical models when the latent variables are considered as parameters in the model (Li et al., 2013b). The implication of using cDIC as model selection has been documented via simulations for financial volatility models (Chan and Grant, 2016a). The authors concluded that in contrast to mDIC, cDIC tends to select overly complex models. For overdispersed count data,Millar (2009) pointed out that the conditional-level DIC is an unreliable tool for model selection, while the same is true for the conditional WAIC (Millar, 2018). Merkle et al. (2018) advocated the use of marginal information criteria for item response models, and show that mWAIC corresponds to leave-one-cluster-out, whereas cWAIC corresponds to leave-one-unit-out.

While we focus in this paper on Bayesian model selection, we note that also in the frequentist paradigm the performance of the conditional versus marginal model selection criteria has been compared extensively. A broad overview of a wide range of model selection criteria for the LMM is discussed in Müller et al. (2013) for model selection in a frequentist content, including conditional and marginal information criteria. A short section in that paper is devoted to the Bayesian paradigm. Further, Fan et al. (2014) showed that the marginal AIC (mAIC) is asymptotically equivalent to the leave-one-cluster-out cross-validation while the conditional AIC (cAIC) is asymptotically equivalent to the leave-one-observation-out cross-validation. Srivastava and Kubokawa (2010) derived three conditional AICs and showed theoretically and by simulations that their proposals outperform cAIC and mAIC of Vaida and Blanchard (2005).Finally, Säfken et al. (2018) introduce the R-package cAIC4 for the calculation of the cAIC for LMMs estimated with lme4. To determine the marginal criteria extra computations are needed, which renders them less popular.

In practice, researchers are often not aware of the difference between the marginal and conditional version of the information criteria, therefore, rely on default software (Merkle et al., 2018). That is why we have set up a simulation study that compares the performance of the two versions of the selection criteria for LMMs with longitudinal data. The first set of simulations makes use of the classical model LMM assumptions, i.e. when the random effects and measurement errors have a normal distribution. In the second set of simulations, we have simulated from LMMs with a skewed-normal and t-distribution for the random effects and measurement errors. Finally, we considered settings were we select both fixed and random effect jointly. All these sets of simulations clearly show the superiority of the marginal selection criteria. Moreover, in the analysis of a real data set, we again illustrate that the conditional criteria choose the least appropriate LMM. In order to promote the use of the marginal criteria for LMMs, we have written R software for the LMMs considered in our simulation study that can easily be combined with classical Bayesian software to compute the criteria mDIC, mPSBF and mWAIC for LMMs.

The rest of the article is organized as follows. In Section 3.2 we present the classical linear mixed model for longitudinal data. In Section 3.3 we treat the skew-normal LMM. The model selection criteria are introduced in Section 3.4 and the difference between conditional and marginalized versions is discussed in Section 3.5. In Section 3.6 we compare the criteria in an extensive simulation study, in order to give some practical recommendations. We also compared alternative versions of DIC and WAIC as suggested in the literature. In the same section we discuss the simulation results when the normality assumption in the LMM is relaxed. A comparison of the conditional and marginal criteria on a real data set is done in Section 3.7. We give concluding remarks in Section 3.8.

3.2 The linear mixed-effects model

The classical LMM Laird and Ware (1982) for longitudinal data can be expressed as

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\epsilon}_i, \tag{3.1}$$

where $\mathbf{Y}_i$ is an m_i-dimensional response vector of measurements for the i-th subject, $(i = 1, \dots, n)$. $\mathbf{X}_i$ and $\mathbf{Z}_i$ are $m_i \times p$ and $m_i \times q$-dimensional covariate matrices, respectively, and $\boldsymbol{\beta}$ is a p-dimensional vector of fixed effects. The residual component vector $\boldsymbol{\epsilon}_i$ is distributed as $N_{m_i}(\mathbf{0}, \boldsymbol{\Sigma}_i)$, where $\boldsymbol{\Sigma}_i$ is an $m_i \times m_i$ positive-definite covariance matrix. It is usually assumed that $\boldsymbol{\Sigma}_i = \sigma_\epsilon^2\mathbf{I}_{m_i}$, where

$\mathbf{I}_{m_i}$ denotes the identity matrix of dimension m_i.

The q-dimensional random-effects vectors $\mathbf{b}_i$ are assumed independent from the residuals and distributed as $N_q(\mathbf{0}, \mathbf{D})$, where $\mathbf{D}$ is a $q \times q$ positive-definite covariance matrix. Model (3.1) is called a mixed-effects model because it combines the fixed-effects structure $\boldsymbol{\beta}$ with the subject-specific random effects $\mathbf{b}_1, \dots, \mathbf{b}_n$. The LMM is advantageous because the data are not required to be balanced, and additionally, the within- and between-individual variations can be explicitly modeled through $\boldsymbol{\Sigma}_i$ and $\mathbf{D}$, respectively.

In the frequentist setting, the model parameters are estimated from the marginalized model for the response, after integrating out the random effects (Verbeke and Molenberghs, 2000). The marginalized distribution has a closed form for model (3.1), namely

$$p(\mathbf{y}_i|\boldsymbol{\beta}, \mathbf{D}, \boldsymbol{\Sigma}_i) = N_{m_i}(\mathbf{X}_i\boldsymbol{\beta}, \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i' + \boldsymbol{\Sigma}_i). \tag{3.2}$$

In the Bayesian framework, inference is usually based on the hierarchical formulation of the model. In the first hierarchical stage, the response follows the conditional distribution $p(\mathbf{y}_i|\boldsymbol{\beta}, \boldsymbol{\Sigma}_i, \mathbf{b}_i) = N_{m_i}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = N_{m_i}(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i, \boldsymbol{\Sigma}_i)$, whilst in the second stage, the subject-specific effects are specified with distribution $p(\mathbf{b}_i|\mathbf{D}) = N_q(\mathbf{0}, \mathbf{D})$.

3.3 The skew-normal linear mixed model

A $m-$dimensional random vector $\mathbf{Y}$ follows a m-variate skew-normal (SN) distribution with location vector $\boldsymbol{\mu}_0 \in \mathbb{R}^m$, $m \times m$ positive definite scale matrix $\mathbf{H}$ and $m \times q$ skewness matrix $\boldsymbol{\Delta}$, if its density function is given by

$$\begin{aligned} f\left(\mathbf{y}|\boldsymbol{\mu}_0, \mathbf{H}, \boldsymbol{\Delta}\right) =& 2^q \boldsymbol{\phi}_m\left(\mathbf{y}|\boldsymbol{\mu}_0, \mathbf{H} + \boldsymbol{\Delta}\boldsymbol{\Delta}'\right) \\ & \times \boldsymbol{\Phi}_q\left(\boldsymbol{\Delta}'\left(\mathbf{H} + \boldsymbol{\Delta}\boldsymbol{\Delta}'\right)^{-1}(\mathbf{y} - \boldsymbol{\mu}_0)|\mathbf{0}, \left(\mathbf{I}_q + \boldsymbol{\Delta}'\mathbf{H}^{-1}\boldsymbol{\Delta}\right)^{-1}\right), \end{aligned} \tag{3.3}$$

where $\boldsymbol{\phi}_m$ and $\boldsymbol{\Phi}_q$ are the density function and the cumulative distribution functions of the m-dimensional and q-dimensional normal distribution, respectively. If we substitute $\boldsymbol{\Delta} = \mathbf{0}$, equation (3.3) reduces to the usual symmetric multivariate distribution $N_m(\boldsymbol{\mu}_0, \mathbf{H})$.

Arellano-Valle and Genton (2005) denote $\mathbf{Y} \sim SN_{m,q}(\boldsymbol{\mu}, \mathbf{H}, \boldsymbol{\Delta})$ and $\mathbf{Y} \sim SN_m(\boldsymbol{\mu}, \mathbf{H}, \boldsymbol{\Delta})$ when $m = q$. Also, when $m = q$, $\boldsymbol{\Delta} = \text{diag}\left(\delta_1, \dots, \delta_m\right)$ and $\mathbf{H}$ diagonal, equation (3.3) reduces to the multivariate skew-normal distribution,

(see e.g. Sahu et al., 2003). In practical settings, when the response and the covariate are highly skewed distributed, it might be more realistic to assume a multivariate SN for both random effects and measurement error (Huang and Dagne, 2012).
The classical LMM (3.1) can be extended by assuming that

$$\mathbf{b}_i \sim SN_q\left(\mathbf{0}, \mathbf{D}, \mathbf{\Delta}_b\right) \qquad \text{and} \qquad \boldsymbol{\epsilon}_i \sim SN_{m_i}\left(\mathbf{0}, \mathbf{\Psi}_i, \mathbf{\Delta}_{\epsilon_i}\right), \qquad i = 1, \ldots, n,$$

all independent. This results in the following skew-normal linear mixed model (SNLMM):

$$\begin{aligned} \mathbf{y}_i | \mathbf{b}_i, \boldsymbol{\beta}, \mathbf{\Psi}_i, \mathbf{\Delta}_{\epsilon_i} &\sim SN_{m_i}\left(\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i, \mathbf{\Psi}_i, \mathbf{\Delta}_{\epsilon_i}\right) \\ \mathbf{b}_i | \mathbf{D}, \mathbf{\Delta}_b &\sim SN_q\left(\mathbf{0}, \mathbf{D}, \mathbf{\Delta}_b\right), \end{aligned}$$

where $\mathbf{D} = \mathbf{D}(\boldsymbol{\alpha})$ is a dispersion matrix, usually associated with the between-units variances, with $\boldsymbol{\alpha}$ unknown parameters in $\mathbf{D}$. In addition, $\mathbf{\Delta}_{\epsilon_i}$ and $\mathbf{\Delta}_b$ are diagonal matrices with unknown elements $\delta_{\epsilon_{i1}}, \ldots, \delta_{\epsilon_{im_i}}$ and $\delta_{b_1}, \ldots, \delta_{b_q}$, respectively. These components correspond to the skewness parameters. The marginal version of the SNLMM was shown by Arellano-Valle et al. (2007) to be equal to

$$f_{Y_i}(\mathbf{y}_i|\mathbf{\Theta}, \boldsymbol{\vartheta}) = 2^{m_i+q}\phi_{n_i}(\mathbf{y}_i|\boldsymbol{X}_i\boldsymbol{\beta}, \mathbf{\Psi}_i)\mathbf{\Phi}_{m_{i+q}}(\boldsymbol{\mu}_{2i} - \mathbf{\Gamma}_i\boldsymbol{\mu}_{1i}|\mathbf{0}, \mathbf{R}_i + \mathbf{\Gamma}_i\mathbf{\Lambda}_i\mathbf{\Gamma}_i'),$$

where for $i = 1, \ldots, n$:

$$\mathbf{\Psi}_i = (\delta_\epsilon^2 + \sigma_\epsilon^2)\mathbf{I}_{m_i} + \mathbf{Z}_i(\mathbf{\Delta}_b^2 + \mathbf{D})\mathbf{Z}_i', \qquad \mu_{1i} = \frac{\mathbf{\Lambda}_i\mathbf{Z}_i'(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})}{\delta_\epsilon^2 + \sigma_\epsilon^2},$$

$$\mu_{2i} = \left(\frac{\delta_\epsilon}{\sqrt{\sigma_\epsilon^2(\delta_\epsilon^2 + \sigma_\epsilon^2)}}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})\right), \qquad \mathbf{\Gamma}_i = \begin{pmatrix} \frac{\delta_\epsilon}{\sqrt{\sigma_\epsilon^2(\delta_\epsilon^2+\sigma_\epsilon^2)}}\mathbf{Z}_i \\ -\mathbf{\Delta}_b(\mathbf{\Delta}_b^2 + \mathbf{D})^{-1} \end{pmatrix},$$

$$\mathbf{R}_i = \begin{pmatrix} \mathbf{I}_{mi} & \mathbf{0} \\ \mathbf{0} & (\mathbf{I}_q + \mathbf{\Delta}_b\mathbf{D}^{-1}\mathbf{\Delta}_b)^{-1} \end{pmatrix}, \qquad \mathbf{\Lambda}_i = \left((\mathbf{\Delta}_b^2 + \mathbf{D})^{-1} + \frac{\mathbf{Z}_i'\mathbf{Z}_i}{\delta_\epsilon^2 + \sigma_\epsilon^2}\right).$$

Note that Arellano-Valle et al. (2007) also suggested a skew-t distribution whereby the basic Gaussian distribution is replaced by the t-distribution.

3.4 Bayesian criteria for model selection

Let $\boldsymbol{\theta}$ represent all model parameters of the LMM. For the marginal LMM, this includes the fixed effects and the parameters making up the covariance matrix of the random effects augmented with skewness parameters for the SNLMM. With the conditional LMM the random effects are part of $\boldsymbol{\theta}$. Further, we denote the collected (longitudinal) responses by $\mathbf{y}$ and the obtained covariate values by the matrix $\mathbf{X}$. The posterior distribution is $p(\boldsymbol{\theta} \mid \mathbf{y}, \mathbf{X}) = p(\mathbf{y} \mid \boldsymbol{\theta}, \mathbf{X}) p(\boldsymbol{\theta}) / p(\mathbf{y} \mid \mathbf{X})$. Since the posterior distribution does not have a closed form for the LMM, it is approximated using MCMC methods. Namely, K (dependent) values $\boldsymbol{\theta}^1, \ldots, \boldsymbol{\theta}^K$ are sampled from the posterior distribution. The true posterior summary measures can then be approximated by their sampled versions.

When describing longitudinal data, a set of well-justified models can be established with different specifications for the fixed effects, random effects, covariance structure of the random effects and measurement error. Therefore, a model selection procedure is necessary to find an adequate model that explains current and future data. A variety of model selection procedures has been proposed in the Bayesian framework, but there is no consensus about the best criterion. Here we discuss the most popular criteria; they are also relatively easy to compute in practice.

3.4.1 The pseudo-Bayes factor

The Bayes factor (BF) could be viewed as the Bayesian equivalent of the likelihood ratio test. The Bayes factor can be used for testing the hypothesis that $\mathbf{y}$ is generated by model M_1 with parameters $\boldsymbol{\theta}_1$ versus the alternative model M_2 with parameters $\boldsymbol{\theta}_2$. Hereby BF measures the change from prior to posterior odds in favor of the null model, namely

$$\mathsf{BF}_{1,2} = \frac{p(\mathsf{M}_1 \mid \mathbf{y})}{1 - p(\mathsf{M}_1 \mid \mathbf{y})} = \frac{p(\mathsf{M}_1 \mid \mathbf{y})}{p(\mathsf{M}_2 \mid \mathbf{y})} = \frac{p(\mathbf{y} \mid \mathsf{M}_1)\, p(\mathsf{M}_1)}{p(\mathbf{y} \mid \mathsf{M}_2)\, p(\mathsf{M}_2)},$$

where $p(\mathsf{M}_1)$ and $p(\mathsf{M}_2)$ are the prior model probabilities, commonly set as $p(\mathsf{M}_1) = p(\mathsf{M}_2) = 0.5$. In that case, the Bayes factor in favor of model M_1 is given by $\mathsf{BF}_{1,2} = p(\mathbf{y} \mid \mathsf{M}_1)/p(\mathbf{y} \mid \mathsf{M}_2)$ where $p(\mathbf{y} \mid \mathsf{M}_r) = \int p(\mathbf{y} \mid \boldsymbol{\theta}_r, \mathsf{M}_r)\, p(\boldsymbol{\theta}_r \mid \mathsf{M}_r)\, d\boldsymbol{\theta}_r$ for $r = \{1, 2\}$. The use of the Bayes factor is, however, limited in practice since it has been shown to be quite sensitive to the choice of the prior distributions $p(\boldsymbol{\theta}_r \mid \mathsf{M}_r)$ and is not defined for improper priors (e.g. Gelfand and Dey, 1994).

Several alternatives for BF have been suggested to reduce the impact of $p(\boldsymbol{\theta}_r \mid \mathsf{M}_r)$. One proposal is PSBF, which is based on the partitions of the data set as follows. For the ith subject, one partitions the data set into a learning set $\mathbf{y}_L = \{\mathbf{y}_i : i \in L\}$ and a testing set $\mathbf{y}_T = \{\mathbf{y}_i : i \in T\}$ (Geisser and Eddy, 1979), whereby the testing and learning parts are defined respectively as $T = \{i\}$ and $L = \{1, ..., i-1, i+1, ..., n\}$. The pseudo-Bayes factor in favor of model M_1 with respect to model M_2 is then obtained as

$$\mathsf{PSBF}_{1,2} = \frac{\prod_{i=1}^{n} p(\mathbf{y}_i \mid \mathbf{y}_{(i)}, \mathsf{M}_1)}{\prod_{i=1}^{n} p(\mathbf{y}_i \mid \mathbf{y}_{(i)}, \mathsf{M}_2)},$$

where $\mathbf{y}_{(i)}$ is the total sample without $\mathbf{y}_i$. The component $p(\mathbf{y}_i \mid \mathbf{y}_{(i)}, \mathsf{M}_r)$ is the probability of observing $\mathbf{y}_i$ given the model M_r fitted with all observations in the sample except $\mathbf{y}_i$. Thus, the PSBF makes use of pseudo-marginal likelihoods in the numerator and denominator instead of the classical marginal likelihoods. The product terms are called conditional predictive ordinates (CPOs) Gelfand and Dey (1994). For the ith subject under model M_r, $\mathsf{CPO}_{r,i}$ is defined as $\mathsf{CPO}_{r,i} = p(\mathbf{y}_i \mid \mathbf{y}_{(i)}, \mathsf{M}_r)$. $\mathsf{CPO}_{r,i}$ is computed from the sampled values $\boldsymbol{\theta}_r^1, \ldots, \boldsymbol{\theta}_r^K$ under model M_r as follows:

$$\mathsf{CPO}_{r,i} \approx \left[\frac{1}{K}\sum_{k=1}^{K} \frac{1}{p(\mathbf{y}_i \mid \boldsymbol{\theta}_r^k, \mathsf{M}_r)}\right]^{-1}.$$

This statistic can be highly unstable for a very small value of the likelihood (Raftery et al., 2007). To ensure stability, different approaches have been prescribed in the literature (Raftery et al., 2007; Gelfand and Dey, 1994; Dey et al., 1997; Congdon, 2005). However, there is no perfect approach due to computational issues (Lachos et al., 2013).

The log-pseudo marginal likelihood is then for each model equal to $\mathsf{LPML}_r = \sum_{i=1}^{n} \log(\mathsf{CPO}_{r,i})$. Therefore, the $\mathsf{PSBF}_{1,2}$ in favor of model M_1 respect to model M_2 can be computed as

$$\mathsf{PSBF}_{1,2} = \exp(\mathsf{LPML}_1 - \mathsf{LPML}_2).$$

3.4.2 The deviance information criterion

The DIC suggested by Spiegelhalter et al. (2002) is based on the predictive accuracy of the estimated model defined as

$$\mathsf{DIC} = -2\log p(\mathbf{y}|\bar{\boldsymbol{\theta}}) + 2p_{DIC}, \tag{3.4}$$

where p_{DIC} corresponds to the effective number of parameters, given by

$$p_{DIC} = -2\ E_{\boldsymbol{\theta}|\mathbf{y}}[\log p(\mathbf{y}|\boldsymbol{\theta})] + 2\log[p(\mathbf{y}|\bar{\boldsymbol{\theta}})],$$

which quantifies the number of parameters to be estimated after incorporating the prior information into the model. As seen above, the point estimator is the posterior mean of the parameters, but other estimates such as the median have also been suggested. Defining the deviance as $D(\boldsymbol{\theta}) = -2\log\{p(\mathbf{y}|\boldsymbol{\theta})\} + 2\log\{f(\mathbf{y})\}$, the effective number of parameters can alternatively be written as $p_D = \overline{D(\boldsymbol{\theta})} - D(\bar{\boldsymbol{\theta}})$ where $\overline{D(\boldsymbol{\theta})}$ is the posterior mean of the deviance.

For practical purposes, we can ignore $f(\mathbf{y})$. The mean deviance $\overline{D(\boldsymbol{\theta})}$ can be approximated by $\frac{1}{K}\sum_{k=1}^{K} D(\boldsymbol{\theta}^k)$ and the plug-in deviance $D(\bar{\boldsymbol{\theta}})$ by $D(\frac{1}{K}\sum_{k=1}^{K}\boldsymbol{\theta}^k)$. This criterion is popular because it is easy to compute once we have an MCMC sample and can be directly obtained in several Bayesian packages such as WinBUGS. However, DIC has been criticised, (see Spiegelhalter et al., 2014) for details. For instance, DIC is not invariant to non-linear transformations of $\boldsymbol{\theta}$ and negative values for p_{DIC} can occur in some cases.

3.4.3 The widely applicable information criterion

The widely applicable information criterion (WAIC) (Watanabe, 2010) is a fully Bayesian estimator that averages over the posterior distribution of $\boldsymbol{\theta}$ instead of conditioning on a point estimator $\hat{\boldsymbol{\theta}}(\mathbf{y})$ as done for DIC. For a future observation $\tilde{\mathbf{y}}_i$, this criterion measures the predictive accuracy of the model based on the log-posterior predictive distribution $\log p_{\boldsymbol{\theta}|\mathbf{y}}(\tilde{\mathbf{y}}_i)$ of the parameter vector $\boldsymbol{\theta}$. Since $\tilde{\mathbf{y}}_i$ is unknown, predictive accuracy is defined by the expected log-predictive distribution (elpd) as

$$\text{elpd}_i = E_f[\log p_{\boldsymbol{\theta}|\mathbf{y}}(\tilde{\mathbf{y}}_i)] = \int \log p_{\boldsymbol{\theta}|\mathbf{y}}(\tilde{\mathbf{y}}_i) f(\tilde{\mathbf{y}}_i) d\tilde{\mathbf{y}}_i,$$

where f is the unknown distribution under the true model. For each observation of a new data set, elpd is computed to establish the predictive accuracy of that data set. This is called the expected log-pointwise predictive density (elppd) defined as elppd $= \sum_{i=1}^{n} E_f[\log p_{\boldsymbol{\theta}|\mathbf{y}}(\tilde{\mathbf{y}}_i)]$.

Predictive accuracy can also be defined with a point estimate $\hat{\boldsymbol{\theta}}(\mathbf{y})$, often $\hat{\boldsymbol{\theta}}(\mathbf{y}) = E(\boldsymbol{\theta}|\mathbf{y})$, as the expected log predictive distribution given the point estimator elpd$_{\hat{\boldsymbol{\theta}}(\mathbf{y})} = E_f(\log p(\tilde{y}|\hat{\boldsymbol{\theta}}(\mathbf{y})) = \int \log p_{\boldsymbol{\theta}|\mathbf{y}}(\tilde{\mathbf{y}}_i) f(\tilde{\mathbf{y}}_i) d\tilde{\mathbf{y}}_i$. The log pointwise predictive distribution (lppd) based on the observed data is calculated as follows

$$\text{lppd} = \log \prod_{i=1}^{n} p_{\boldsymbol{\theta}|\mathbf{y}}(\mathbf{y}_i) = \sum_{i=1}^{n} \log \int_{\boldsymbol{\theta}} p(\mathbf{y}_i|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta}.$$

In practice, lppd can be estimated using an MCMC sample from the posterior distribution as

$$\widehat{\mathsf{lppd}} = \sum_{i=1}^{n} \log \left[\frac{1}{K} \sum_{k=1}^{K} p(\mathbf{y}_i | \boldsymbol{\theta}^k) \right].$$

With the WAIC criterion, the expected log pointwise predictive density elppd is estimated as the log pointwise predictive distribution lppd with a bias correction $\widehat{\mathsf{elppd}}_{WAIC} = \widehat{\mathsf{lppd}} - p_{WAIC}$. The measure p_{WAIC} corresponds to an estimate of the effective number of parameters given by

$$p_{WAIC} = 2 \sum_{i=1}^{n} \left[\log \left(\frac{1}{K} \sum_{k=1}^{K} p(\mathbf{y}_i | \boldsymbol{\theta}^k) \right) - \frac{1}{K} \sum_{k=1}^{K} \log \; p(\mathbf{y}_i | \boldsymbol{\theta}^k) \right].$$

Note that, WAIC can be alternatively expressed as

$$\mathsf{WAIC} = -2\widehat{\mathsf{lppd}} + 2p_{WAIC},$$

similar to DIC in (3.4).

One of the strengths of WAIC is its invariability to the scale of the model parameters, which implies that WAIC does not change when $\boldsymbol{\theta}$ is replaced by $\boldsymbol{\psi} = h(\boldsymbol{\theta})$, with h a strictly monotone function.

3.5 Marginal and conditional criteria

In practice, the choice between conditional and marginal information criteria should be motivated by the aim of the study (Vaida and Blanchard, 2005). Most often, this means that the marginal model selection criteria should be used since they estimate the predictiveness of the model when new clusters (in longitudinal studies, this implies new subjects) are involved, whereas the conditional criteria estimate the predictiveness of the model when new elements in the cluster (in longitudinal studies, new observations from the existing subjects) are involved.

Nevertheless, when it comes to selecting the correct LMM it might still be that conditional criteria do a good job. In other words, it might be that the relative ordering of preference models is basically the same for both the conditional and marginal criteria. All of these comments apply to all three considered model selection criteria, but since cDIC is obtained automatically in most Bayesian software, it is the standard criterion in practice. Therefore, the literature shows some focus on DIC when examining the performance of conditional and marginal criteria. Despite the popularity of DIC, many have shown that the asymptotic justification of DIC (Spiegelhalter et al., 2002) does not hold for hierarchical models, see e.g. Li et al. (2013a).

3.6 Simulation studies

We have carried out three simulation studies. In the first two studies we based the simulated data on two classical data sets: the Potthoff and Roy data set (Potthoff and Roy, 1964) and the Jimma Infant Growth study (Lesaffre et al., 1999). They were chosen because the first is representative for a balanced longitudinal study, while for the second study the time points are (somewhat) irregular and subjects drop out from the study. Using the fitted LMMs as population models, the performance of the conditional and marginal versions of DIC, PSBF and WAIC are contrasted using simulations.

mDIC can be obtained from a WinBUGS run by working with the marginal model instead of the hierarchical model. To avoid specifying the marginal model in the estimation process, an R function was implemented, which computes the marginalized version of DIC, PSBF and WAIC for a Gaussian, skew-normal and skew-t distribution of the random effects and measurement error. This R function takes the parameters sampled in the MCMC procedure from any Bayesian package and calculates the marginalized version using the closed form (3.2) and its extensions allowing for skew-normal and skew-t distributions. In addition, the conditional version of the three criteria is also computed by this function.

The main objective of the simulation study is to assess how well PSBF, DIC and WAIC select the correct model. According to the *minimum value* strategy, the model with the minimum value for the criterion is selected. Several simulation studies examining the performance of AIC and BIC, see e.g. Lesaffre and Lawson (2012), suggest to select the more complex model only if they differ in the criterion value with more than 5. This will be referred to as the *absolute difference* strategy. We will apply this strategy to all criteria. However, there is no evidence that this criterion is justified outside DIC.

3.6.1 The data sets and population models

In the dental study analyzed by Potthoff and Roy (1964), the distance in (mm) from the pituitary to the pterygomaxillary fissure was measured at years 8, 10, 12, and 14 on 11 girls and 16 boys. We fitted the following linear mixed model as a function of *age* and *sex* (0= Female, 1=Male):

$$y_{ij} = \beta_0 + \beta_1 \mathsf{sex}_i + \beta_2 \mathsf{age}_{ij} + b_{0i} + \epsilon_{ij}, \quad (i = 1, \ldots, 27; j = 1, \ldots, 4), \quad (3.5)$$

where y_{ij} is the distance (mm) measure of child i at time j and b_{0i} is a random intercept assumed to follow $b_{0i} \sim N(0, \sigma_b^2)$. Using the SAS procedure MIXED

(Littell et al., 2007), we obtained the following maximum likelihood estimates: $\hat{\beta}_0 = 24.9688$, $\hat{\beta}_1 = 1.4831$, $\hat{\beta}_2 = -2.3210$, $\hat{\sigma}_b^2 = 2.0495$ and $\hat{\sigma}_\epsilon^2 = 3.2668$.

These values were used as true parameters in this simulation study. The Jimma Infant Growth data set is based on the growth characteristics of about 8000 live births from South-West Ethiopia examined between September 1992 and September 1993. The growth characteristics height, weight and arm circumference of the babies were examined approximately every 60 days, but there were occasional deviations from the planned visits. Also, some children dropped out from the study for a variety of reasons such as relocation of their parents during the study or death of the child. This creates an unbalanced structure for the data. For the purpose of this simulation study, we have taken weight as response with covariates *age* and *sex* (0=Girls, 1=Boys) of the child, and *age of the mother at delivery* (agem). The details of the original analysis can be found in Lesaffre et al. (1999, 2000) where a sample of 495 children was selected to fit the model. This subset will also be the basis for this simulation study. The weight evolves in a non-linear way. To make use of an LMM, the time variable age was transformed into $\text{newage}_{ij} = \sqrt{\text{age}_{ij}} - (\text{age}_{ij} + 1) - 0.02 \times \text{age}_{ij}$ using fractional polynomials Lesaffre et al. (2000). Initially, our population model is based on the following random intercept and slope model:

$$y_{ij} = \beta_0 + \beta_1 \text{sex}_i + \beta_2 \text{newage}_{ij} + \beta_3 \text{agem}_i + b_{0i} + b_{1i} \times \text{newage}_{ij} + \epsilon_{ij}, \quad (3.6)$$

assuming $(b_{0i}, b_{1i})' \sim N(\mathbf{0}, \boldsymbol{D})$. Again, the estimates from this model (see Appendix) are used as the true values for the parameters in the simulation.

3.6.2 Simulation study 1

In the first simulation study, we consider the most popular setting of assuming normality for the random effects and measurement error. We believe that it is essential to show the performance of the selection criteria in this most popular setting. The performance of the model selection criteria may depend on whether the models differ in the fixed components or the random effects structure. Therefore, we examined the performance of the conditional and marginal criteria under two scenarios. For each of the two data sets we considered two scenarios. In *Scenario I* we assumed that the random effects structure is known but that the considered models differ from the true model in the fixed part. For *Scenario II* we assumed that the fixed part is known but the random effects part is unknown. Regarding the prior distributions, we assigned independent vague normal priors, $N(0, 1000^2)$ for the regression coefficients and a vague inverse gamma prior for the residual variance, i.e. $\sigma^2 \sim IG(0.001, 0.001)$. The conditionally conjugate prior for the random-effects covariance matrix is the inverse Wishart distribution, but this choice has been shown to be problematic when the number of clusters (here subjects) is small (Gelman, 2006; Quintero and Lesaffre, 2017). Therefore, we have taken uniform priors $U(0, 100)$ for the standard deviation of the random effects, (see Gelman, 2006). For the models with at least random intercept and slope, we assigned a uniform prior distribution $U(-0.5, 0.5)$ for all pairwise correlations between random effects to ensure positive definiteness of the covariance matrix $\mathbf{D}$ (Plummer, 2011) following a proof in Coakley and Rokhlin (2013).

The balanced case: the Potthoff and Roy data set

As indicated above, we have considered two scenarios:

Scenario I: We assumed that the random effects structure is correct and considered models that differ in the fixed part. Besides the true data-generating model (3.5), we considered an overspecified model, which includes the interaction of age with sex and an underspecified model, which ignores the effect of sex. Hence, the alternative models are

- $y_{ij} = \beta_0 + \beta_1 \text{age}_{ij} + \beta_2 \text{sex}_i + \beta_3 \text{age}_{ij} \times \text{sex}_i + b_{0i} + \epsilon_{ij}$ (overspecified),
- $y_{ij} = \beta_0 + \beta_1 \text{age}_{ij} + b_{0i} + \epsilon_{ij}$ (underspecified).

Scenario II: We assumed that the fixed structure is correct and considered models that differ in the random effects. The overspecified model includes an additional random slope whereas the underspecified alternative ignores the random intercept in the data, more specifically

- $y_{ij} = \beta_0 + \beta_1 \text{age}_{ij} + \beta_2 \text{sex}_i + b_{0i} + b_{1i} \times \text{age}_{ij} + \epsilon_{ij}$ (overspecified),
- $y_{ij} = \beta_0 + \beta_1 \text{age}_{ij} + \beta_2 \text{sex}_i + \epsilon_{ij}$ (underspecified).

We simulated 500 data sets based on model (3.5). The covariate age was taken as in the original data set and sex was generated from a Bernoulli distribution with probability of success equal to 0.6, where 0.6 is the proportion of boys in the original data set. All the models in this simulation study were estimated based on three chains of $15,000$ iterations (discarding the first $5,000$ as a burn-in) and thinning equal to 10. Convergence of the MCMC samples was assessed with the Brooks-Gelman-Rubin (BGR) diagnostic. In cases where

BGR was larger than 1.1, a new MCMC sample was selected with $10,000$ extra iterations until obtaining convergence.

In Table 3.1, we present for each criterion and for the two selection strategies, the percentage of times the correct, the overspecified or the underspecified model was chosen. The performance of the marginalized criteria is clearly better than the conditional counterparts in all cases.

For instance, when using the *minimum value* selection rule, in most cases the percentage of correct selection for the marginalized version is almost twice that of the conditional counterpart.

In addition, note that for the *absolute difference* rule in Scenario I, the percentage of correct model selections for the conditional version of DIC and of WAIC is basically zero. This strategy seems to work well also for PSBF and WAIC in Scenario II, but not in Scenario I. In Scenario II, the conditional versions of DIC, PSBF and WAIC favor overspecified models with additional random effects as also observed in Chan and Grant (2016a) for financial volatility models.

Table 3.1*: Simulation study 1: Performance of the Bayesian model selection criteria for the Potthoff & Roy data set.*

		Minimum value			Absolute difference		
Scenario	criteria	Over	Correct	Under	Over	Correct	Under
I	cDIC	18.6	67.6	13.8	2.4	1.0	96.6
	mDIC	16.8	76.4	6.8	1.4	55.2	43.4
	cPSBF	27.0	43.0	30.0	18.6	29.8	51.6
	mPSBF	17.6	75.2	7.2	2.8	65.2	32.0
	cWAIC	19.8	31.0	49.2	2.6	0.0	97.4
	mWAIC	18.8	75.0	6.2	1.4	58.4	40.2
II	cDIC	46.2	53.8	0.0	10.4	89.6	0.0
	mDIC	15.0	85.0	0.0	0.6	99.4	0.0
	cPSBF	52.4	47.6	0.0	32.0	68.0	0.0
	mPSBF	14.4	85.6	0.0	1.2	98.8	0.0
	cWAIC	63.2	36.8	0.0	16.0	84.0	0.0
	mWAIC	18.0	82.0	0.0	0.8	99.2	0.0

The unbalanced case: the Jimma Infant growth study

Again we considered two scenarios:

Scenario I: We assumed that the random effects structure is correct and considered the following models that differ in the fixed part parameters, namely

- Model (3.6) and including the interaction newage × sex (overspecified),
- Model (3.6) but ignoring the covariate sex (underspecified).

Scenario II: We assumed that the covariates in the fixed part are correct and considered the following models that differ in the random effects structure, i.e.

- Model (3.6) and including an additional random slope for newage^2 (overspecified),
- Model (3.6) but ignoring the random slope for newage (underspecified).

Table 3.2: *Simulation study 1: Performance of the Bayesian model selection criteria for the Jimma Infant Growth data set.*

		Minimum value			Absolute difference		
Scenario		Over	Correct	Under	Over	Correct	Under
I	cDIC	34.4	34.0	31.6	15.2	29.0	55.8
	mDIC	21.2	58.0	20.8	0.8	32.4	66.8
	cPSBF	33.0	32.8	34.2	47.0	31.8	21.2
	mPSBF	21.0	57.8	21.2	3.0	44.0	53.0
	cWAIC	36.2	31.2	32.6	14.4	26.4	59.2
	mWAIC	21.2	58.2	20.6	0.8	32.6	66.6
II	cDIC	63.2	36.8	0.0	43.2	56.8	0.0
	mDIC	26.4	73.6	0.0	0.2	99.8	0.0
	cPSBF	55.2	44.8	0.0	51.8	48.2	0.0
	mPSBF	28.0	72.0	0.0	2.8	97.2	0.0
	cWAIC	66.0	34.0	0.0	49.2	50.8	0.0
	mWAIC	27.4	72.6	0.0	0.2	99.8	0.0

We generated 500 data sets from model (3.6). The covariate age was taken as in the original data set (i.e 8,10,12,14) and sex was generated from a Bernoulli distribution with probability of success equal to 0.6, where 0.6 is the proportion of boys in the original data set. The age of the mother was generated from a normal distribution agem$_i \sim N(24.49, 6.29)$ and we have taken $0, 60, 120, \ldots, 360$ days as the moments of measurements. We created an unbalanced data set by allowing subjects to drop out randomly at days 240, 300 or 360.

As shown in Table 3.2, the marginalized criteria strongly outperform their conditional counterparts in both scenarios and selection strategies. We see again for Scenario II that all conditional criteria support the overspecified alternative with an additional random slope and that in this scenario the *absolute difference* strategy also works for PSBF and WAIC. With the *minimum value* rule, the probability of correctly selecting the data-generating model is about 1/3 with the conditional criteria. Hence, carrying out model selection based on the conditional criteria performs worse than selecting the models at random.

3.6.3 Simulation study 2: additional simulations for the balanced case

We first evaluated the sensitivity of the results to some changes in the population model based on the Potthoff and Roy data. First, we varied the signal-to-noise ratio in model (3.5) by setting the value of σ_ϵ^2 to be $\frac{1}{4}$, $\frac{1}{2}$, 1, 2 and 4 times of the estimated residual variance as specified in Section 3.6.1. Table 3.3 displays the results on model selection. Again, the marginal criteria outperform their conditional counterparts irrespective of the scenario and selection strategy. Note that the performance of mDIC decreases with increasing residual variance and using the *absolute difference* strategy.

Second, we varied the number of subjects in the study as 25, 50, 75 and 100. As shown in Table 3.4, the marginal criteria perform best regardless of the sample size. Note also that the performance of the marginal criteria increases with increasing sample size in both scenarios and selection strategies, which is not the case for the conditional criteria. For instance, the percentage of correct model selection for cDIC decreases with sample size for Scenario II with both selection rules. Our results are in line with the findings in Li et al. (2013b), who pointed out asymptotic problems with cDIC. Our simulation study also indicates that cWAIC is not better in this sense.

We additionally evaluated the model selection performance for alternative versions of DIC and WAIC. We denote as DIC$_1$ the criterion advocated in Spiegelhalter et al. (2002) where the complexity (p_{DIC1}) is defined in Section 3.4.2. The alternative version DIC$_2$ is the approximation to DIC$_1$ (Gelman et al., 2004a). The complexity penalty (p_{DIC2}) is a function of the variance of the deviance calculated as

$$p_{DIC2} = 2\text{var}_{\boldsymbol{\theta}|\boldsymbol{y}}(\log\{p(\mathbf{y}|\boldsymbol{\theta})\}). \tag{3.7}$$

Further, we modified DIC by letting the penalty term depend on the sample size. It has been suggested in Jones (2011) that the penalization should be defined based on the effective sample size n_e, which depends on the within-subjects error structure. In the context of the LMM, statistical software like SAS defines n_e as the total number of (independent) subjects, i.e. $n_e = n$. Otherwise, n_e is defined as the number of total data points, $n_e = n_T$.

We defined the following DIC criteria as DIC$_3$ and DIC$_4$ with effective degrees of freedom defined as $p_{DIC3} = \log(n)\, p_{DIC1}$ and $p_{DIC4} = \log(n_T)\, p_{DIC1}$, respectively. These modifications are more a BIC-type as pointed out by a referee, however, we believe that it will be a useful exercise to evaluate their performance in this context.

The effective number of parameters of WAIC can be estimated in two ways (Gelman et al., 2014); p_{WAIC1} as defined in Section 3.4.3 and the alternative version p_{WAIC2} given as the variance of the log posterior distribution as

$$p_{WAIC2} = \sum_{i=1}^{n} \text{var}_{\boldsymbol{\theta}|\boldsymbol{y}}(\log p(\boldsymbol{y}_i|\boldsymbol{\theta})).$$

We notice from Table 3.5 that Spiegelhalter's DIC (DIC$_1$) outperforms DIC$_2$ for the conditional versions. This may be expected since the alternative definition (3.7) is explicitly based on approximate posterior normality, which is likely not satisfied in the hierarchical version of the model. The marginal versions of DIC$_1$ and DIC$_2$ perform similarly.

As expected, DIC$_4$ penalizes model complexity more heavily than DIC$_3$. Regardless of the selection strategy, we observed that by increasing the penalization, the percentage of correct model selection decreases under the marginal versions and increases under the conditional versions.

As for the different versions of WAIC, we observed that the percentage of correct selection for WAIC$_2$ is slightly higher in the conditional version whereas the performance of the marginal versions is similar irrespective of the scenario. *Absolute difference*, however, is not a good alternative to the conditional version of DIC and WAIC alternatives.

Table 3.3: *Simulation study 2: Percentage correct selection when changing the residual variance in the Potthoff & Roy data set.*

Scenario		Minimum value					Absolute difference				
	Criteria	0.25	0.5	1	2	4	0.25	0.5	1	2	4
I	cDIC	64.6	70.2	77.0	77.8	79.2	0.6	1.2	3.2	10.8	24.6
	mDIC	81.6	83.0	83.0	82.8	82.0	93.0	92.8	92.6	88.6	78.6
	cPSBF	31.8	36.6	40.8	58.4	68.0	30.3	38.2	39.0	39.6	39.4
	mPSBF	91.2	94.0	83.2	90.8	87.8	95.4	97.8	93.0	97.8	94.0
	cWAIC	41.4	36.6	39.4	38.4	39.0	0.4	0.2	0.2	0.4	0.2
	mWAIC	81.2	81.6	82.4	82.0	81.6	92.2	93.0	92.8	89.0	79.0
II	cDIC	44.4	47.4	50.8	51.6	55.4	86.2	86.2	87.2	88.4	89.0
	mDIC	80.4	82.4	83.6	85.4	86.4	99.2	99.4	99.6	99.6	90.2
	cPSBF	60.4	58.4	44.8	62.2	73.4	52.0	55.8	65.8	67.5	69.6
	mPSBF	83.8	86.8	84.2	84.0	83.8	98.7	97.9	97.6	91.2	86.4
	cWAIC	34.4	32.8	34.2	36.6	36.2	81.4	81.8	83.4	81.0	82.6
	mWAIC	77.6	81.0	82.6	82.0	82.4	97.6	99.2	99.2	99.0	92.2

3.6.4 Simulation study 3: extra simulation for possible extensions of LMM

Simulation study: jointly selection of both fixed and random effects

Depending on the data at hand, researchers are usually faced with the challenge of choosing the correct model. It is therefore important to select a parsimonious model that fits the data accurately. Since there is minimal agreement on which criteria to choose for Bayesian model selection, we evaluated the performance of the marginal and conditional criteria in choosing the correct model among other alternative models. Based on Potthoff & Roy data, we generated 500 data sets from Equation (3.5) and considered five possible alternative models for the data. We considered, namely, (i) different scale of the covariates (ii) distributional assumptions not satisfied for either or both random-effects and measurement error (iii) the nature of measurement error (heteroscedastic or heteroscedastic) (iv) wrong random effects structure. The following models were considered jointly with the model given by Equation (3.5).

- C1: The model generating data specified in Equation (3.5).
- C2: Equation (3.5) with age replaced by age^2 and including an additional random slope for age.

Table 3.4*: Simulation study 2: Percentage correct selection when changing the sample size in the Potthoff & Roy data set.*

Scenario		Minimum value				Absolute difference			
	Criteria	25	50	75	100	25	50	75	100
I	cDIC	67.6	77.0	79.0	80.6	1.0	3.2	7.2	19.0
	mDIC	76.4	83.0	84.2	82.8	52.2	92.6	98.4	99.0
	cPSBF	43.0	40.8	49.4	45.4	0.0	0.4	0.8	0.0
	mPSBF	75.2	83.0	84.4	83.0	83.1	93.0	93.2	96.1
	cWAIC	31.0	39.4	41.4	43.8	0.0	44.8	44.6	40.6
	mWAIC	75.0	82.4	83.8	82.0	56.2	92.8	98.8	98.8
II	cDIC	53.8	50.8	47.4	41.0	89.6	87.2	87.2	84.8
	mDIC	85.0	83.6	86.2	83.8	99.2	99.6	99.2	99.4
	cPSBF	47.6	44.8	47.8	53.0	65.2	65.8	66.2	65.8
	mPSBF	85.6	84.2	86.0	83.0	90.2	97.6	97.6	97.9
	cWAIC	36.8	34.2	34.2	31.8	83.8	83.4	83.2	82.6
	mWAIC	82.0	82.6	84.6	80.2	99.4	99.2	99.2	99.3

- C3: Equation (3.5) age replaced by age^2.
- C4: Equation (3.5) age replaced by $\log$(age).
- C5: Equation (3.5) with the normality assumption for random effects replaced by the skew-normal assumption.
- C6: Equation (3.5) with the normality assumption for random effects replaced by the skew-normal assumption and heteroscedastic measurement error is assumed.

As seen in Table 3.6, the marginal criteria select the data-generating model (C1) in about 70% of the times contrary to the conditional criteria which select the true model in about 10% of the time. It is interesting to note that the conditional criteria select C5 (the model that assumes a skew-normal distribution for the random effects) in about 65% while the marginal criteria choose C5 in about 2%. The results show the superiority of the marginal criteria in selecting the true data-generating model.

Simulation study: normality assumption for the random effects and measurement errors are relaxed

We also assessed the performance of the model selection criteria when the normality assumption for the random effects and measurement errors are relaxed.

For this simulation study, we generated 500 data sets from the model.

$$y_{ij} = \beta_0 + x_i\beta_1 + t_{ij}\beta_2 + b_{0i} + \epsilon_{ij}, i = 1, \ldots, n = 200, j = 1, \ldots, 6 \qquad (3.8)$$

where $t_{ij} = j$, $\beta_1 = 2$, $\beta_2 = 1$ and $\epsilon_{ij} \sim SN_1(0, 0.5^2, 4)$. First, we assumed that $\beta_0 + b_{0i} \sim N(4, 4)$, i.e, $\beta_0 = 4$ and $b_{0i} \sim N(0, 4)$. In addition, to show the advantages of the skew-normal distribution for the random effect it is penchant to accommodate skewness.

Second, we have taken the previous one except now we generated the $\beta_0 + b_{0i}$ according to *Gamma*$(2, 1)$ distribution (as done also in Arellano-Valle et al. (2007) and Lachos et al. (2010)) with probability density $f(x) = x\exp(-x)$ yielding a highly skewed distribution. The subject-specific covariate x_i is binary with $x_i = 1$ if $i \leq n/2$ and is zero otherwise, while t_{ij} represents a covariate with values varying within individuals and the same for all individuals.

For each of the 500 simulated data sets, model (3.8) was fit under alternative models as described in Section 3.6.2. We sampled 7000 iterations after discarding the initial 3000 iterations. The thinning factor was at 7 to avoid correlation problems in the generated chains.

The following vague priors were assigned: $\beta \sim N(0, 10^2)$, $\sigma^2_\epsilon \sim IG(0.001, 0.001)$, $\sigma^2_b \sim IG(0.001, 0.001)$, $\delta_\epsilon \sim N(0, 10^2)\mathbb{I}\delta_\epsilon > 0$, $\delta_b \sim N(0, 10^2)\mathbb{I}\delta_b > 0$. The marginal distribution corresponding to Equation (3.8) is expressed in the closed form, as seen in Section 3.3.

The simulation results shown in Table 3.7 confirm the results obtained above under the Gaussian distribution. Finally, we repeated the above simulation when (i) both random effects and random error have a skew-normal distribution and when (ii) the random error follows a $t(3)$ distribution. The results (not shown) confirm the above simulation results.

3.7 Application

The Nigerian indigenous chicken (NIC) data set describes the longitudinal evolution of the body weight (BW) of chickens of different breeds raised in a university experimental farm. Four hundred and sixteen chickens were measured every week from hatching up to 20 weeks. The study aimed to evaluate the growth of different chicken breeds. Here we considered two classes of progenies.

Two hundred and seventy chickens were produced from the same parent stock (pure breed), while 146 chickens have different parents (cross breed). The rational for the study and the experimental design can be found in Adeleke et al. (2011).

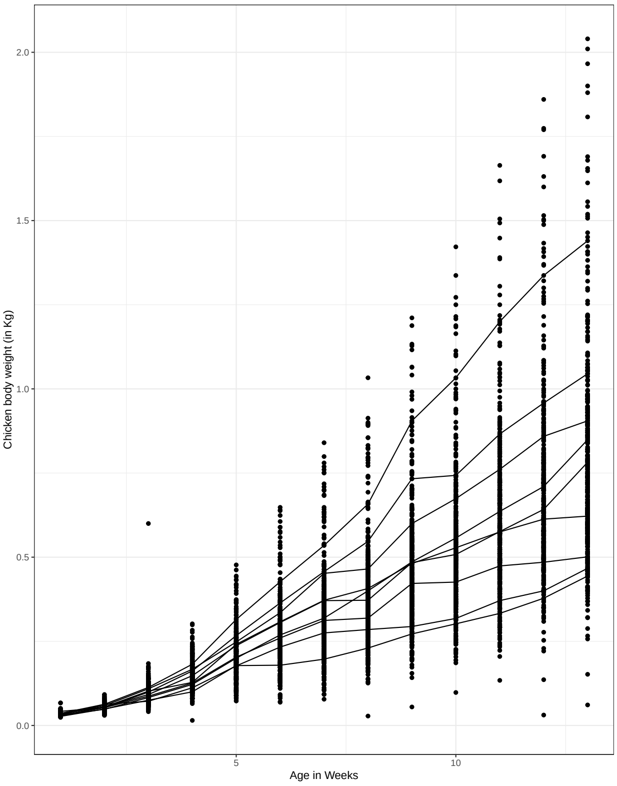

Figure 3.1*: Nigerian indigenous chicken data set: Longitudinal profiles of body weight for 416 chickens highlighting 10 randomly chosen chickens*

See Figure 3.1 for the evolution of weights of the chickens over time. Assuming a quadratic growth model with subject-specific random intercept and slopes, we fitted an LMM model to the weight at the jth measurement time of the ith chicken as

$$y_{ij} = \beta_0 + \beta_1 breed_i + \beta_2 age_{ij} + \beta_3 age_{ij}^2 + b_{0i} + b_{1i} age_{ij} + b_{2i} age_{ij}^2 + \epsilon_{ij}, \quad (3.9)$$

where y_{ij} is the chicken body weight (kg); breed$_i$ is the breed indicator (1=pure breed, 2=cross breed), the age$_{ij}$ represents the age (standardized).

For the purpose of this study, we limited the chicken's age to 13 weeks since after that age a considerable amount of chicken died. Thus, $\mathbf{x}_{ij} = (1, \text{breed}_i, \text{age}_{ij}, \text{age}_{ij}^2)'$, $\mathbf{b}_i = (b_{0i}, b_{1i}, b_{2i})'$ and $\mathbf{Z}_{ij} = (1, \text{age}_{ij}, \text{age}_{ij}^2)$, $i = 1, \ldots, 416$, $j = 1, \ldots, 13$.
We first used model (3.9) together with the classical Gaussian assumptions as model to fit the weights of the chickens over time, and we refer to this as Model 3.9(a).

Based on the model fit, Figure 3.2 shows histograms and the corresponding Q-Q plots of the standardization posterior means of $\mathbf{b}_i$ and ϵ_{ij}, whereby the posterior means were divided by their corresponding posterior standard deviations. The plots show that there is apparently a non-normal pattern for subject-specific intercepts and slopes.

Also, the residual plot suggests deviation from normality. We note that such plots may be difficult to interpret because the shrinkage effect depends on the number of measurements per subject,(see e.g. Verbeke and Lesaffre, 1996). But here there were no missing responses up to week 13 and standardisation was applied.

Nevertheless, these plots triggered us to consider three additional models with the same fixed effects structure but differing in the error and random effects distribution:

- **Model 3.9(b):** LMM with a univariate skew-normal distribution for measurement error and a trivariate Gaussian distribution for the random effects.
- **Model 3.9(c)**: LMM with model with a trivariate skew normal random effects with Gaussian measurement error.
- **Model 3.9(d):** LMM with a univariate skew-normal distribution for measurement error and a trivariate skew-normal distribution for the random effects.

The vague priors used are the same as those described in Section 3.6.4. We used 25,000 iterations after discarding the first 10,000 and thinning was set to 10. Convergence of the MCMC samples was assessed with the BGR criteria. Resulting parameter estimates are shown in Table 3.8.

It can be observed from Table 3.8 that the conditional criteria support Model 3.9(b), which seems to be an incorrect model based on Figure 3.2. In contrast, the marginal criteria favor Model 3.9(d), which appears to be also the most appropriate model here. We further evaluated the effect of the quadratic term in the fixed and random effects.

The results (results not shown) of both versions of the criteria show that age^2 is more important in the random effects part than in the fixed part and there is an agreement between the conditional and the marginal criteria on this.

Table 3.5*: Simulation study 2: Performance of alternative criteria for the Potthoff & Roy data set.*

		Minimum value			Absolute difference		
Scenario	Criteria	Over	Correct	Under	Over	Correct	Under
I	$cDIC_1$	18.6	67.6	13.8	2.4	1.0	96.6
	$cDIC_2$	11.8	36.0	52.2	1.6	0.0	98.4
	$cDIC_3$	3.2	85.0	11.8	0.6	22.8	76.6
	$cDIC_4$	4.2	40.4	55.4	1.4	19.6	79.0
	$cWAIC_1$	19.8	31.0	49.2	2.6	0.0	97.4
	$cWAIC_2$	16.8	41.2	42.0	2.6	0.0	97.4
	$mDIC_1$	16.8	76.4	6.8	1.4	55.2	43.4
	$mDIC_2$	16.8	73.8	9.4	1.4	52.2	46.4
	$mDIC_3$	1.8	65.4	32.8	0.2	34.0	65.8
	$mDIC_4$	2.8	53.2	44.0	0.2	24.0	75.8
	$mWAIC_1$	18.8	75.0	6.2	1.4	58.4	40.2
	$mWAIC_2$	17.8	75.2	7.0	1.4	56.2	42.4
II	$cDIC_1$	46.2	53.8	0 .0	10.4	89.6	0.0
	$cDIC_2$	0.6	99.2	0.2	0 .0	99.4	0.6
	$cDIC_3$	0.0	47.8	52.2	0.0	36.8	63.2
	$cDIC_4$	0.0	0.8	99.2	0 .0	0.6	99.4
	$cWAIC_1$	63.2	36.8	0.0	16.0	84.0	0.0
	$cWAIC_2$	55.8	44.2	0.0	10.4	89.6	0.0
	$mDIC_1$	15.0	85.0	0.0	0.6	99.4	0.0
	$mDIC_2$	8.0	92.0	0.0	0.2	99.8	0.0
	$mDIC_3$	2.4	97.6	0.0	0.2	99.6	0.2
	$mDIC_4$	0.4	99.6	0.0	0.0	99.2	0.8
	$mWAIC_1$	18.0	82.0	0.0	0.8	99.2	0.0
	$mWAIC_2$	15.4	84.6	0.0	0.8	99.2	0.0

Table 3.6*: Simulation study 3: Percentage of times the criteria selection select the required model described in Section 3.6.4 in the Potthoff &Roy data set.*

	Model					
Criteria	C1	C2	C3	C4	C5	C6
cDIC	12.8	7.0	3.6	4.0	70.6	2.0
cWAIC	13.2	8.4	8.0	4.6	64.2	1.6
cPSBF	10.8	10.6	6.0	5.8	66.8	0.0
mDIC	76.2	18.4	1.2	2.8	1.4	0.0
mWAIC	67.4	20.4	2.2	3.0	4.2	2.8
mPSBF	74.8	8.6	11.4	3.4	1.8	0.0

Table 3.7: *Simulation study 3: Performance of the Bayesian model selection criteria for Gamma(2,1) for random error and N(0,4) for random effect.*

		Minimum Value			Absolute difference		
Scenario	Criteria	Over	Correct	Under	Over	Correct	Under
I	cDIC	29.6	43.2	27.2	39.8	60.2	0.0
	mDIC	13.0	60.8	26.2	22.4	77.6	0.0
	cPSBF	59.0	28.2	12.8	46.6	52.4	1.0
	mPSBF	11.0	67.4	21.6	44.2	55.8	0.0
	cWAIC	25.4	51.4	23.2	32.6	67.4	0.0
	mWAIC	11.0	62.4	26.6	20.2	79.8	0.0
II	cDIC	18.2	26.4	55.4	38.2	61.8	0.0
	mDIC	18.2	64.4	17.4	15.6	84.4	0.0
	cPSBF	19.2	56.4	37.2	47.2	51.4	1.4
	mPSBF	14.6	70.2	15.2	19.2	78.8	2.0
	cWAIC	15.6	20.4	64.0	32.2	67.8	0.0
	mWAIC	18.2	66.0	15.8	14.4	85.6	0.0

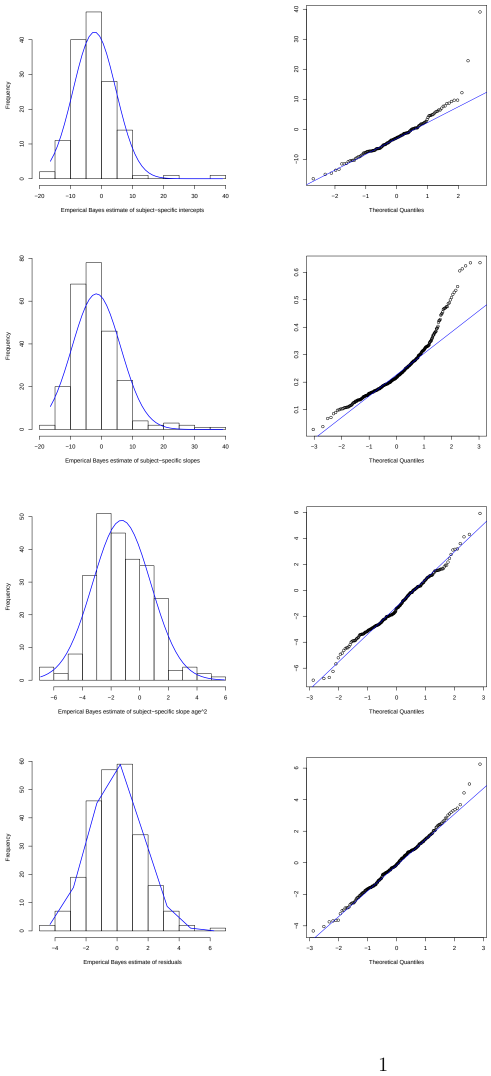

Figure 3.2: *Nigerian indigenous chicken data set: Histogram and normal Q-Q plots for standardised posterior means of random effects based on Model 3.9(a): Subject-specific intercepts in the first row, subject-specific slope of age in the second row, subject-specific slope for the age^2 in the third row and residual in the fourth row.*

Table 3.8: *Nigeria indigenous chicken data set: Posterior mean (regression coefficients) & median (variance parts),* 95% *probability intervals and the conditional and marginal criteria under the four fitted models, see Section 6.6*

	Model 9 a			**Model 9 b**			**Model 9 c**			**Model 9 d**		
	Estimate	2.50%	97.50%	Estimate	2.50%	97.50%	Estimate	2.50%	97.50%	Estimate	2.50%	97.50%
β_0	0.335	0.321	0.349	0.369	0.284	0.848	0.359	0.353	0.374	0.315	0.299	0.329
β_1	-0.008	-0.014	-0.001	-0.009	-0.018	0.000	-0.028	-0.030	-0.021	-0.029	-0.035	-0.023
β_2	0.239	0.229	0.249	0.308	0.227	0.853	0.235	0.231	0.245	0.232	0.221	0.242
β_3	0.031	0.027	0.034	0.046	0.028	0.223	0.031	0.030	0.032	0.030	0.028	0.031
δ_{b1}	-	-	-	-	-	-	0.003	0.001	0.009	0.003	0.000	0.009
δ_{b2}	-	-	-	-	-	-	0.002	0.001	0.007	0.002	0.000	0.007
δ_{b3}	-	-	-	-	-	-	0.002	0.001	0.007	0.002	0.000	0.007
δ_ϵ	-	-	-	0.051	0.048	0.054				0.060	0.055	0.064
$d11$	0.013	0.011	0.015	0.013	0.011	0.319	0.015	0.014	0.017	0.014	0.012	0.016
$d12$	0.010	0.009	0.012	0.010	0.008	0.318	0.007	0.001	0.040	0.008	-0.012	0.031
$d13$	0.001	0.000	0.001	0.000	0.000	0.098	0.005	-0.002	0.023	0.004	-0.019	0.024
$d22$	0.010	0.008	0.011	0.010	0.008	0.383	0.008	0.003	0.122	0.009	0.001	0.085
$d23$	0.002	0.001	0.002	0.002	0.001	0.123	-0.003	-0.011	0.002	-0.003	-0.069	0.002
$d33$	0.001	0.001	0.001	0.001	0.001	0.039	0.008	0.004	0.081	0.006	0.001	0.068
σ_ϵ	0.001	0.001	0.001	0.000	0.000	0.001	0.002	0.002	0.002	0.001	0.001	0.001
cDIC		-19117.4			-19809.10			-19710.85			-18574.70	
cWAIC		-19782.2			-20361.42			-19117.48			-20128.30	
cplppd		-15242.3			-15945.86			-15414.23			-16113.33	
mDIC		-16821.6			-15673.10			-17269.46			-17362.04	
mWAIC		-16808.5			-15472.20			-17488.41			-17511.63	
mlppd		-16665.4			-16965.43			-16765.43			-17165.43	

3.8 Discussion

We have compared three Bayesian selection criteria in the context of LMM for longitudinal data. In addition, we extended these settings to the skew-normal and t(3) distribution for random effects and measurement error. The simulation studies show that the marginal criteria outperform their conditional counterparts. Our results confirm the results of Chan and Grant (2016a) for volatility models, Merkle et al. (2018); Millar (2018); Li et al. (2016) for item response models and Quintero and Lesaffre (2018) in hierarchical models.

It is important to remark that calculating the marginalized criteria does not represent an additional computational effort for LMM since the marginalized likelihood can be written in a closed form at least for a number of important distributions for the random effects and measurement errors. However, for generalized linear mixed models computing the marginalized likelihood is more involved and numerical integration methods are needed (Quintero and Lesaffre, 2018). The performance of the conditional criteria will be examined in a subsequent paper. We examined two selection rules: *minimum value* and *absolute difference* for all criteria. However, our results did not show justification for *absolute difference* outside DIC.

In our simulation study, the performance for the marginalized versions of DIC, WAIC and PSBF is similar. However, in contrast to DIC, WAIC and PSBF have the advantage of being non-invariant to non-linear transformations of the parameters in focus. For this reason, our advice is to base model selection on the marginal versions of WAIC or PSBF.

Nevertheless, our R function computes both the marginal and conditional versions of all three selection criteria with no additional computational efforts. The function can be downloaded from https://ibiostat.be/online-resources/bayesian. Another useful exercise is to evaluate the performance of the selection criteria when varying the vague prior for the covariance matrix of the random effects. This is under current examination.

Chapter 4

Model selection for Bayesian linear mixed models with longitudinal data: Sensitivity to the choice of priors

This chapter has been published as:

Ariyo, O., Lesaffre, E., Verbeke, G., and Quintero, A. (2019). **Model selection for Bayesian linear mixed models with longitudinal data: Sensitivity to the choice of priors**. *Communications in Statistics: Simulation and Computation* 0 (0), 1-25.

Abstract

We explore the performance of three popular Bayesian model-selection criteria when vague priors are used for the covariance parameters of the random effects in a linear mixed effects model (LMM) using an extensive simulation study. In a previous paper, we have shown that the conditional selection criteria perform worse than their marginal counterparts. It is known that for some 'vague' priors, their impact on the estimated model parameters can be non-negligible e.g for the priors of the covariance matrix of the random effects in a longitudinal LMM. We evaluate here the impact of vague priors for the covariance matrix of the random effects on selecting the correct LMM using classical Bayesian selection criteria. We consider marginal and conditional criteria. For the random intercept case, we assign different vague priors to the variance parameters. With two or more random effects, we considered five different specifications of Inverse-Wishart (IW) prior, five different separation priors and a joint prior. The results show again the better performance of the marginal over the conditional criteria and the superiority of joint and separation priors over IW in all settings. We also illustrate the performance of the selection criteria on a practical data set.

4.1 Introduction

The linear mixed-effects model (LMM) is a popular model to analyse longitudinal data with a Gaussian response, especially when the outcomes have been recorded at irregular time points. The model consists of fixed effects and random effects. The fixed effects represent the effect of covariates on the population average, while the random effects represent individual-specific deviations in profiles and account for the correlation among responses from the same individual. Selecting the appropriate LMM implies determining the appropriate fixed effects part and random effects part such that the model fits the current and future data well.

In a Bayesian framework, there is little agreement on the appropriate model selection criteria. Three criteria are currently popular in practice. The deviance information criterion (DIC) is by far the most popular criterion because it can be easily obtained with the popular Bayesian software packages WinBUGS and OpenBUGS. The pseudo-Bayes factor (PSBF) and the widely applicable information criterion (WAIC) are increasingly in use but are not automatically obtained in the classical Bayesian packages, except for WAIC which is provided by Stan (Carpenter et al., 2017). These three model selection criteria may be computed on the hierarchical specification of the LMM, i.e. given the random effects. This then leads to the conditional version of the selection criteria.

However, the marginal version of the LMM, i.e. the model averaged over the distribution of the random effects, can be analytically determined. Selection criteria based on this marginal likelihood are then referred to as marginal selection criteria. It is most popular, but also in general easier, to fit the hierarchical version of the LMM in Bayesian software thereby making use of the data augmentation algorithm. Indeed, the conditional version of DIC is provided by most Bayesian statistical packages, but also for the other selection criteria the conditional version is easy to compute from the generated Markov Chain Monte Carlo (MCMC) samples. Consequently, marginal versions of the selection criteria are basically never reported.

The conditional criteria have, however, been criticized in the literature (Chan and Grant, 2016a; Ariyo et al., 2019b; Merkle et al., 2018). Theoretical arguments and simulation results point out that model selection based on conditional criteria is inferior to model selection based on marginal criteria. This was for instance shown in Ariyo et al. (2019b) for the LMM.
Here, we examine the impact of vague priors on the model parameters on the performance of the model selection criteria. Given the inferior results of the conditional selection criteria, we are particularly interested to see whether the marginal selection criteria highly depend on the chosen vague priors for the model parameters. However, since we realize that the conditional selection criteria will remain popular despite the theoretical and empirical evidence, we also checked the impact of the vague priors on the conditional selection criteria.

While for the fixed effects most often normal priors with a large variance are chosen, there is no standard choice for the vague prior of the variance terms of the random effects in LMMs (Kass et al., 2006). The impact of a vague prior on the posterior distributions can also be more pronounced when the data set is small and/or the number of units contributing to the estimation of the between-unit variation is small (Lambert et al., 2005). In this situation, Lambert et al. (2005) argued that informative prior distributions are required.

When the LMM involves two or more random effects, a prior on their covariance matrix is required. The Inverse Wishart (IW) distribution is the natural choice for a covariance matrix due to its conditional conjugacy. However, problems have been reported with the use of the IW prior as it assumes the same amount of prior information for every variance parameter. More importantly, it assumes a prior relationship between the variances and correlations (Alvarez et al., 2014). These issues have a larger impact when the dimension of the covariance matrix increases.

Several alternative priors for the covariance matrix have been suggested in the literature. Firstly, the IW prior has been given an hierarchical structure. Secondly, various priors have been suggested separating the priors on the variance and correlation parameters. Such separation priors has been shown in the literature to be more efficient than the classical IW prior (Alvarez et al., 2014; Huang et al., 2013). Among the merits of separation priors is their flexibility in incorporating informative prior information. However, things become somewhat more complicated with three or more random effects because certain restrictions must be imposed in order to ensure positive definiteness of the covariance matrix (Hurtado Rúa et al., 2015; Barnard et al., 2000; Wei and Higgins, 2013; Huang et al., 2013). Other priors in which the variance terms of both the measurement errors and the variance-covariance of random effects are modelled jointly have been suggested.These priors have been shown to reduce bias and improve efficiency in the posterior inference (Demirhan and Kalaylioglu, 2015; Kalaylioglu and Demirhan, 2017). Just like for separation priors, certain restrictions are needed to ensure the positive-definite of the covariance matrix.

The aim of this study is to ascertain if the choice of the vague prior, especially on variance and covariance parameters, is important for model selection. More specifically, we wish to measure how much different vague priors impact the marginal selection criteria, but given their popularity we also checked this for the conditional criteria.

The remainder of this article is organized as follow. In Section 4.2 we introduce the Bayesian linear mixed model for longitudinal data. We present the model selection criteria in Section 4.3. In Section 4.4 previous findings are discussed while in Section 4.5 we explore the vague prior for the covariance matrix of the random parameters. In Section 4.6, we assess the sensitivity of different vague covariance priors on random effects on the performance of the above-mentioned model selection criteria using a simulation study. An illustration on a practical data set is shown in Section 4.7. Main conclusions and a discussion are given in Section 4.8.

4.2 The linear mixed-effects model (LMM)

Let $\mathbf{Y}_i = (y_{m_i i} \ldots, y_{m_i,i})^T$ be an m_i-dimensional response vector of (longitudinal) measurements for the i-th (independent) individual, $\mathbf{X}_i$ and $\mathbf{Z}_i$ are $(m_i \times p)$ and $(m_i \times q)$-dimensional covariate matrices, respectively and $\boldsymbol{\beta}$ a p-dimensional vector of fixed effects. The classical LMM is then given as (Laird and Ware, 1982)

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\epsilon}_i, (i = 1, \dots, n), \tag{4.1}$$

with the residual component vector $\boldsymbol{\epsilon}_i \sim N_{m_i}(\mathbf{0}, \boldsymbol{\Sigma}_i)$, where $\boldsymbol{\Sigma}_i$ is an $(m_i \times m_i)$ positive-definite covariance matrix with $\boldsymbol{\Sigma}_i = \sigma_\epsilon^2 \mathbf{I}_{m_i}$ where $\mathbf{I}_{m_i}$ denotes the identity matrix of dimension m_i. The $q \times 1$ random effects vectors are also assumed normally distributed, i.e. $\mathbf{b}_i \sim N_q(\mathbf{0}, \mathbf{D})$, where $\mathbf{D}$ is a $(q \times q)$ positive-definite covariance matrix.

Model (4.1) is called the linear mixed-effects model because it combines the fixed-effects structure $\boldsymbol{\beta}$ with the subject-specific random effects $\mathbf{b}_1, \dots, \mathbf{b}_n$. Inference may be focused on the regression coefficients $\boldsymbol{\beta}$, the unit-specific coefficients $\mathbf{b}_i$ or the variance components ($\boldsymbol{\Sigma_i} = \sigma^2 I_{m_i}$and $\mathbf{D}$). Model (4.1) is the hierarchical version of the LMM, which provides the conditional LMM likelihood.

The marginal version of the LMM is obtained as follows. Let $f(\mathbf{Y}_i|\mathbf{b}_i)$ and $f(\mathbf{b}_i)$ be the (Gaussian) density functions of $\mathbf{Y}_i$ and random effects respectively, then the marginal density function of $\mathbf{Y}_i$ is given by $f(\mathbf{Y}_i|\boldsymbol{\beta}, \sigma^2, \mathbf{D}) = \int f(\mathbf{Y}_i|\mathbf{b}_i, \boldsymbol{\beta}, \sigma^2) f(\mathbf{b}_i|\mathbf{D}) d\mathbf{b}_i$. It can easily be shown that, with Gaussian densities, the marginal version of (4.1), is a multivariate normal distribution given by

$$\mathbf{Y}_i \sim N_{m_i}(\mathbf{X}_i\boldsymbol{\beta}, \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^T + \boldsymbol{\Sigma}_i), (i = 1, \dots, n). \tag{4.2}$$

The Bayesian LMM is obtained when prior distributions are given for all model parameters. Hence, additional to model (4.1) or equivalently model (4.2) we specify priors for $\boldsymbol{\beta}, \mathbf{D}$ and σ^2. Classical choices for these priors are: $\boldsymbol{\beta} \sim N_p(\boldsymbol{\beta}_0, \mathbf{B}_0), \mathbf{D} \sim IW(k, \mathbf{V})$ and $\sigma^{-2} \sim$ Gamma(v_0, δ_0).

Typically, one needs MCMC methods to estimate the model parameters, such as Gibbs sampling or Metropolis-Hastings algorithm (Geman and Geman, 1984; Lesaffre and Lawson, 2012).

4.3 Bayesian model selection

Model selection is an important step in a statistical modelling exercise. In a frequentist context one distinguishes model selection for nested models versus model selection with non-nested models. In the first case formal tests, most often likelihood ratio tests, are used, while in the second case typically information criteria such as Akaike's Information Criterion (AIC) and Bayesian Information Criterion (BIC) are in use. In a Bayesian context, the same model selection criteria apply for nested and non-nested models.

The Deviance Information Criterion (DIC) is an adaptation of AIC to the Bayesian context. DIC is the most popular Bayesian model selection criterion, because it has been implemented in the popular software WinBUGS, and later also in other popular Bayesian software such as OpenBUGS.
However, the literature has been critical about its theoretical foundations. This is sometimes reflected in practice when the associated degrees of freedom, p_{DIC}, is estimated negative thereby making the criterion useless (Spiegelhalter et al., 2014). As a result, there has been increasing interest to use other criteria, such as the pseudo-Bayes factor (PSBF) and the Widely Available Information Criterion (WAIC). WAIC has been recently advocated as having a similar flavor as DIC but with better properties (Watanabe, 2010; Millar, 2018). There is, however, still no consensus about the best criteria for model selection in a Bayesian context.
For reasons of completeness, we will discuss the three most popular Bayesian model selection criteria in more detail.

4.3.1 The deviance information criterion

The deviance information criterion (Spiegelhalter et al., 2002) was developed for Bayesian model selection and is derived from AIC by replacing frequentist concepts by their Bayesian counterparts. As such, DIC expresses the predictive accuracy of the model in a Bayesian way. The frequentist mean is replaced by the posterior mean of the model parameter, i.e. $\bar{\boldsymbol{\theta}} = E(\boldsymbol{\theta}|\boldsymbol{y})$, and frequentist integration is replaced by Bayesian integration. DIC is then defined as

$$\text{DIC} = -2\log p(\mathbf{y}|\bar{\boldsymbol{\theta}}) + 2p_{DIC}, \tag{4.3}$$

where p_{DIC} corresponds to the effective number of parameters, given by

$$p_{DIC} = -2\ E_{\boldsymbol{\theta}|\mathbf{y}}[\log p(\mathbf{y}|\boldsymbol{\theta})] + 2\log[p(\mathbf{y}|\bar{\boldsymbol{\theta}})],$$

which quantifies the number of parameters to be estimated after incorporating the prior information into the model.
From (4.3) it is clear that low values of DIC indicate a better fit of the model to the data. DIC is popular in practice because it is a by-product of the MCMC calculations and implemented in popular Bayesian software. Namely, with the deviance given by $\overline{D(\boldsymbol{\theta})} = -2\log p(\mathbf{y}|\boldsymbol{\theta})$, p_{DIC} and DIC can be approximated by making use of $\boldsymbol{\theta}^1, \ldots, \boldsymbol{\theta}^K$, which are the sampled values of $\boldsymbol{\theta}$ from a converged MCMC chain.

We then have $p_{DIC} = \overline{D(\boldsymbol{\theta})} - D(\bar{\boldsymbol{\theta}})$ and DIC $= D(\bar{\boldsymbol{\theta}}) + 2p_{DIC}$, where $\overline{D(\boldsymbol{\theta})} \approx \frac{1}{K}\sum_{k=1}^{K} D(\boldsymbol{\theta}^k)$ and $D(\bar{\boldsymbol{\theta}}) \approx D(\frac{1}{K}\sum_{k=1}^{K}\boldsymbol{\theta}^k)$. We note that there are different versions of DIC implemented in the popular Bayesian packages, where the difference is primarily due to a different definition of p_{DIC}.

DIC (and p_{DIC}) have received considerable criticism in the statistical literature. First of all, DIC is not invariant to monotonic parameter transformations, i.e. DIC changes value when based on $\boldsymbol{\psi} = h(\boldsymbol{\theta})$ rather than on $\boldsymbol{\theta}$. Furthermore, it has been shown that the asymptotic properties upon which DIC is based, are not fulfilled in hierarchical models (Li et al., 2013b), see also Section 4.4. Note also that it is not clear how to compute DIC when there are missing responses. For this reason, different versions of DIC have been explored in Celeux et al. (2006).

4.3.2 The pseudo Bayes factor

A natural Bayesian selection mechanism is to choose the model with the largest posterior probability. Suppose that there are L models $M_1, \ldots, M_L$ to choose from, with prior probabilities $p(M_1), \ldots, p(M_L)$, respectively. The posterior probability of model M_ℓ is determined by computing the marginal likelihoods $p(\mathbf{y}|M_\ell) = \int p(\mathbf{y}|\boldsymbol{\theta}_\ell, M_\ell)\, p(\boldsymbol{\theta}_\ell|M_\ell)\, d\boldsymbol{\theta}_\ell \quad (\ell = 1, \ldots, L)$ and is given by

$$p(M_\ell|\mathbf{y}) = \frac{p(M_\ell)p(\mathbf{y}|M_\ell)}{\sum_k p(M_k)p(\mathbf{y}|M_k)} = \frac{p(M_\ell)BF_{1,2}[M_\ell : M_b]}{\sum_k p(M_k)BF_{1,2}[M_k : M_b]}, \tag{4.4}$$

where $BF_{1,2}[M_\ell : M_b]$ is the Bayes factor, which compares model M_ℓ to a reference model M_b and is given by

$$BF_{1,2}[M_\ell : M_b] = \frac{p(\mathbf{y}|M_\ell)}{p(\mathbf{y}|M_b)}.$$

The classical Bayes factor is difficult to use in practice because: (1) the marginal likelihood is not defined for improper priors, (2) priors must be well chosen otherwise the classical Lindley-Bartlett paradox (Bernardo, 1980) comes into play and (3) its computation can be very demanding, sometimes even worse than computing the posterior distribution (Lesaffre and Lawson, 2012)[p. 273]. Several versions of the original Bayes factor have been suggested to make the computations feasible and practical. A popular version is the pseudo-Bayes factor (PSBF), where the numerator and denominator in (4.4) are replaced by the product of the marginal likelihoods over all subjects, whereby the marginal likelihood for the ith subject is evaluated in $\mathbf{y}_i$ and is based on the posterior of the model parameters obtained from all other subjects, i.e. from $\mathbf{y}_{(i)}$. This yields a ratio of two pseudo-likelihoods, each being the product of n conditional predictive ordinates (CPOs)(Gelfand and Dey, 1994). The CPO for subject i under model M_ℓ is the probability of observing $\mathbf{y}_i$ given model M_ℓ fitted with all observations in the sample except for $\mathbf{y}_i$ i.e $\mathsf{CPO}_{i,\ell} = p(\mathbf{y}_i|\mathbf{y}_{(i)}, M_\ell)$. The conditional predictive ordinate can be approximated making use of the converged MCMC sample $\boldsymbol{\theta}^1, \ldots, \boldsymbol{\theta}^K$ as follows:

$$\mathsf{CPO}_{i,\ell} \approx \left[\frac{1}{K}\sum_{k=1}^{K}\frac{1}{p(\mathbf{y}_i|\boldsymbol{\theta}_\ell^k, \mathsf{M}_\ell)}\right]^{-1}.$$

To compute the PSBF, the log-pseudo likelihood (LPML) for each model is computed by summing up $\mathsf{CPO}_{i,\ell}$ across the n subjects, i.e, $\mathsf{LPML}_\ell = \sum_{i=1}^n \log(\mathsf{CPO}_{i,\ell})$. Then to compare two models M_1 and M_2, the pseudo-Bayes factor $\mathsf{PSBF}_{1,2}$ favors model M_1 to model M_2 when $\mathsf{PSBF}_{1,2} = \exp(\mathsf{LPML}_2 - \mathsf{LPML}_1) < 1$. Note, that we have adapted the original definition of PSBF in order that small values imply better models. In contrast to DIC, the PSBF is invariant to monotonic parameter transformations.

4.3.3 The widely applicable information criterion

The widely applicable information criteria (WAIC)(Watanabe, 2010) measures the predictive accuracy of the model based on the log-posterior predictive distribution $\log p_{\boldsymbol{\theta}|\mathbf{y}}(\tilde{\mathbf{y}}_i)$ of the parameter vector $\boldsymbol{\theta}$ for a future observation $\tilde{\mathbf{y}}_i$. The predictive accuracy for a future unknown $\tilde{\mathbf{y}}_i$ is expressed by the log-predictive distribution (elpd) as $\mathsf{elpd}_i = E_f[\log p_{\boldsymbol{\theta}|\mathbf{y}}(\tilde{\mathbf{y}}_i)] = \int \log p_{\boldsymbol{\theta}|\mathbf{y}}(\tilde{\mathbf{y}}_i) f(\tilde{\mathbf{y}}_i) d\tilde{\mathbf{y}}_i$, and f is the unknown distribution under the true model. The measure of predictive accuracy can also be described with a point estimate $\bar{\boldsymbol{\theta}}$, often taken equal to $E(\boldsymbol{\theta}|\mathbf{y})$, as the expected log predictive distribution given the point estimator $\mathsf{elpd}_{\bar{\boldsymbol{\theta}}} = E_f(\log p(\tilde{y}|\bar{\boldsymbol{\theta}}))$. The log pointwise predictive distribution (lppd) based on the observed data $\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_n$, is calculated as follows $\mathsf{lppd} = \log \prod_{i=1}^n p_{\boldsymbol{\theta}|\mathbf{y}}(\mathbf{y}_i) = \sum_{i=1}^n \log \int_{\boldsymbol{\theta}} p(\mathbf{y}_i|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta}$. In practice, lppd can be estimated with the converged MCMC sample $\boldsymbol{\theta}^1, \ldots, \boldsymbol{\theta}^K$ from the posterior distribution as $\widehat{\mathsf{lppd}} = \sum_{i=1}^n \log\left[\frac{1}{K}\sum_{k=1}^K p(\mathbf{y}_i|\boldsymbol{\theta}^k)\right]$. The expected log pointwise predictive density elppd is estimated as the log pointwise predictive distribution lppd with a bias correction using the WAIC criterion $\widehat{\mathsf{elppd}}_{WAIC} = \widehat{\mathsf{lppd}} - p_{WAIC}$. The measure p_{WAIC} corresponds to the estimate of the effective number of parameters given by $p_{WAIC} = 2\ \sum_{i=1}^n \left[\log\left(\frac{1}{K}\sum_{k=1}^K p(\mathbf{y}_i|\boldsymbol{\theta}^k)\right) - \frac{1}{K}\sum_{k=1}^K \log\ p(\mathbf{y}_i|\boldsymbol{\theta}^k)\right]$. WAIC can be alternatively expressed as $\widehat{\mathsf{lppd}} = \sum_{i=1}^n \log\left[\frac{1}{K}\sum_{k=1}^K p(\mathbf{y}_i|\boldsymbol{\theta}^k)\right]$. WAIC $=$

$-2\widehat{\text{lppd}} + 2p_{WAIC}$. As for PSBF, WAIC does not change when $\boldsymbol{\theta}$ is replaced by $\boldsymbol{\psi} = h(\boldsymbol{\theta})$, with h a strictly monotone function. As for DIC, smaller values of WAIC indicate a better model.

4.4 Previous findings

In this paper, we evaluate the dependence of vague priors on the performance of Bayesian selection criteria for the linear mixed model. As seen in, e.g. Quintero and Lesaffre (2018), the selection criteria can be based on the hierarchical or conditional version of the LMM given by model (4.1) or on the marginal version of the LMM based on model (4.2). In the first case, one speaks of a conditional selection criterion. We have for the above three popular criteria the conditional DIC (cDIC), the conditional PSBF (cPSBF) and the conditional WAIC (cWAIC). In the second case, we have the marginal versions of the criteria denoted here as: mDIC, mPSBF and mWAIC. It has been argued that the choice of likelihood (conditional or marginal) should be motivated by the aim of the study (Vaida and Blanchard, 2005). For example, in a clinical trial that evaluates a new drug on patients enrolled within an hospital, cDIC may be used to conduct model selection if the interest lies on the efficacy of the new drug at the hospitals of the study. However, mDIC is the appropriate model selection criterion when one wishes to evaluate the efficacy of the new drug in all hospitals.

In the statistical literature, there is evidence of the better performance of the marginal model selection criteria. In a slightly different setting, Chan and Grant (2016a) observed in a simulation study that cDIC usually selects an overfitted model but that mDIC performs better. In general hierarchical models, Quintero and Lesaffre (2018) concluded via a simulation study that mDIC selects (much) more often the correct model than cDIC. They also provided R software to compute the marginal criteria via a dedicated sampling algorithm. The same result was obtained for cWAIC by Millar (2018) and therefore he recommended to use mWAIC. There is also evidence that the same is true for an item response model (Li et al., 2016; Millar, 2018; Merkle et al., 2018). Furthermore, Ariyo et al. (2019b) compared the conditional and marginal versions of DIC, PSBF and WAIC for the Gaussian LMM, the skew-normal LMM (SNLMM) and the skew-t linear mixed model (STLMM) via an extensive simulation study. Both the balanced as well as the unbalanced case was studied for longitudinal data. Both the random intercept case as well as the 2- and 3-dimensional case for the random effects part were considered. The simulation results showed a strong advantage of the marginal criteria in selecting the true data-generating model.

Since the marginal likelihood for the LMM, SNLMM and STLMM have a closed form it is relatively easy to compute these marginal criteria. In addition, the model selection performance of the conditional criteria decreases with increasing sample size (increasing number of random effects), while the performance of the marginal criteria improves with increase in sample size. Furthermore, to facilitate the computations of the marginal criteria in practice, R functions were developed, which can be downloaded from https://ibiostat.be/online-resources/bayesian.
In the course of this study, it was observed that the choice of the prior may affect the ability of the criteria to select the appropriate data-generating model, especially for small sample sizes. The appropriateness of vague priors is most often checked by evaluating their effect on estimation, but here we check their impact on the performance of DIC, PSBF and WAIC in selecting the correct model. More specifically, we wish to check for the conditional and marginal criteria: (1) the impact of vague priors on selecting the best model and (2) whether there is a best vague prior in this context. In principle, we could have limited ourselves to the marginal version of the criteria since the conditional criteria showed repeatedly not to perform well. But, since many will continue to use the conditional criteria because of their practical advantage, the above aims are of practical interest.

4.5 Vague prior distributions for the LMM

An essential step in statistical modelling is the choice of the appropriate statistical model for the data at hand. This is not only an essential step, but also a notoriously complex part of statistical modelling involving statistical tools and substantive knowledge. In this paper we look at model selection in a Bayesian context. That is, we assume that we have a rather limited number of models to choose from. In the model selection step it is customary to choose vague priors for the model parameters. In contrast, informative priors are typically chosen when an appropriate model is already available. For a LMM (vague) priors must be specified for the fixed effects and the variance components. For the fixed effects we have taken vague normal priors. Here we focus on the vague priors for the covariance matrix of the random effects. We consider the univariate case of a random intercept and the multivariate case of several random effects.

4.5.1 Vague priors for the random intercept

Various vague priors have been suggested for the level-2 variance of the Gaussian hierarchical model. This model is a special case of the LMM with only a random intercept. In that case, (4.1) can be written as

$$\boldsymbol{Y}_i \sim N(X_i\boldsymbol{\beta} + \mathbf{1}_{m_i}b_i, \sigma_\epsilon^2) \quad i = 1, \ldots, n. \tag{4.5}$$

with $\mathbf{1}_{m_i}$ is a $m_i \times 1$ vector of ones, and where the random intercept $b_i \sim N(0, \sigma_b^2)$. The improper prior $p(\sigma^2) \propto 1/\sigma^2$, suggested by Jeffreys for the simple case of $N(\mu, \sigma^2)$ yields an improper posterior for model (4.5) if applied to σ_b^2. This was recognized long time ago, see e.g Lesaffre and Lawson (2012). In the early days of the development and use of WinBUGS, this improper prior was replaced by $\sigma_b^2 \sim$ IG$(0.001, 0.001)$, where IG$(\varepsilon, \varepsilon)$ refers to an inverse gamma distribution with two parameters equal to ε. Later on, it was realised that the posterior on σ_b^2 depends much on the choice of the value of ε. This was a trigger to suggest alternative vague but proper priors for σ_b^2. We note that, in contrast to above Jeffreys prior, the proper vague priors depend on the scale of the data. Hence, the vague prior distributions for σ_b listed below are not invariant to change of scale in the data. The following vague but proper priors for σ_b^2 have been considered in the literature:

1. $\frac{1}{\sigma_b^2} \sim$ Gamma$(0.001, 0.001)$. This was a popular prior distribution for variance terms used initially in the WinBUGS Examples I and II documents (Lunn et al., 2000);

2. $\log(\sigma_b^2) \sim$ Uniform$(-10, 10)$. This prior distribution was suggested in the analysis of cluster randomized trials (Spiegelhalter, 2001);

3. $\frac{1}{\sigma_b^2} \sim$ Pareto$(1, 0.001)$. This prior was suggested in genetic epidemiology models (Burton et al., 1999; Scurrah et al., 2000) and is equivalent to Uniform$(0, 1000)$ on the variance scale;

4. $\sigma_b \sim$ Uniform$(0, 100)$. This prior was recommended by Spiegelhalter et al. (2004);

5. $\sigma_b \sim$ half$-$t$(0, 1, 1)$. Gelman (2006) suggested the use of half-t prior with *df=1* (half-Cauchy) on the standard deviation when the number of groups is small. Since a half-t prior appear to be completely harder to work with (Huang et al., 2013), we give the precision parameter a scaled gamma distribution which is equivalent to a half-Cauchy prior (with mean zero) on the standard deviation (Wand et al., 2011).

Note that the above priors are appropriate for the scale of the simulated data, but also for the scale of the data in the analysis of the chicken data set in Section 4.7. Furthermore, the prior for the variance of the measurement error, σ_ϵ^2 is given an IG$(0.001, 0.001)$ prior, which is a classical choice.

4.5.2 Vague priors for the covariance matrix of the random effects

Specifying an appropriate prior for a covariance matrix has been the topic of intensive research in the last two decades. The mathematically convenient prior for a covariance matrix is given by the Inverse Wishart (IW) distribution. This prior is often used in Bayesian modelling for an unknown covariance matrix due to its conditional conjugacy and its implementation in most of the Bayesian statistical software, but there are practical problems with this prior.

In next subsections we review the problems involved with the Inverse Wishart prior, then we discuss some generalizations of this prior to improve convergence properties and its ability to represent (absence of) prior knowledge in an appropriate manner. Note that the same inverse gamma prior for σ_ϵ^2 will be taken as in Section 4.5.1.

The Inverse Wishart prior and variations

The conditional conjugate prior for the covariance matrix $\boldsymbol{D}$ in the linear mixed model (4.1) is the IW distribution (Lesaffre and Lawson, 2012; Schervish, 2012)

$$\boldsymbol{D} \sim IW(k, \boldsymbol{V}),$$

where $\boldsymbol{V}$ is a $q \times q$ positive semi-definite scale matrix and $k(\geq q)$ is the *df*. $\boldsymbol{V}$ is used to position the IW distribution in the parameter space, and k sets the certainty about the prior information in the scale matrix (Hurtado Rúa et al., 2015). For instance, to obtain a minimally informative prior, $k \approx q$ appears appropriate (Gelman et al., 2014; Gelman and Hill, 2007). When $k = q + 1$, the marginal distribution of the correlations is uniform, but their joint distribution is not (Tokuda et al., 2011). Further, the larger k, the more informative is the IW distribution (Gelman et al., 2014; Gelman and Hill, 2007). In JAGS, the standard choice is to take small values for the diagonal elements of $\boldsymbol{V}$ with the degrees of freedom set equal to the dimension of the matrix. However, setting the diagonal elements to larger values also influences the position of the IW (Schnell et al., 2016). In other words, specifying an IW prior distribution requires balancing the size of $\boldsymbol{V}$ and the value of k, but it is not clear how to choose the diagonal elements in $\boldsymbol{V}$. In addition, varies studies have shown that the IW prior is problematic, namely: (1) there is over-dependence in the posterior distribution of the covariance matrix when data is sparse (i.e small number of clusters), (Quintero and Lesaffre, 2017; Gelman, 2006); (2) the uncertainty for all variances is controlled by a single degree of freedom parameter (Gelman et al., 2004b); (3) there is a priori dependence between the standard deviations and the correlation (Tokuda et al., 2011) and (4) the marginal distribution for the variances has low density in a region near zero (Gelman, 2006). In addition, convergence may be difficult with the IW prior. This triggered Gelman et al. (2008) to suggest parameter expansion techniques, which primarily improve the convergence of the MCMC algorithm.

Variations of the classical IW prior have been suggested to improve convergence of the MCMC computations, and to better express (absence of) prior information. O'Malley and Zaslavsky (2008) suggested the scaled Inverse Wishart prior, which is based on the IW prior but with additional parameters to better specify the prior information on the variances. Another variation is suggested by Huang et al. (2013), who suggested an hierarchical Inverse Wishart prior for $\boldsymbol{D}$:

$$\begin{aligned} \boldsymbol{D} \mid d_1, \ldots, d_q &\sim \mathsf{IW}(v + q - 1, 2v\mathsf{diag}(1/d_1, \ldots, 1/d_q)), \\ d_k &\sim \mathsf{IG}(1/2, 1/A_k^2), k = 1, \ldots, q, \end{aligned} \tag{4.6}$$

where diag$(1/d_1, \ldots, 1/d_q)$ denotes a diagonal matrix with $1/d_1, \ldots, 1/d_q$ on the diagonal and v, $A_1, \ldots, A_q$ are positive scalars. The authors showed that (4.6) produces half-t(v, A_k) distributions for each standard deviation of $\boldsymbol{D}$ and that it is a matrix generalisation of the half-t prior of Gelman (Gelman, 2006). Large values of A_k imply a weakly informative prior on standard deviations as in Gelman (2006). Huang et al. (2013) also showed that the choice of $v = 2$ leads to marginal uniform distributions for correlation terms $\rho_{j,k}$, $j \neq k$. This prior will be evaluated in our simulated study and will be referred to as *HIW prior*, more specifically as HIW$(v, \boldsymbol{A})$, with $\boldsymbol{A} = \{A_1, \ldots, A_q\}$. The performance of both variations on the IW prior has been evaluated in a simulation study (Alvarez et al., 2014), who concluded that these priors show good performance and are definitely much better than the classical IW when the true variance is small relative to the prior mean, which holds for larger sample sizes.

Separation strategies for modelling covariance matrices

Another class of priors are based on the separation strategy, first suggested by Barnard et al. (2000). The idea is to decompose the variance covariance matrix $\boldsymbol{D}$ as $\boldsymbol{D} = \boldsymbol{S}^{\frac{1}{2}}\boldsymbol{R}\boldsymbol{S}^{\frac{1}{2}}$, where $\boldsymbol{S}^{\frac{1}{2}}$ is a diagonal matrix with standard deviations as elements and $\boldsymbol{R}$ is a $q \times q$ matrix of correlations. The next two vague priors are based on this separation technique. The two correlation priors will be combined with uniform priors on [0,100] for the elements of $\boldsymbol{S}$, i.e. the variances. In the first proposal, the correlation matrix $\boldsymbol{R}$ is factorised as $\boldsymbol{R} = \boldsymbol{L}^T\boldsymbol{L}$, where $\boldsymbol{L}$ is a $q \times q$ upper-triangular matrix. A prior is then placed on the $q(q+1)/2$ elements in $\boldsymbol{L}$, i.e. the Cholesky factors $L_{ij}\ (i = 1, \ldots, q, i \leq j)$. The following prior ensures unconstrained estimation of variance-covariance matrix and that the positive semi-definite condition is satisfied (Wei and Higgins, 2013):

$$\begin{aligned} \boldsymbol{L}_{1j} &\sim U(-1,1), \\ \boldsymbol{L}_{jj} &= \sqrt{1 - \sum_{i=1}^{j-1} L_{ij}^2}, \\ \boldsymbol{L}_{ij} &= U\left(-\sqrt{1 - \sum_{i=1}^{j-1} L_{kj}^2}, \sqrt{1 - \sum_{i=1}^{j-1} L_{ij}^2}\right), i < j \end{aligned} \tag{4.7}$$

for $j = 2, \ldots, q$, with $L_{11} = 1$ to ensure uniqueness. This prior will be referred to as the *Chol* prior.

Another approach is to use the spherical decomposition of the correlation matrix first suggested by Pinheiro and Bates (1996). In this approach, the Cholesky decomposition is parametrized by sine and cosine functions as follows. Setting $L_{11} = 1$ and let $k = 2, \ldots, q$, we have

$$\begin{aligned} L_{k1} &= \text{cos}(\phi_{k2}), \\ L_{k2} &= \text{sin}(\phi_{k2})\text{cos}(\phi_{k3}), \\ &\vdots \\ L_{k,k-1} &= \text{sin}(\phi_{k2})\text{sin}(\phi_{k3})\ldots\text{cos}(\phi_{kk}), \\ L_{k,k} &= \text{sin}(\phi_{k2})\text{sin}(\phi_{k3})\ldots\text{sin}(\phi_{kk}). \end{aligned}$$

Uniform $(0, \pi)$ priors, with $\pi = 3.1415$, are given to the ϕ_{km} parameters in-order to ensure the uniqueness of the spherical parametrization. Note that the (i, j)th element of $\boldsymbol{R}$ is the inner product $\boldsymbol{L}_i^T\boldsymbol{L}_j$ and $\boldsymbol{L}_k^T\boldsymbol{L}_k = 1$, where $\boldsymbol{L}_k$ is the k^{th} column of $\boldsymbol{L}$. This prior is referred to as *Spherical* prior.

Daniels and Kass (1999) proposed a separation prior that puts a distribution on the correlations so that they will end up shrinking toward 0. To this end, they proposed a normal distribution for the Fisher's *z*-transform on each of the $q(q-1)/2$ correlations ρ: $z(\rho) = \frac{1}{2}\log(\frac{1+\rho}{1-\rho})$. To guarantee a positive define matrix $\boldsymbol{D}$, the foregoing normal distributions on the *z*-transformed correlations needs to be truncated over the relevant values of the correlations (Daniels and Kass, 1999). For a single ρ, assumed here and representing compound symmetry, a half-normal distribution for $z(\rho)$ is assumed. When the correlations are allowed to differ, the constraints to satisfy positive definiteness are more complicated. Further, the authors assigned a prior on the unknown variance σ^2_ρ and flat priors on the diagonal elements of $\boldsymbol{D}$. Christiansen and Morris (1997) considered a hyper-prior on $\sigma^2_{\rho_b}$, with $\pi(\sigma^2_{\rho_b}) \propto (c + \sigma^2_{\rho_b})^{-2}$ and c is a constant that represents a variance. For instance, c can be set to be $\frac{1}{n-3}$, the variance of the *Fisher-z* transformation (Hurtado Rúa et al., 2015). Similar to the approach of Hurtado Rúa et al. (2015), we assigned IG(0.1,0.1) prior for variance parameters and a truncated normal distribution prior for $z(\rho)$. We refer this prior as the *Fisher-z* prior.

Barnard et al. (2000) proposed the separation strategy whereby the $q \times q$ correlation matrix $\boldsymbol{R}$ has a joint uniform distribution on [-1,1]^q. However, the effective algorithm to draw $\boldsymbol{R}$ uniformly is computationally demanding for $q \geq 3$ due to the positive definite constraint. We used here the approach of Tokuda et al. (2011), which is based on the results shown by Joe (2006). He proved that a q-dimensional positive definite correlation matrix $\boldsymbol{R} = (\rho_{ij})_{i,j=1,\ldots,q}$ can be written in terms of the correlations $\rho_{i,i+1}$ and the partial correlations $\rho_{ij;i+1,\ldots,j-1}$ for $(j-1) \geq 2$. These parameters can take independently values in the $[-1, 1]$.

Therefore, he concluded that one can generate a random positive definite correlation matrix by choosing independent distributions $\boldsymbol{F}_{ij}$, $1 \leq j \leq q$ for these parameters (correlations and partial correlations). An appropriate choice for $\boldsymbol{F}_{ij}$ leads to a joint density for $\rho_{ij:1 \leq i < j \leq q}$ that is proportional to $det(\boldsymbol{R})^{\eta-1}$, where $\eta > 0$. When $\eta = 1$, Lewandowski et al. (2009) proved that the marginal distribution of each correlation is a symmetric translated Beta(q/2,q/2)-distribution on the interval [-1,1]. Consequently, the marginal distribution of each correlation becomes more concentrated around zero as q increases in order to satisfy the positive definite constraint. Note also that Joe proved that his algorithm is able to sample from a joint uniform distribution on [-1,1]^q. Tokuda et al. (2011) visualized the implied distribution of $\boldsymbol{D}$ and they observed that for this prior the correlations are a priori independent of the standard deviations. There are several options for the prior distribution on the diagonal elements of $\boldsymbol{D}$ (O'Malley and Zaslavsky, 2008). Here, we assigned Gelman's folded half-t (Gelman et al., 2008) priors for elements of D. In our simulations we considered $q = 1, 2, 3$, then in each case we sampled (each) partial correlation from a translated Beta$(q/2, q/2)$ on [-1,1]. We refer to this prior as the *Partial* prior.

A joint prior for error variance and random effects variance-covariance matrix

Often, priors for error variance and variance-covariance matrix of the random effect are independently modelled. However, it has been shown by Demirhan and Kalaylioglu (2015) and by Kalaylioglu and Demirhan (2017) that a joint prior for these variance terms is more appropriate. Hence, we compare the performance of the conditional and marginal version of the criteria when variance terms are given a joint prior. Kalaylioglu and Demirhan (2017) utilized Cholesky decomposition to separate the random effects variance-covariance $\mathbf{D} = \mathbf{L}\mathbf{L}^T$, where $\mathbf{L}$ is a $q \times q$ lower-triangular matrix. Further, the authors vectorized the diagonal and non-zero off-diagonal matrix $\mathbf{L}$ and the resulting column vectors are denoted by L_1 and L_2, respectively. Additionally, they considered a joint prior distribution for $(L_1^T, L_2^T, \sigma_\epsilon^2, \sigma_b^2)^T)$ if the response variable is continuous and $(L_1^T, L_2^T, \sigma_b^2)^T)$ for a dichotomous /polychotomous response. Furthermore, a multivariate distribution prior was assigned to the vector of log-transformed error variances, log-transformed L_1 and untransformed L_2. For theoretical details of this approach, the reader should consult Demirhan and Kalaylioglu (2015); Kalaylioglu and Demirhan (2017). To ensure positive-definite of $\mathbf{D}$, priors on L_1 need to be positive while L_2 is left unconstrained. The authors' multivariate priors on $\mathbf{D}$ and σ_ϵ^2 are as follows:

$$\Big(\log(L_1), L_2, \log(\sigma_\epsilon^2) \sim F(\delta, v, \lambda, \xi)\Big)^T,$$

and represent the generalized multivariate log gamma (G-MVLG) (Demirhan and Hamurkaroglu, 2011). We refer to this prior as the G-MVLG prior.

4.6 Simulation study

We carried out simulation studies anchored on two longitudinal data sets. Two simulation studies were considered. In the first study, the guiding data set is based on the well-known balanced dental growth study of Potthoff & Roy (Potthoff and Roy, 1964). Measurements were taken on the jaw bi-annually from children between 8 and 14 years of age. The second study is based on the Jimma Infant Survival study, which was designed to evaluate the risk factors affecting infant survival in the Jimma town located in Ethiopia (Lesaffre et al., 1999). This data set is unbalanced due to missing responses, babies that dropped out of the study or died during the study.

In these simulation studies, we used two selection strategies based on (1) *minimum value* and (2) *absolute difference*. For the minimum value strategy, we selected the model having the lowest selection criterion. For the absolute difference strategy, the simplest model was selected when the absolute difference between these models is less than five. This has been suggested in the literature for AIC and BIC, but also for DIC (Lesaffre and Lawson, 2012). We used the same threshold for WAIC and PSBF, however, our previous work (Ariyo et al., 2019b) did not show justification for *absolute difference* outside DIC. Therefore, we report

for WAIC and PSBF only the results using *minimum value*. For both simulation substudies, convergence was evaluated using the Brooks-Gelman-Rubin (BGR) statistic (Brooks and Gelman, 1998; Gelman et al., 1992). All model parameters in the simulation study were estimated based on three chains of 15,000 iterations after discarding the first 7,000 iterations as burn-in. The thinning factor was set at 10. When BGR was larger than 1.1, further sampling was performed until BGR < 1.1. The JAGS code used in this study is provided in the Supporting Materials. Further details on the simulation settings are given below.
The aims of the simulation studies are ultimately to provide practical guidelines. More specifically, we are interested in:

- The impact of the particular choice of the vague prior on the conditional and marginal version of the selection criteria. This is the main aim of this paper;
- Which of the criteria to choose in practice, taking also into account that DIC has some undesirable properties, such as non-invariance to parameter transformations and that sometimes $p_{DIC} < 0$ so that we cannot use DIC in that case;
- The difference in performance of the conditional and marginal version of the criteria. Previously, it has been shown that model selection should be done on the marginal criteria, but it is not immediately clear whether the priors affect the two versions of the criteria equally;
- If the conditional criteria are to be used, whether certain vague priors can still induce good performance of the conditional criteria;
- The impact of the sample size on the above conclusions.

4.6.1 The balanced case: the Potthoff and Roy data set

In the Potthoff & Roy study, changes in pituitary-pterygomaxillary distances during growth of a child were examined at years 8, 10, 12 and 14 on 11 girls and 16 boys who underwent orthodontic treatment. The following LMM was fitted to the data as a function of age and sex (0= girls, 1=boys):

$$Y_{ij} = \beta_0 + \beta_1 \text{sex}_i + \beta_2 \text{age}_{ij} + b_{0i} + \epsilon_{ij}, \quad (i = 1, \ldots, 27; j = 1, \ldots, 4), \quad (4.8)$$

where Y_{ij} is the distance (mm) measure of the ith child at time j, b_{0i} is a random intercept with $b_{0i} \sim N(0, \sigma_b^2)$ and $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$. The following restricted maximum likelihood estimates: $\hat{\beta}_0 = 24.97$, $\hat{\beta}_1 = 1.48$, $\hat{\beta}_2 = -2.32$, $\hat{\sigma}_b^2 = 2.05$ and $\hat{\sigma}_\epsilon^2 = 3.27$ were obtained and used as true parameters in the simulation study. We then considered two scenarios.

- **Scenario I**: We assumed that the random effects structure is known and considered models that differ in the fixed effects part. Besides the true data-generating model (4.8), we considered an overspecified model, which includes an interaction term age*sex and an underspecified model, which omits sex from the model.
- **Scenario II**: We assumed that the fixed structure is known and considered models that differ in the random effects. The overspecified model includes an additional random slope whereas the underspecified alternative ignores the random intercept in the data.

Data generation and prior specifications

Twenty simulation settings were considered for each of the four different sample sizes $n =$ (5, 10, 25, 100) and five signal-to-noise ratios ($\frac{1}{4}$, $\frac{1}{2}$,1,2 and 4 times the residual variance). Each time 500 data sets were generated from model (4.8). For each of the simulation settings, the regression coefficients were given a vague normal prior. Namely, $\beta_j \sim \mathsf{N}(0, 10^6) \quad (j = 0, 1, 2)$. Further, seven prior distributions for variance terms were assigned. Each time, three models (correct, over-and-under specified models) were fitted to evaluate the performance of both the marginal and conditional versions of the Bayesian model selection. We have taken the following vague priors for σ_b:

1. $\frac{1}{\sigma_b^2} \sim$ Gamma(a, a), $(a = 0.001$ and $a = 0.1)$;
2. $\log(\sigma_b^2) \sim$ Uniform(a, b), $(a, b) = (-10, 10)$;
3. $\frac{1}{\sigma_b^2} \sim$ Pareto(a, b), $(a, b) = (1, 0.001)$;
4. $\sigma_b \sim$ Uniform(a, b), $(a, b) = (0, 100)$;
5. $\sigma_b \sim$ half-t$(0, s, 1)$, $s = (1, 0.75)$.

The motivation of the choices of a, b and s is given in Section 4.5.1
We focused on the independent prior distributions in which the variance terms of the random intercept and measurement errors are modelled independently to evaluate the performance of both versions of the criteria. As these priors are commonly used in the literature. The impact of joint prior will be examined in the subsequent section.

Simulation results

Table 4.1 shows the percentage of correct selection for different sample sizes under Scenario I . The results show that the impact of the vague priors on the marginal criteria is minimal, but their impact on the conditional criteria is considerable. This conclusion holds irrespective of the sample size. However, the performance for the conditional criteria improves with increasing sample size. In addition, among the three conditional criteria, DIC is best for higher sample sizes (25 and 100), competing even with the marginal criteria for sample size 100. This in an inconsistent manner. For smaller sample sizes, the half-t prior performed best for the marginal version of all criteria. However, it is not clear which prior outperforms across the different settings and sample sizes. In Table 4.2 the percentages of correct selection for different sample sizes under Scenario II are shown. We observed that a uniform prior for log(variance) performs well for both versions of DIC and PSBF, but the conditional WAIC performs poorly. The poor performance of the conditional WAIC is also seen with the other priors. Again, regardless of the scenario, the marginal criteria outperform the conditional criteria. Their performance increases with sample size while the conditional criteria often select over-specified models (not shown here). However, there is no clear winner among the marginal criteria in this scenario.

4.6.2 The unbalanced case: the Jimma Infant Growth study

The second dataset is obtained from an Ethiopian study designed to evaluate risk factors affecting infant survival (Lesaffre et al., 1999). The growth characteristics of the babies were examined approximately every 60 days, but there were occasional deviations from the planned visits. For the purpose of this analysis, we have taken weight as response with covariates age and sex (0=girls, 1=boys) of the child, and age of the mother at delivery ($agem$). The details of the original analysis can be found in Lesaffre et al. (1999, 2000) where a sample of 495 children was selected to fit the model. This subset will also be the basis for this simulation study. As suggested, Lesaffre et al. (2000) the time variable age was transformed into $\text{newage}_{ij} = \sqrt{\text{age}_{ij}} - (\text{age}_{ij} + 1) - 0.02 \times \text{age}_{ij}$ in model to fit a LMM to the weight profiles. We select our model generating data to be

$$Y_{ij} = \beta_1 + \beta_2 \text{sex}_i + \beta_3 \text{newage}_{ij} + \beta_4 \text{agem}_i + b_{0i} + b_{1i} \times \text{newage}_{ij} + \epsilon_{ij}, \quad (4.9)$$

assuming $(b_{0i}, b_{1i}) \sim N_2(\mathbf{0}, \boldsymbol{D})$ and $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$. The following parameter values were obtained by analysis Jimma data: $\hat{\beta_1} = 2.8581$, $\hat{\beta_2} = 0.1518$, $\hat{\beta_3} = 0.8865$, $\hat{\sigma_\epsilon} = 0.3465$ and $\mathbf{D} = \begin{pmatrix} 0.6813 & -0.0414 \\ -0.0414 & 0.0450 \end{pmatrix}$ where these parameters are used as population parameters for the simulated data set. 500 data sets were generated from model (4.9) with the covariate sex was generated from a Bernoulli distribution with probability of success equal to 0.51, which is the proportion of boys in the data set. The age of the mother was generated from a normal distribution agem$_i \sim N(24.49, 6.29)$ and we have taken $0, 60, 120, \ldots, 360$ days as the moments of measurements. The alternative models considered for each scenario are described below.

- **Scenario I**: We assumed that the random effects structure is known and considered the following models that differ in the fixed part parameters, namely
  - Model (4.9) and including an additional interaction (newage × sex) (overspecified),
  - Model (4.9) but ignoring the sex covariate (underspecified).
- **Scenario II**: We assumed that the covariates in the fixed part are known and considered the following models that differ in the random effects structure, i.e.
  - Model (4.9) and including an additional random slope for newage^2 (overspecified),
  - Model (4.9) but ignoring the random slope for newage (underspecified).

Data generation and prior specifications

With the above specifications for generating data, we considered twenty-four (24) simulation settings for five sample sizes. These settings correspond to twelve different prior choices for the covariance matrix for the two scenarios described above. For each of the settings and sample sizes, we generated 500 datasets from model (4.9). Model (4.9) is then fit using each of the following prior specifications:

- **Prior (1)**: Six specifications of the IW conditional conjugate prior described in Section 4.5.2 of the form $IW(\textit{df}, \mathbf{V})$ using *df*$= q, q+1, q+2$, and $\mathbf{V} = c\boldsymbol{I}_q$ for $c \in \{0.001, 1\}$, where $\boldsymbol{I}_q$ denotes $q \times q$ identity matrix. This is a relatively commonly used informative IW (Schnell et al., 2016)
- **Prior (2)**: Five separation strategies (Section 4.5.2) for covariance matrix $\mathbf{D}$.

- **Prior (3)**: (**G-MVLG**) For joint variance prior $(\log(L_1), L_2, \log(\sigma_\epsilon^2)) \sim$ GMVLG$(0.7, 1.42, \lambda, \xi)$ with $\lambda = (0.3, 0.3, 0.3, 0.4)^T$ and $\xi = (0.25, 0.35, 0.25, 0.1)^T$. This is a non-informative prior and the hyper-parameter values were selected to impose uncertainty on the variance parameters (Kalaylioglu and Demirhan, 2017).

Simulation results

Table 4.3 shows the performance of different specifications of the IW prior to Jimma Infant Survival dataset. Regardless the scenario and sample size, the performance of the conditional and marginal criteria varies with changing *df* and $\mathbf{V}$. We observed that both criteria perform better with a larger value of $\mathbf{c} = 1$ regardless of the value of *df*. The IW prior *df*=*q* and $\mathbf{c} = 1$ performed relatively better in both scenarios.

The performance of the IW prior deteriorates with increasing dimension of the random effects covariance matrix, see also Quintero and Lesaffre (2017). Additionally, to choose an appropriate scale matrix and degrees of freedom is not straightforward, since inconsistent performance is seen. Other disadvantages of the IW prior distribution have been discussed and alternatives proposed (Wei and Higgins, 2013; Barnard et al., 2000; Schuurman et al., 2016; Daniels and Pourahmadi, 2002; Pourahmadi, 1999; Lu and Ades, 2009). Therefore, we considered the effect of some separation priors for the conditional and marginal versions of PSBF, DIC and WAIC.

Table 4.4 shows the performance of both versions of selection criteria using a commonly used IW prior compared with the separation priors and a joint prior for different sample sizes. Since the Cholesky and spherical decomposition performed similarly, the results of the spherical decomposition are omitted here. For both scenarios, and especially for small sample sizes, the joint prior and separation priors outperformed the classical IW prior. In addition, for both versions of the criteria the impact of the sample size is less pronounced with a joint prior and the separation priors than for the IW prior. This result agrees with the conclusion in Alvarez et al. (2014), i.e. that the classical IW prior is less effective when compared with a separation prior.

Further, in Scenario I, the approach based on the *Fisher-z* transformation performed best for both the conditional and marginal versions of the criteria. For scenario II, there is no significant difference between the *HIW* prior based on the approach proposed by Huang et al. (2013) and G-MVLG prior. For both scenarios, the G-MVLG prior and separation priors gave better performance for the conditional criteria when compared with IW prior. Additionally, the impact of sample sizes is less in both G-MVLG and separation prior compared with IW prior.

While there is no best vague prior in both scenarios, we conclude that if the conditional version of the criteria is to be used, then the G-MVLG prior, hierarchical or separation priors are to be used. In fact, the use of IW prior to conditional criteria is strongly discouraged especially for smaller sample sizes. But, again the marginal version of the criteria outperformed the conditional criteria in all scenarios and sample sizes.

Table 4.1: *Potthoff & Roy dataset (Scenario I): Sensitivity of the performance of the conditional and the marginal selection criteria to choose the correct LMM by varying the prior distribution on variance terms for different sample sizes*

		Sample sizes			
Prior	criteria	5	10	25	100
$\frac{1}{\sigma_b^2} \sim \text{Gamma}(0.0001, 0.0001)$	cDIC	26.8	27.2	64.0	79.4
	cPSBF	35.0	38.8	39.2	50.6
	cWAIC	21.4	24.8	62.6	72.4
	mDIC	55.2	60.8	77.4	83.2
	mPSBF	41.4	53.6	75.4	84.2
	mWAIC	46.8	59.4	75.2	83.2
$\frac{1}{\sigma_b^2} \sim \text{Gamma}(0.1, 0.1)$	cDIC	55.4	56.2	64.6	79.8
	cPSBF	55.0	44.8	43.2	43.0
	cWAIC	53.8	55.6	56.8	65.4
	mDIC	56.0	63.6	78.0	83.2
	mPSBF	44.0	56.2	76.6	84.0
	mWAIC	46.8	61.2	75.6	83.2
$\log(\sigma_b^2) \sim \text{Uniform}(-10, 10)$	cDIC	25.0	27.4	67.2	70.0
	cPSBF	36.6	38.4	45.8	46.8
	cWAIC	33.2	38.4	62.0	68.8
	mDIC	58.8	61.6	78.0	82.8
	mPSBF	40.0	54.6	76.4	83.8
	mWAIC	48.8	57.6	75.8	82.6
$\log(\sigma_b^2) \sim \text{Uniform}(0.001, 100)$	cDIC	46.2	53.0	68.0	78.8
	cPSBF	43.8	45.0	46.8	48.0
	cWAIC	53.4	48.0	66.6	70.2
	mDIC	57.8	60.8	77.8	83.6
	mPSBF	40.0	58.0	75.6	84.6
	mWAIC	49.0	58.2	75.6	83.8
$\frac{1}{\sigma_b^2} \sim \text{Pareto}(1, 0.0001)$	cDIC	32.8	52.4	69.2	78.8
	cPSBF	33.4	44.4	43.6	42.4
	cWAIC	42.2	50.0	64.4	73.4
	mDIC	51.2	54.0	75.6	83.0
	mPSBF	40.8	51.2	76.0	83.0
	mWAIC	46.4	51.4	74.0	82.6
$\sigma_b \sim \text{Uniform}(0, 100)$	cDIC	31.6	46.6	69.4	79.4
	cPSBF	43.0	42.6	43.4	42.8
	cWAIC	41.4	45.6	61.6	71.8
	mDIC	55.0	56.4	77.6	83.2
	mPSBF	28.2	53.8	76.8	84.8
	mWAIC	33.8	55.6	75.8	83.8
$\sigma_b \sim \text{t}(0, 0.75, 1)$	cDIC	41.2	63.0	64.0	71.0
	cPSBF	31.8	43.7	52.0	64.8
	cWAIC	40.0	67.4	69.0	70.2
	mDIC	68.5	76.8	79.6	84.2
	mPSBF	61.4	74.8	75.2	80.6
	mWAIC	66.2	74.0	76.2	81.0

Table 4.2: *Potthoff & Roy dataset (Scenario II): Sensitivity of prior distribution on variance terms on selecting the correct model (%) for different sample sizes and criteria DIC,PSBF,WAIC evaluated on conditional and marginal version of LMM*

		Sample sizes			
Prior	criteria	5	10	25	100
$\frac{1}{\sigma_b^2} \sim \text{Gamma}(0.0001, 0.0001)$	cDIC	27.8	42.6	50.8	44.2
	cPSBF	32.8	39.2	46.8	57.8
	cWAIC	22.6	37.6	46.6	43.0
	mDIC	41.0	54.0	82.4	82.6
	mPSBF	37.2	51.2	80.2	80.2
	mWAIC	42.6	57.4	76.8	81.2
$\frac{1}{\sigma_b^2} \sim \text{Gamma}(0.1, 0.1)$	cDIC	54.6	50.6	52.2	49.2
	cPSBF	54.2	59.4	56.6	51.2
	cWAIC	46.4	40.0	46.0	28.4
	mDIC	52.0	72.8	81.4	84.6
	mPSBF	64.6	79.0	80.8	84.4
	mWAIC	59.8	76.6	77.8	82.6
$\log(\sigma_b^2) \sim \text{Uniform}(-10, 10)$	cDIC	53.2	55.2	46.6	41.0
	cPSBF	45.4	46.4	44.6	46.2
	cWAIC	48.8	45.0	51.4	48.8
	mDIC	40.8	66.6	82.4	82.8
	mPSBF	51.0	63.4	80.2	81.6
	mWAIC	52.6	68.0	79.0	80.6
$\log(\sigma_b^2) \sim \text{Uniform}(0.001, 100)$	cDIC	67.0	77.2	83.8	97.4
	cPSBF	66.8	80.6	77.0	94.4
	cWAIC	54.8	53.6	39.8	37.4
	mDIC	65.6	89.0	100.0	100.0
	mPSBF	75.0	92.6	100.0	100.0
	mWAIC	67.4	90.8	100.0	100.0
$\frac{1}{\sigma_b^2} \sim \text{Pareto}(1, 0.0001)$	cDIC	59.2	68.2	56.8	40.6
	cPSBF	61.4	66.4	52.6	47.0
	cWAIC	53.4	41.6	29.0	19.0
	mDIC	45.0	76.6	86.0	85.8
	mPSBF	40.0	87.4	86.0	85.2
	mWAIC	48.8	84.4	84.4	83.8
$\sigma_b \sim \text{Uniform}(0, 100)$	cDIC	61.4	62.2	53.8	42.2
	cPSBF	59.2	53.0	46.8	44.0
	cWAIC	51.4	42.0	36.6	35.0
	mDIC	48.8	76.4	84.8	85.6
	mPSBF	61.0	82.4	85.2	84.0
	mWAIC	50.4	80.4	81.6	83.4
$\sigma_b \sim \text{t}(0, 0.75, 1)$	cDIC	40.2	56.0	55.8	61.4
	cPSBF	41.2	43.7	47.4	50.0
	cWAIC	41.2	43.8	57.4	60.4
	mDIC	71.2	85.8	87.4	92.6
	mPSBF	70.6	84.4	86.4	86.9
	mWAIC	70.4	81.4	86.2	91.8

Table 4.3: *Jimma Infant Survival dataset (Scenario I ($q = 2$) and Scenario II ($q = 3, 2$ and 1): Performance of Bayesian model selection with six specifications of Inverse Wishart conjugate prior for over-specified, correct and under-specified respectively.*

					Criteria			
Sample size	Scenario		cDIC	cPSBF	cWAIC	mDIC	mPSBF	mWAIC
10	I	$df = q, \mathbf{V} = 0.001$	49.0	47.8	41.8	59.0	59.0	56.8
		$df = q, \mathbf{V} = 1$	62.8	44.6	58.0	82.8	81.8	89.8
		$df = q+1, \mathbf{V} = 0.001$	46.6	37.4	37.6	58.4	59.4	54.8
		$df = q+1, \mathbf{V} = 1$	60.4	47.8	55.0	80.0	77.4	76.8
		$df = q+2, \mathbf{V} = 0.001$	51.8	42.2	41.6	58.2	63.2	55.8
		$df = q+2, \mathbf{V} = 1$	57.6	43.4	57.5	79.0	76.0	53.3
	II	$df = q, \mathbf{V} = 0.001$	61.0	52.6	63.2	63.2	69.4	68.2
		$df = q, \mathbf{V} = 1$	62.4	65.6	45.8	86.0	88.6	97.0
		$df = q+1, \mathbf{V} = 0.001$	58.0	44.4	58.2	57.4	52.8	61.0
		$df = q+1, \mathbf{V} = 1$	59.2	67.8	45.6	86.6	88.4	87.2
		$df = q+2, \mathbf{V} = 0.001$	60.4	53.0	57.8	58.0	55.3	62.4
		$df = q+2, \mathbf{V} = 1$	60.2	62.4	46.8	87.2	88.8	86.2
50	I	$df = q, \mathbf{V} = 0.001$	54.0	50.0	52.0	68.0	70.0	68.0
		$df = q, \mathbf{V} = 1$	68.8	72.4	67.8	90.0	90.0	90.0
		$df = q+1, \mathbf{V} = 0.001$	64.4	72.4	67.8	76.4	77.6	76.0
		$df = q+1, \mathbf{V} = 1$	69.4	70.6	68.4	90.0	90.0	90.0
		$df = q+2, \mathbf{V} = 0.001$	77.0	71.4	77.8	79.0	86.6	79.0
		$df = q+2, \mathbf{V} = 1$	68.8	73.0	77.8	90.0	90.0	90.0
	II	$df = q, \mathbf{V} = 0.001$	52.0	49.0	53.0	66.0	68.0	62.0
		$df = q, \mathbf{V} = 1$	53.6	47.8	56.0	80.4	76.4	79.6
		$df = q+1, \mathbf{V} = 0.001$	60.0	44.2	55.4	66.4	71.0	65.4
		$df = q+1, \mathbf{V} = 1$	64.2	42.4	57.4	82.0	75.8	79.0
		$df = q+2, \mathbf{V} = 0.001$	57.8	49.0	78.2	65.2	69.6	64.6
		$df = q+2, \mathbf{V} = 1$	62.0	40.8	51.6	79.6	74.6	75.8

Table 4.4: *Jimma Infant survival dataset: Sensitivity of the performance of the conditional and marginal selection criteria to choose the correct LMM by using separation priors, a joint prior with an IW prior for different sample sizes.*

Scenario		I					II				
Prior		**10**	**25**	**50**	**100**	**200**	**10**	**25**	**50**	**100**	**200**
IW	cDIC	31	44	48	66	63	44	52	58	53	55
	cPSBF	36	40	46	68	69	42	39	37	42	46
	cWAIC	35	42	41	65	69	42	53	58	51	53
	mDIC	52	52	66	72	75	54	63	67	70	72
	mPSBF	52	50	67	72	75	54	54	68	73	73
	mWAIC	53	54	66	73	75	56	62	67	71	72
Chol	cDIC	60	61	61	56	52	70	72	79	78	78
	cPSBF	55	58	59	61	59	69	67	67	70	70
	cWAIC	59	61	60	52	50	66	68	72	76	78
	mDIC	89	89	88	92	96	100	100	100	100	100
	mPSBF	70	71	76	84	87	98	100	100	100	100
	mWAIC	79	79	80	83	87	98	100	100	100	100
HIW	cDIC	60	62	63	64	67	82	84	81	88	88
	cPSBF	60	60	66	66	67	72	76	71	78	77
	cWAIC	61	62	62	63	66	82	83	82	85	88
	mDIC	76	80	82	82	100	98	100	100	100	100
	mPSBF	74	81	85	84	100	96	100	100	100	100
	mWAIC	75	81	85	86	100	99	100	100	100	100
Partial	cDIC	66	61	64	62	61	54	72	62	68	66
	cPSBF	59	61	62	61	64	59	66	63	60	60
	cWAIC	64	64	64	66	67	51	69	63	67	69
	mDIC	76	84	82	86	84	73	84	86	87	80
	mPSBF	75	85	83	86	85	71	80	84	87	81
	mWAIC	83	83	79	86	85	69	80	85	86	80
Fisher-z	cDIC	74	75	80	79	82	71	69	71	80	87
	cPSBF	70	70	73	77	83	61	70	74	77	80
	cWAIC	67	70	74	76	78	67	66	68	77	83
	mDIC	74	79	83	100	100	70	82	98	100	100
	mPSBF	66	71	79	100	100	73	84	94	100	100
	mWAIC	75	74	81	100	100	67	80	96	100	100
G-MVLG	cDIC	62	69	64	66	67	71	86	79	89	89
	cPSBF	64	65	62	67	68	70	79	73	86	97
	cWAIC	61	68	63	68	67	69	85	69	86	87
	mDIC	69	73	80	88	100	100	100	100	100	100
	mPSBF	70	70	81	89	99	100	100	99	100	100
	mWAIC	71	72	83	89	100	99	100	100	100	100

4.7 Analysis of the longitudinal evolution of Nigerian chickens

We analysed of the Nigerian indigenous chicken (NIC) data set and evaluated the sensitivity of separation priors and classical IW prior on the covariance matrix. These data concern the longitudinal evolution of body weight (BW) of chickens of different breeds raised in a university experimental farm. Four hundred and sixteen chickens were measured every week (age) from hatching up to twenty weeks to evaluate the growth of two progenies (breeds) of chicken. A first analysis can be found in Ariyo et al. (2019b). We refer to Adeleke et al. (2011) for the rationale for the study and the experimental design. Figure 4.1a shows the evaluation of weight of the chicken and the average profile over time. The deviations between the observed chickens' body weight and the mean structure are presented in Figure 1b. It will be assumed that

$$Y_{ij} = \beta_0 + \beta_1 breed_i + \beta_2 age_{ij} + b_{0i} + b_{2i} age_{ij} + \epsilon_{ij}, \tag{4.10}$$

where Y_{ij} is the chicken body weight (kg); breed$_i$ is the breed indicator (1=pure breed, 2=cross breed), age$_{ij}$ represents the age (standardised). We limit the chicken's age to 13 weeks since a considerable amount of chicken died after this age.
We fitted the following alternative models:

- Model 1: Linear model in fixed effects and linear in random effects
- Model 2: Quadratic model in fixed effects and linear in random effects
- Model 3: Linear model in fixed effects and quadratic in random effects
- Model 4: Quadratic in fixed effects and quadratic in random effects
- Model 5 : Cubic in fixed effects and cubic in random effects.

The classical IW prior together with two separation (*Fisher-z* and *Chol*) priors and a Hierarchical prior (*HIW*) discussed in Section 4.5.2 were used for the covariance matrix.
Table 4.5 shows that there is some discrepancy in both the marginal and conditional criteria using different priors. The conditional DIC and WAIC select Model 2 using the IW prior. Contrary, the conditional PSBF as well as the marginal version of the criteria selection Model 3. This shows inconsistency in model selection among the conditional criteria when IW prior is used. However, both the models selected by these criteria seem to be incorrect as the average growth curve of the chicken seems quadratic and the individual growth curves differ from the average curve in a quadratic manner (see Figure 4.1). In contrast, all the separation priors as well as joint prior support Model 4 (i.e the presence of quadratic terms in both fixed and random effects) which appears to be the appropriate model here based on Figure 4.1. This confirmed the results of the simulation that separation priors are more efficient than the IW prior.

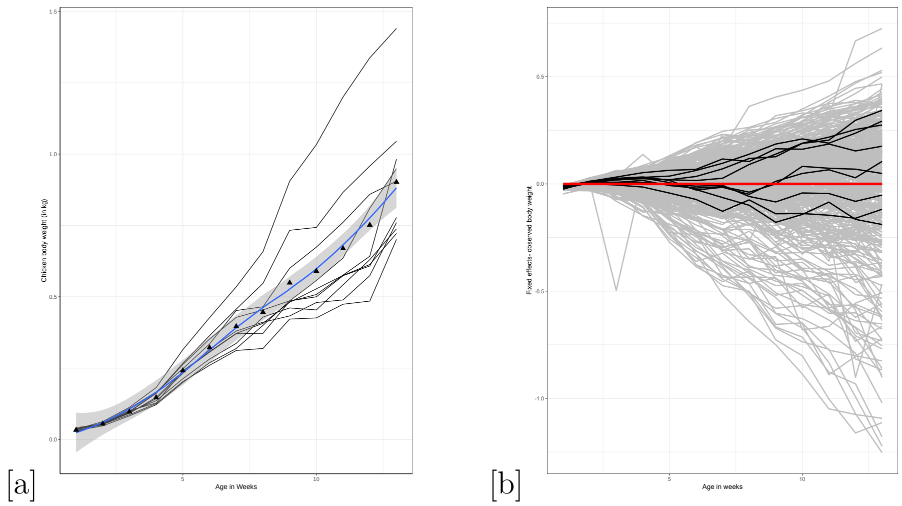

Figure 4.1: *Nigerian indigenous chicken data set: (a) individual and average profiles of 10 randomly selected chickens' body weight obtained by locally weighted regression using ggplot2: (b) the deviation of the 10 randomly selected chickens' body weight from the mean structure.*

Table 4.5: *Nigeria indigenous chicken data set: Sensitivity of the performance of the conditional and marginal selection criteria using separation priors and joint prior with an IW prior.*

	Criteria	IW	HIW	Fisher-z	Chol	G-MVLG
Model 1	cDIC	-15129.8	-14791.2	-15191.6	-15189.5	-14901.8
	cWAIC	-15497.7	-15334.3	-15478.1	-15470.4	-15014.2
	clpml	-13938.7	-12995.1	-14058.4	-14073.7	-14913.1
	mDIC	-13648.5	-12768.9	-13695.0	-12376.1	-13012.0
	mWAIC	-13639.5	-12759.1	-13685.7	-12358.9	-13085.8
	mlpml	-13609.6	-12727.2	-13653.9	-12333.5	-13043.4
Model 2	cDIC	-16726.7	-16855.0	-16755.0	-16755.0	-16045.7
	cWAIC	-17110.1	-17089.3	-17049.3	-17049.3	-16042.3
	clpml	-15401.3	-15576.0	-15516.0	-15516.0	-16519.1
	mDIC	-15093.9	-15131.8	-15121.8	-15121.8	-15501.8
	mWAIC	-15079.7	-15117.1	-15107.1	-15107.1	-15410.9
	mlpml	-15036.7	-15090.6	-15060.6	-15060.6	-15462.2
Model 3	cDIC	-16095.7	-19095.7	-18709.6	-18702.6	-18612.3
	cWAIC	-16776.4	-19776.4	-19485.6	-19482.6	-19810.8
	clpml	-17114.1	-17914.1	-16117.7	-16111.7	-16132.0
	mDIC	-16509.1	-16579.1	-15365.3	-15365.3	-15369.0
	mWAIC	-16505.7	-16565.7	-15351.6	-15351.6	-15350.9
	mlplm	-16504.3	-16524.3	-15310.2	-15310.2	-15320.7
Model 4	cDIC	-16186.8	-19476.8	-18796.3	-19492.1	-20047.4
	cWAIC	-16851.8	-20034.2	-19533.5	-20026.3	-20126.8
	clplm	-16777.5	-17609.5	-16193.5	-17593.9	-17784.0
	mDIC	-16104.2	-16788.4	-16632.9	-16878.2	-16897.4
	mWAIC	-16314.9	-16770.5	-16618.3	-16663.6	-16709.3
	mlpml	-16466.7	-16719.0	-16572.4	-16601.3	-16700.4
Model 5	cDIC	-16176.8	-16676.8	-16476.8	-16476.8	-16421.0
	cWAIC	-16034.2	-17434.2	-17034.2	-17034.2	-17114.6
	clppd	-16609.5	-17609.5	-17609.5	-17609.5	-17709.1
	mDIC	-16108.4	-16208.4	-16498.4	-16738.4	-16715.0
	mWAIC	-16400.5	-16200.5	-16470.5	-16620.5	-16631.3
	mlpml	-16509.0	-16309.0	-16469.0	-16710.0	-16731.8

4.8 Conclusion

We have performed simulation studies to determine if the choice of the vague prior for the variance or covariance matrix of the random effects in a longitudinal study is of great importance in model selection. In addition, we assessed whether different vague prior distributions have a different effect on the conditional and marginal version of DIC, PSBF and WAIC. We made use of vague priors that were proposed in the literature. While the considered scenarios are still somewhat limited in scope, the performance of the criteria in our simulation study allows already for some clear conclusions.

The results can be broadly summarized as follows for the variance of the random intercept. The choice of the vague prior impacted both versions of the criteria but the impact is much less for the marginal version than for the conditional version of the criteria. In addition, the conditional criteria performed in an inconsistent manner often selecting over-specified models while the marginal version of the criteria showed much less dependence to the choice of parameter values of the prior and often selected the correct model. For longitudinal mixed models that involve two or more random effects, the joint prior, the hierarchical prior and the separation priors all outperformed the classical IW prior. These priors are also the choice when the conditional version of the criteria are to be taken. We noted, to our surprise, that cWAIC was significantly poorer in some cases than the two other criteria.

Finally, we believe that a sensitivity analysis is necessary when using prior distributions that are intended to be vague for the level 2 variance parameters. This is especially important for small sample sizes. For models with more than one random effect, the joint prior, the hierarchical prior and separation priors are to be chosen for both the conditional and marginal versions of the criteria. For large sample sizes, the classical IW prior can still be used for model selection for computational convenience. Finally, the marginal version of the criteria outperformed the conditional version of the criteria, as was earlier recommended in the literature, (see Chan and Grant, 2016a; Li et al., 2016; Quintero and Lesaffre, 2018; Merkle et al., 2018; Millar, 2018; Ariyo et al., 2019b). We have added evidence to this recommendation in the context of longitudinal mixed models, which constitutes an important class of models in biomedical research.

Chapter 5

Bayesian model selection for longitudinal count data

This chapter has been submitted as:

Ariyo,O., Lesaffre, E., Verbeke, G., and Quintero, A. (2020). **Bayesian model selection for longitudinal count data**. *Communications for Statistical Applications and Methods*.

Abstract

We have compared three popular Bayesian model selection criteria for generalised linear mixed-effects models (GLMMs) for longitudinal count data. Two versions of these criteria can be used: (1) the conditional version, computed given the random effects and (2) the marginal version, based on the likelihood averaged over the random effects. Despite the theoretical evidence that the marginal criteria are more appropriate, applied statisticians most often use the conditional criteria due to their availability in most software but also because the marginal criteria are not available in closed form. We have written an R function that computes the marginal model selection criteria for GLMMs, especially for longitudinal Poisson models and their extensions, based on samples from the posterior distribution. We illustrate via a simulation study that the marginal criteria are to be preferred. Our procedure involves extra sampling on top of Markov chain Monte Carlo sampling. Therefore, we examined the amount of additional sampling needed to obtain a stable estimate of the marginal criteria. Finally, we illustrate the advantages of the marginal criteria on a public data set of patients who have epilepsy.

5.1 Introduction

In a longitudinal study, subjects are monitored over time. Such a study type allows to discover baseline or time-varying characteristics that have an impact on the outcome of interest. Generalised linear mixed models (GLMMs) are one of the most popular tools to analyse various types of outcomes (continuous, binary, counts) measured repeatedly over time. The GLMM (McCullagh, 1989) is a generalisation of the linear mixed model including both fixed and random effects with a response having a distribution in the exponential family. In the frequentist approach, the model parameters are estimated by integrating out the random effects from the likelihood. Most often, this is done under the assumption of Gaussian random effects. The integral is then evaluated using non-adaptive or adaptive Gaussian quadrature methods. In contrast, in the Bayesian approach, the random effects are most often estimated together with the fixed effects. This implies that Bayesian computations are based on the conditional likelihood, which is the likelihood of the data given the random effects.

To find an appropriate GLMM for a (longitudinal) data set, one makes use in the frequentist approach of the likelihood ratio test for nested models or information criteria, such as AIC and BIC, for non-nested models. In the Bayesian approach, the same model selection criteria are used for both nested and non-nested models. One of such criteria to select between two models is Bayes' factor (Kass and Raftery, 1995), defined as the ratio of the marginal likelihoods (marginalised over the prior of the model parameters) of the two competing models. While the Bayes' factor is an elegant Bayesian tool, there are serious issues with its computation in practice. Namely, it turns out that computing the Bayes' factor proved to be at least as difficult as computing the posterior distribution, cannot be computed with improper priors and is quite sensitive to the choice of the prior distribution. To overcome this problem, the pseudo-Bayes factor (PSBF) (Gelfand and Dey, 1994) has been suggested. To compute the PSBF one updates an (improper) prior to a proper posterior and calculates the Bayes' factor using the generated posterior as prior.

The most popular Bayesian model selection criterion is the Deviance Information Criteria (DIC) (Spiegelhalter et al., 2002). The DIC aims to estimate the predictive ability of the fitted model to future samples from the same population, and like AIC and BIC, it represents a trade-off between the model fit and model complexity. However, the theoretical basis of DIC is not clear, and several objections and alternatives have been formulated by the discussants of Spiegelhalter et al. (2002), see also Celeux et al. (2006) and Spiegelhalter et al. (2014). Recently, Watanabe's Widely Applicable Information Criterion (WAIC) (Watanabe, 2013) has been proposed. WAIC has been singled out as a worthy successor of DIC (Spiegelhalter et al., 2014). We consider PSBF, DIC and WAIC in this paper since there is little agreement in the statistical literature on the choice of these criteria for model selection.

Model selection criteria may be based on the conditional likelihood (given the random effects) resulting in conditional criteria or on the marginal likelihood (integrating out the random effects) resulting in the marginal criteria. The conditional criteria measure the predictiveness of the model for the subjects included in the current study, whereas the marginal criteria measure the predictiveness of the model for all subjects from the same population in a future study. Vaida and Blanchard (2005) pointed out that the choice of the criteria should be motivated by the research question. This implies that most often the marginal criteria should be used in practice, definitely for longitudinal studies. However, irrespective of that research question, the conditional criteria are most often used in practice because of convenience and their easy availability in most software. This usage has been questioned for (extended) LMMs (Ariyo et al., 2019b,a) as well as for GLMMs (Millar, 2009; Christensen, 2017; Quintero and Lesaffre, 2018; Merkle et al., 2018). Ariyo et al. (2019b) and Ariyo et al. (2019a) explored the performance of the marginal model selection criteria for the LMM and concluded their superior performance over the corresponding conditional criteria. An R program has been written that computes the marginal and conditional versions of PSBF, DIC and WAIC for any LMM based on MCMC output from a fitted model. However, for a GLMM, there is no closed-form for the likelihood. Hence, there is the need for an approach to computing the marginal model selection criteria for non-closed-form likelihoods such as for GLMMs.

Numerical methods have been developed that compute the marginal criteria for non-closed form likelihoods. For example, Chan and Grant (2016b) proposed fast algorithms for computing the marginal DIC (mDIC) for a variety of high dimensional latent variable models and show that mDIC has much smaller numerical standard errors compared to the DIC based on the conditional likelihood (cDIC). Likewise, Chan and Grant (2016a) proposed importance-sampling algorithms for computing mDIC under a variety of stochastic volatility models. In the INLA package, developed by Rue et al. (2009) for latent Gaussian models, the marginal posterior is computed by integrated nested Laplace approximations. In that case, mDIC is derived directly from these approximate marginal likelihoods. Up to now, all these methods make use of a Gaussian assumption for the random effects. In this paper, we consider a general method for computing the marginal criteria in a random-effects model for count data. The computational procedure is based on the replication sampling approach in combination with importance sampling. The strategy (for GLMM) is similar to that of Quintero and Lesaffre (2018) who generalised the methodology of Chan and Grant (2016a).

This paper aims to show the superiority of the marginal criteria for Poisson mixed-effects models, which are a special case of a GLMM. Especially in some practical settings such as (i) when there is overdispersion in the counts and/or too many zeros and (ii) when the number of repeated measurements is relatively small to the number of independent variables. Overdispersion has been shown to produce problematic results (see for example Chen and Wehrly, 2016; van Smeden et al., 2016, 2019) and to affect the performance of the selection criteria (Fitzmaurice, 1997; Howe et al., 2019). We have written an easy-to-use R function that computes the marginal selection criteria for GLMMs via sampling techniques to promote their usage among practitioners. Finally, we illustrate these procedures in simulation studies motivated by a well-known data set of patients suffering from epilepsy.

The structure of the paper is as follows. Section 5.2 presents the general GLMM and introduces the Poisson mixed-effects model. In Section 5.3, we discuss the conditional and marginal selection criteria in generality. Section 5.4 presents and evaluates the sampling methods for the computation of marginal criteria of a GLMM. In Section 5.5, we discuss extensions of the Poisson mixed-effects model that deal with overdispersion in the repeated counts. In Section 5.6, our approach is illustrated in the well-known longitudinal epilepsy data set. Different simulation settings and scenarios are presented in Section 5.7. In the same section, we compare the performance of the conditional and marginal model selection criteria and evaluate the performance of the sampling techniques in computing the marginal criteria. The article concludes with a general discussion in Section 5.8.

5.2 Generalised linear models with cluster-specific effects

5.2.1 The generalised linear model

A random variable Y follows a distribution in the exponential family if its density is of the form

$$f(y) \equiv f(y \mid \lambda, \phi) = \exp\left\{\phi^{-1}[y\lambda - \zeta(\lambda)] + c(y, \phi)\right\},$$

where λ and ϕ are termed "natural (canonical) parameter" and "dispersion parameter", respectively for unknown functions $\zeta(\cdot)$ and $c(\cdot, \cdot)$. It is is well-known that

$$E(Y) = \mu = \zeta^{'}(\lambda), \tag{5.1}$$

and

$$Var(Y) = \sigma^2 = \phi\zeta^{''}(\lambda). \tag{5.2}$$

This implies that the mean and variance are related through $\sigma^2 = \phi\zeta''[\zeta'^{-1}(\mu)] = \phi\nu(\mu)$, with variance function $\nu(\cdot)$ describing the mean-variance relationship. Suppose that for the ith subject $(i = 1, \ldots, n)$ a covariate vector $\mathbf{x}_i$ is available and that given $\mathbf{x}_i$, the response Y_i of that subject has the above exponential distribution with mean μ_i and that $\eta(\mu_i) = \mathbf{x}_i^T\boldsymbol{\beta}$. Then the above defines a generalised linear model for the response, denoted as GLM.

For some GLMs such as the binomial, Poisson and exponential distributions the mean and variance parameters are forced to depend on a single parameter. However, in some applications, this assumption may be overly restrictive. A number of extensions to the Poisson model have been proposed by Hinde and Demétrio (1998); Breslow (1984); Lawless (1987) and Molenberghs and Verbeke (2005) that accommodate overdispersion, i.e. when the variance of the counts (much) exceeds their mean. Note that one way to deal with overdispersion is to allocate an overdispersion parameter $\phi \neq 1$ so that (5.2) produces $Var(Y) = \phi\nu(\mu)$.

This leads to a quasi-likelihood approach, see Molenberghs et al. (2007). Here we consider parametric generalisations of the Poisson model.

5.2.2 The generalised linear mixed model

The generalised linear mixed model extends the GLM by adding random effects, and thereby becomes a tool to analyse non-Gaussian repeated measurements, see e.g. Verbeke and Molenberghs (2000). Let Y_{ij} be the jth outcome $(j = 1, \ldots, m_i)$ measured on the ith subject $(i = 1, \ldots, n)$, then a GLMM for Y_{ij} is defined as a GLM conditional on random effects.
More specifically, we assume that, in analogy with Section 5.2.1, conditionally upon a $q-$dimensional random-effect $\mathbf{b}_i \sim N_q(\mathbf{0}, \mathbf{D})$, where $N_q(\cdot)$ is a q-variate variate normal distribution with mean vector $\mathbf{0}^T = (0, \ldots, 0)$, of dimension $q \times 1$, and $\mathbf{D}$ is a $q \times q$ positive-definite covariance matrix. The outcomes Y_{ij} are distributed independently with densities of the form

$$f_i(y_{ij} \,|\, \mathbf{b}_i, \boldsymbol{\beta}, \phi) = \exp\left\{\phi^{-1}[y_{ij}\lambda_{ij} - \zeta(\lambda_{ij})] + c(y_{ij}, \phi)\right\},$$

with

$$\eta[\zeta^{'}(\lambda_{ij})] = \eta(\mu_{ij}) = \eta[E(y_{ij} \,|\, \mathbf{b}_i, \boldsymbol{\lambda})] = \mathbf{x}_{ij}^T \boldsymbol{\beta} + \mathbf{z}_{ij}^T \mathbf{b}_i,$$

where $\mathbf{x}_{ij}$ and $\mathbf{z}_{ij}$ are $p-$dimensional and $q-$dimensional vectors of known covariates, respectively. Further, $\boldsymbol{\beta}$ is a $p-$dimensional vector of unknown fixed effects parameters, ϕ is a dispersion parameter and $\eta(\cdot)$ is a known link function. The distribution for the random effects $f(\mathbf{b}_i \,|\, \mathbf{D})$ is most often specified as $N_q(\mathbf{0}, \mathbf{D})$.

In this paper we aim to illustrate the performance of the Bayesian model selection criteria on longitudinal count data. As a start, usually the Poisson distribution is taken for counts. With repeated measures, the Poisson mixed-effects model (PMM) in the context of a longitudinal study becomes

$$\begin{aligned} Y_{ij} &\sim Poi(\lambda_{ij} \,|\, \mathbf{b}_i), \ (i = 1, \ldots, n; j = 1, \ldots, m_i) \\ \log(\lambda_{ij}) &= \mathbf{x}_{ij}^T \boldsymbol{\beta} + \mathbf{z}_{ij}^T \mathbf{b}_i, \\ \mathbf{b}_i &\sim N_q(\mathbf{0}, \mathbf{D}). \end{aligned}$$

In Section 5.5, we will describe some extensions of this model to deal with overdispersion.

5.3 Bayesian model selection criteria

In this paper, we aim to illustrate the performance of the marginal and conditional versions of DIC, WAIC and PSBF on a Poisson mixed effects model and its extensions described in Section 5.5. The conditional version of these selection criteria is based on the conditional likelihood incorporating the random effects, i.e. they are based on the conditional likelihood $p(\mathbf{y} \,|\, \mathbf{\Theta}, \mathbf{b}) \equiv L(\mathbf{\Theta}, \mathbf{b} \,|\, \mathbf{y}) = \prod_i L(\mathbf{\Theta}, \mathbf{b}_i \,|\, \mathbf{y}_i)$, with $\mathbf{y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$ the total set of responses, $\mathbf{\Theta}$ the model parameters (fixed effects $\boldsymbol{\beta}$ and the variance parameters of the random effects, i.e. the elements of $\mathbf{D}$) and $\mathbf{b} = \{\mathbf{b}_1, \ldots, \mathbf{b}_n\}$ the total set of random effects. In contrast, the marginal criteria are based on the marginal likelihood, which is simply the conditional likelihood integrated over the distribution of the random effects, i.e. the marginal likelihood is given by

$$L(\mathbf{\Theta} \,|\, \mathbf{y}) = \prod_{i=1}^{n} L(\mathbf{\Theta} \,|\, \mathbf{y}_i) = \prod_{i=1}^{n} \int L(\mathbf{\Theta}, \mathbf{b}_i \,|\, \mathbf{y}_i)\, p(\mathbf{b}_i \,|\, \mathbf{\Theta})\, d\mathbf{b}_i.$$

Most often, integration over the distribution of the random effects requires numerical procedures, such as (non)-adaptive Gaussian quadrature methods. For the sake of completeness, we will now briefly describe the general definition of the three considered model selection criteria, and apply their definitions to GLMMs. Below ψ represents $\{\mathbf{\Theta}, \mathbf{b}\}$ for the conditional version of the model selection criteria, while it represents $\mathbf{\Theta}$ for the marginal version of the selection criteria. Let also $D(\psi)$ represent the deviance of the model evaluated in ψ, i.e. $D(\psi) = -2 \log p(\mathbf{y} \,|\, \psi) + 2 \log f(\mathbf{y})$, with $f(\mathbf{y})$ represents the likelihood of a saturated model. The latter term is omitted in calculations and also here.

The deviance information criterion is then defined as DIC $= D(\overline{\psi}) + 2p_{DIC}$, where $D(\overline{\psi})$ is the deviance (often) evaluated at the posterior mean. p_{DIC} is called effective number of parameters of the model and is a contrast of the posterior mean of the deviance $\overline{D(\psi)}$ with $D(\overline{\psi})$ and is equal to $p_{DIC} = \overline{D(\psi)} - D(\overline{\psi})$. Both DIC and p_{DIC} can be approximated from an MCMC run with a converged chain $\psi^1, \ldots, \psi^K$. Namely, $\overline{D(\psi)} \approx \frac{1}{K} \sum_{k=1}^{K} D(\psi^k)$ and $D(\overline{\psi}) \approx D(\frac{1}{K} \sum_{k=1}^{K} \psi^k)$. In practice, one chooses the model with the smallest DIC value. As with AIC/BIC a difference of 5 to 10 of DIC values is needed to make a justified choice between models, but now DIC and p_{DIC} are also subject to sampling variability complicating the model choice obviously.

The conditional version of DIC (cDIC) is obtained by plugging the conditional deviance into the expression of DIC, and by taking the posterior mean of $(\mathbf{\Theta}, \mathbf{b})$. The associated effective degrees of freedom is denoted then as p_{cDIC}. The marginal version of DIC (mDIC) is obtained by plugging in the marginal deviance into the expression of DIC together with the posterior mean of $\mathbf{\Theta}$ (which is the same as for the conditional likelihood). We denote the effective degrees of freedom now as p_{mDIC}. Note that the marginal deviance is the posterior mean of the log of the conditional likelihoods averaged over the distribution of the random effects, i.e.

$$\mathbf{E}_{\boldsymbol{\Theta}\,|\,y}\Big[-2\log p\,(\mathbf{y}_i\,|\,\boldsymbol{\Theta})\Big]=\sum_{i=1}^{n}\mathbf{E}_{\boldsymbol{\Theta}\,|\,\mathbf{y}}\Big[-2\log\mathbf{E}_{\mathbf{b}_i\,|\,\boldsymbol{\Theta}}\big[p(\mathbf{y}_i\,|\,\boldsymbol{\Theta},\mathbf{b}_i)\big]\Big].$$

While DIC is still the most popular model selection criterion, it has been criticised that it lacks a good theoretical motivation, but it also suffers from some practical problems. For instance, DIC is not invariant to non-linear transformations of the parameters in $\boldsymbol{\Theta}$ and p_{DIC} can become negative in which case DIC cannot be used. The latter happens especially with the conditional DIC (Spiegelhalter et al., 2014). For further details and illustrations of such problems, one can look in Lesaffre and Lawson (2012).
To accommodate certain problems with DIC, Watanabe (2010) has suggested the Widely Applicable Information Criterion abbreviated as WAIC. Indeed, the advantage of WAIC over DIC is that it is invariant to non-linear transformations of the parameters and the associated degrees of freedom cannot be negative. WAIC is an approximation to minus twice the expected log pointwise predictive density (elppd) for new data $\widetilde{y}_{ij}$. Hence, the elppd is given as

$$elppd=-2\sum_{i=1}^{n}\mathbf{E}_{\widetilde{\mathbf{y}}_i}\log\Big[\mathbf{E}_{\boldsymbol{\Theta},\mathbf{b}\,|\,\mathbf{y}}p\,(\widetilde{\mathbf{y}}_i\,|\,\boldsymbol{\Theta})\Big]. \qquad (5.3)$$

When the responses y_{ij} are independent given the random effects (e.g. when there is no serial correlation), then the above expression can be written as:

$$elppd=-2\sum_{i=1}^{n}\sum_{j=1}^{m_i}\mathbf{E}_{\widetilde{y}_{ij}}\log\Big[\mathbf{E}_{\boldsymbol{\Theta},\mathbf{b}\,|\,\mathbf{y}}p\,(\widetilde{y}_{ij}\,|\,\boldsymbol{\Theta},\mathbf{b}_i)\Big]. \qquad (5.4)$$

From expressions (5.3) and (5.4) one can see that the logarithm of the pointwise (posterior) predictive density for a future response $\widetilde{\mathbf{y}}_i$, i.e. lppd $=\mathbf{E}_{\boldsymbol{\Theta},\mathbf{b}\,|\,\mathbf{y}}p\,(\widetilde{\mathbf{y}}_i\,|\,\boldsymbol{\Theta},\mathbf{b}_i)$, is averaged over the distribution of new responses. Based on a converged chain $\Big\{\boldsymbol{\Theta}^1,\ldots,\boldsymbol{\Theta}^K,\mathbf{b}_1^1,\ldots,\mathbf{b}_1^K,\ldots,\mathbf{b}_n^1,\ldots,\mathbf{b}_n^K\Big\}$ the conditional WAIC can be computed as

$$\text{cWAIC}=-2\sum_{i=1}^{n}\log\left[\frac{1}{K}\sum_{k=1}^{K}p(\mathbf{y}_i\,|\,\boldsymbol{\Theta}^k,\mathbf{b}_i^k)\right]+2p_{\text{cWAIC}}, \qquad (5.5)$$

with $p_{\text{cWAIC}}=2\left(\sum_{i=1}^{n}\log\left[\frac{1}{K}\sum_{k=1}^{K}p(\mathbf{y}_i\,|\,\boldsymbol{\Theta}^k,\mathbf{b}_i^k)\right]-\frac{1}{K}\sum_{k=1}^{K}\log\ p(\mathbf{y}_i\,|\,\boldsymbol{\Theta}^k,\mathbf{b}_i^k)\right).$

The WAIC of the marginal model, i.e. the marginal WAIC, is then computed by

$$\mathsf{mWAIC} = -2\sum_{i=1}^{n}\log\left[\frac{1}{K}\sum_{k=1}^{K}p(\mathbf{y}_i\,|\,\mathbf{\Theta}^k)\right] + 2p_{\mathsf{mWAIC}}, \qquad (5.6)$$

with $p_{\mathsf{mWAIC}} = 2\left(\sum_{i=1}^{n}\log\left[\frac{1}{K}\sum_{k=1}^{K}p(\mathbf{y}_i\,|\,\mathbf{\Theta}^k)\right] - \frac{1}{K}\sum_{k=1}^{K}\log\;p(\mathbf{y}_i\,|\,\mathbf{\Theta}^k)\right)$. The Bayes factor is equal to the ratio of the marginal likelihood of two competing models. The pseudo-Bayes factor is a version of the Bayes factor. Since the Bayes factor is based on the prior distribution of the model parameters, its computation becomes complicated with a vague prior for the parameters. Several "solutions" were proposed to solve this problem and many boil down to applying the vague prior to a part of the data, and then use the resulting posterior as a prior for the calculation of the Bayes factor on the remaining data. The pseudo-Bayes factor deviates from this principle a bit by also involving cross-validation. Suppose we have two models $\mathcal{M}_1$ and $\mathcal{M}_2$ with model parameters ψ_1 and ψ_2, respectively and data $\{\mathbf{y}_1,\ldots,\mathbf{y}_n\}$. The Bayes factor is based on the marginal likelihood, in the sense that the likelihood is marginalised over the prior uncertainty of the model parameters. Namely, this marginal likelihood is given for model $\mathcal{M}$ and parameters ψ (leaving out the model subscript) by:

$$p(\mathbf{y}\,|\,\mathcal{M}) = \int\prod_{i=1}^{n}p(\mathbf{y}_i\,|\,\psi,\mathcal{M})\,p(\psi)\,d\psi. \qquad (5.7)$$

However, (5.7) is not analytically available in general. Therefore, Geisser and Eddy (1979) suggested replacing (5.7) by the pseudo marginal likelihood (PML)

$$\widehat{p}(\mathbf{y}\,|\,\mathcal{M}) = \prod_{i=1}^{n}p(\mathbf{y}_i\,|\,\mathbf{y}_{-i},\mathcal{M}), \qquad (5.8)$$

where $p(\mathbf{y}_i\,|\,\mathbf{y}_{-i},\mathcal{M})$ is called the ith conditional predictive ordinate (CPO_i) and is the predictive density calculated at the observed $\mathbf{y}_i$ given $\mathbf{y}_{-i}$, which is the set of all data except the ith observation. The pseudo-Bayes factor is then obtained by taking the ratio $\widehat{p}(\mathbf{y}\,|\,\mathcal{M}_1)/\widehat{p}(\mathbf{y}\,|\,\mathcal{M}_2)$ to evaluate the preference of model $\mathcal{M}_1$ over model $\mathcal{M}_2$. Low values of this ratio reflect preference of model $\mathcal{M}_2$ based on the current data. The conditional pseudo-Bayes factor (cPSBF) (given random effects) and the marginal pseudo-Bayes factor (mPSBF) (averaged over random effects) are based on (5.8) with the conditional and the marginal likelihood plugged-in, respectively. In practice, one often evaluates the logarithm of expression (5.8), leading to the log pseudo marginal likelihood for model $\mathcal{M}_\ell$ equal to $\mathsf{LPML}_\ell = \sum_{i=1}^{n}\log(\mathsf{CPO}_{i,\ell})$ where

$$\mathsf{CPO}_{i,\ell} \approx \left[\frac{1}{K}\sum_{k=1}^{K}\frac{1}{p(\mathbf{y}_i\mid\mathbf{\Theta}_\ell^k,\;\mathcal{M}_\ell)}\right]^{-1},$$

where $\mathbf{\Theta}_\ell^k$ represents the model parameters for model $\mathcal{M}_\ell$.

5.4 Sampling methods for computing the marginal model selection criteria

The expression of the model selection criteria reveals that expected values over the distribution of the random effects need to be taken. In the simpler case of a linear mixed model, the computations are easy since the marginal LMM can be determined analytically, but this is not the case for a GLMM. For this reason, we explored the use of sampling methods to compute the model selection criteria for a GLMM. Here we combined the replication method, which is sampling from the prior of the random effects, with importance sampling to compute the marginal criteria. The former replaces the integral in $p(\mathbf{y}_i \,|\, \boldsymbol{\Theta}) = \int p(\mathbf{y}_i \,|\, \boldsymbol{\Theta}, \mathbf{b}_i)\, p(\mathbf{b}_i \,|\, \boldsymbol{\Theta})\, d\mathbf{b}_i = \mathbf{E}_{\mathbf{b}_i \,|\, \boldsymbol{\Theta}}[p(\mathbf{y}_i \,|\, \boldsymbol{\Theta}, \mathbf{b}_i)]$ by sampling from the prior distribution of $\mathbf{b}_i$. Importance sampling is a variance reduction technique. In this paper, we compute the marginal version of DIC, WAIC and PSBF based on these sampling techniques. An R function has been written and is available in the supporting materials of the paper.

5.4.1 The replication method

The joint posterior $p(\boldsymbol{\Theta}, \mathbf{b} \,|\, \mathbf{y})$ can be approximated by making use of a MCMC sample $(\boldsymbol{\Theta}^k, \widetilde{\mathbf{b}}^k)$, $(k = 1, \ldots, K)$. Since $p(\mathbf{y}_i \,|\, \boldsymbol{\Theta}) = \int p(\mathbf{y}_i \,|\, \boldsymbol{\Theta}, \mathbf{b}_i)\, p(\mathbf{b}_i \,|\, \boldsymbol{\Theta})\, d\mathbf{b}_i = \mathbf{E}_{\mathbf{b}_i \,|\, \boldsymbol{\Theta}}[p(\mathbf{y}_i \,|\, \boldsymbol{\Theta}, \mathbf{b}_i)]$, the marginal criteria such as mDIC can be based on independent replicates $\widetilde{\mathbf{b}_i^{k,l}}$, $(l = 1, \ldots, L)$ from $p(\mathbf{b}_i \,|\, \boldsymbol{\Theta}^k)$ at each iteration k. To compute the plug-in deviance, we take replicates $\widetilde{\mathbf{b}}_i^m$ from $p(\mathbf{b}_i \,|\, \overline{\boldsymbol{\Theta}})$ $(m = 1, \ldots, M)$ in order to approximate $\sum_{i=1}^n \log[p(\mathbf{y}_i \,|\, \overline{\boldsymbol{\Theta}})] = \sum_{i=1}^n \log \mathbf{E}_{\mathbf{b}_i \,|\, \overline{\boldsymbol{\Theta}}}[p(\mathbf{y}_i \,|\, \overline{\boldsymbol{\Theta}}, \mathbf{b}_i)]$. Thus, the components necessary to compute the marginal criterion mDIC are

$$\begin{aligned} \overline{D(\boldsymbol{\Theta})} &\approx -2 \sum_{i=1}^n \left(\frac{1}{K} \sum_{k=1}^K \log \left[\frac{1}{L} \sum_{l=1}^L p(\mathbf{y}_i \,|\, \boldsymbol{\Theta}^k, \widetilde{\mathbf{b}}_i^{k,l}) \right] \right), \\ D(\overline{\boldsymbol{\Theta}}) &\approx -2 \sum_{i=1}^n \log \left[\frac{1}{M} \sum_{m=1}^M p(\mathbf{y}_i \,|\, \overline{\boldsymbol{\Theta}}, \widetilde{\mathbf{b}}_i^m) \right]. \end{aligned} \tag{5.9}$$

The variability of (5.9) due to the replication method depends on several factors which include: (i) the number of observations in the sample, (ii) the variance of the latent variables induced by $p(\mathbf{b} \,|\, \boldsymbol{\Theta})$, and (iii) the posterior variance of the parameters. In Quintero and Lesaffre (2018), an expression of the variance of mDIC is given. For a small value of Var(mDIC), the proposed estimator (5.9) provides a good approximation to mDIC. However, as pointed out in Quintero and Lesaffre (2018), this variance can be high for large sized clusters and when there are many clusters, which corresponds here to subjects with many repeated observations per subject or many subjects, respectively. In the same spirit, the components necessary to compute marginal criterion mWAIC are

$$
\begin{aligned}
\text{mlppd} &= \frac{1}{K}\sum_{i=1}^{n}\sum_{k=1}^{K}\log\left\{\frac{1}{L}\sum_{l=1}^{L} p\left(\mathbf{y}_i \,|\, \boldsymbol{\Theta}^k, \widetilde{\mathbf{b}}^{k,l}\right)\right\},\\
p_{\text{mWAIC}} &= 2\sum_{i=1}^{n}\left[\log\left\{\frac{1}{M}\sum_{m=1}^{M} p\left(\mathbf{y}_i \,|\, \boldsymbol{\Theta}^m, \widetilde{\mathbf{b}}^{m}\right)\right\} - \frac{1}{K}\sum_{k=1}^{K}\left\{\frac{1}{L}\sum_{l=1}^{L}\log p\left(\mathbf{y}_i \,|\, \boldsymbol{\Theta}^k, \widetilde{\mathbf{b}}^{k,l}\right)\right\}\right],
\end{aligned}
\tag{5.10}
$$

where mlppd is the log pointwise predictive density for the marginal model and p_{mWAIC} is the corresponding effective number of parameters to adjust for overfitting. The mPSBF consists of comparing the (marginal) log-pseudo likelihood (mLPML) for models M_1 and M_2, whereby mLPML is equal to $\sum_{i=1}^{n}\log(\text{mCPO}_{i,\ell})$ for model M_ℓ where

$$
\text{mCPO}_{i,\ell} \approx \left[\frac{1}{K}\sum_{k=1}^{K}\frac{1}{\frac{1}{L}\sum_{l=1}^{L} p\left(\mathbf{y}_i \,|\, \boldsymbol{\Theta}^k, \widetilde{\mathbf{b}}^{k,l}, \mathsf{M}_\ell\right)}\right]^{-1}. \tag{5.11}
$$

The replication method can be based on simple random sampling, but there is gain using instead importance sampling, see e.g. Tran et al. (2016) and Tokdar and Kass (2010) for an overview of the advantages of importance sampling over simple random sampling.

5.4.2 Adequacy of the number of replications

From expressions (5.9), (5.10) and (5.11), it is clear that the number of subjects n in the data set impacts the variability of the estimators. A larger sample size leads to greater variability of $\overline{D(\boldsymbol{\Theta})}$, $D(\overline{\boldsymbol{\Theta}})$, lppd, p_{WAIC} and LPML since these estimators are the sums of the log-likelihoods pertaining to the observation units. In order to approximate well the true marginal model selection criteria, it is important to select L and M appropriately (not too small nor too high).

For DIC, Quintero and Lesaffre (2018) suggested to take $L = 2M/\sqrt{K_{Eff}}$, where
$K_{Eff} = K/(1 + 2\sum_{t=1}\rho_t)$ and $\rho_t = \text{Corr}(\sum_i \log\widehat{p}(\mathbf{y}_i \,|\, \boldsymbol{\Theta}^k), \sum_i \log\widehat{p}(\mathbf{y}_i \,|\, \boldsymbol{\Theta}^k + \mathbf{t}))$.

Among others, these authors suggested to determine L and M such that the standard error for mDIC is smaller 0.5, such that the variability (measured by 95% CI) of mDIC can be expected to be smaller than 1.

Here we checked the adequacy of the choice of L and M for these selection criteria in a numerical exercise, see below. From this exercise we tentatively conclude that when L and M are appropriate for DIC, they are likely to be appropriate for WAIC and LPML.

To illustrate the required number of replications, we performed a small simulation exercise. This is part of the simulation exercise described in more detail in Section 5.7.2. To this end, we have taken a Poisson mixed model. Namely, let Y_{ij} be a count for the ith subject ($i = 1, \ldots, n = 300$) at the jth time point ($j = 1, \ldots, 5$) and b_{0i} the ith random intercept with $b_{0i} \sim N(0, \sigma^2_{b0})$. We allowed for time independent covariates for the ith subject: age at baseline (age$_i$), baseline count (base$_i$), treatment (treat$_i$), interaction baseline count and treatment (basetreat$_i$) and the obvious time dependent covariate time (time$_ij$). That is, we assumed

$$Y_{ij} \sim \text{Poisson}(\lambda_{ij} \mid b_{0i}), \tag{5.12}$$

where

$$\lambda_{ij} = \beta_1 + \beta_2 trt_i + \beta_3 \log(base_i) + \beta_4 visit_{ij} + \beta_5 \log(age_i) + \beta_6 trt_i \times \log(base_i) + b_{0i}.$$

The estimates of the model parameters are given in Section 6.5. It is suggested by Mason et al. (2012) to monitor the stability of the components of (5.9), (5.10) and (5.11) when increasing the number of replications. Figure 5.1 displays the marginal criteria components for the above Poisson model for increasing number of replications M. From this figure we can see that mDIC and mWAIC, and their components stabilise for $M = 8000$. Recall, that for mDIC we have also another basis to decide about its desired value, namely that the standard error of the estimated mDIC should be smaller than 0.5. This is achieved for $M = 8000$ as then the standard error is 0.2. For mWAIC and mLMPL we judged the adequacy of M purely graphically. Note that for mLMPL stability is already achieved with M around 7000.

Further, we evaluated the dependence of the required M on the number of subjects and the number of observations/subject. For this we have considered data sets with 10, 50 and 200 subjects combined with 4, 6 and 10 observations per subject. Each data set was generated according to the above Poisson mixed effects model. From basic principles, Quintero and Lesaffre (2018) concluded that the required M likely increases with increasing number of subjects and/or observations/subject. Note that, as before, $L = 2M/\sqrt{K_{Eff}}$. In Table 5.1, we show the model selection criteria when varying M from 5000 to 10 000. In contrast to the above conclusion, the suspected dependence does not show.

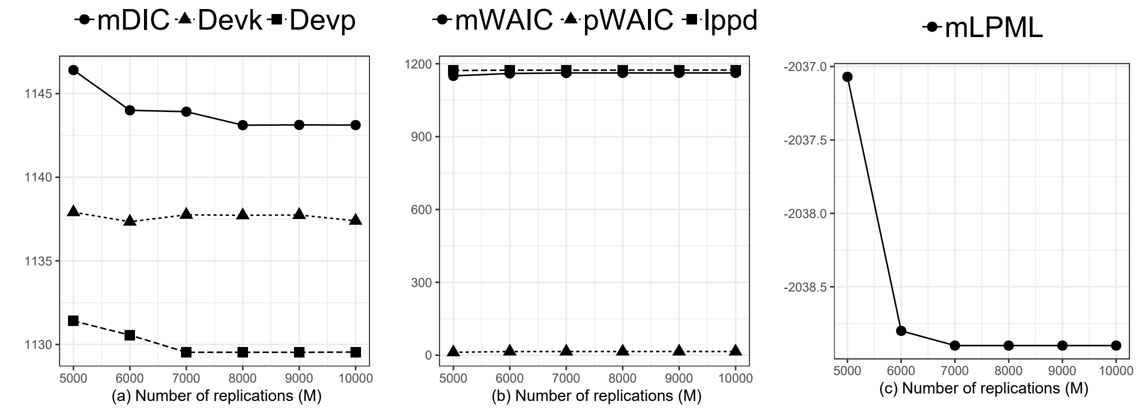

Figure 5.1*: Poisson model (5.12): Dependence of marginal model selection criteria on the number of replications M for: (a) mDIC, Devk=$\overline{D(\mathbf{\Theta})}$ and Devp=$D(\bar{\mathbf{\Theta}})$ as given in (5.9); (b) mWAIC, p$_{mWAIC}$ and mlppd as given in (5.10): (c) mLMPL as given in (5.11)*

Table 5.1: *Marginal selection criteria as a function of M for different numbers of subjects and observations per subject.*

		Number of observations/subject								
		4			6			10		
# of subjects	M	mDIC	mWAIC	mLPML	mDIC	mWAIC	mLPML	mDIC	mWAIC	mLPML
10	5000	114.562	116.762	-47.588	114.762	116.022	-47.359	114.902	116.212	-47.567
	6000	114.662	116.892	-47.365	114.342	116.002	-47.439	114.562	116.046	-47.234
	7000	114.342	116.058	-47.288	113.392	115.122	-47.167	113.212	115.042	-47.162
	8000	113.462	115.478	-47.285	113.220	115.038	-47.169	113.108	115.002	-47.162
	9000	113.262	115.426	-47.285	113.210	115.026	-47.168	113.062	115.002	-47.072
	10000	113.062	115.426	-47.285	113.208	115.026	-47.167	113.062	115.002	-47.084
50	5000	1146.404	1150.210	-537.070	1156.404	1158.210	-537.070	1146.404	1159.210	-547.072
	6000	1144.000	1150.110	-536.800	1154.100	1156.210	-536.800	1144.000	1157.210	-549.802
	7000	1143.917	1149.210	-536.900	1153.017	1152.210	-526.723	1143.917	1152.210	-548.902
	8000	1143.111	1149.110	-536.900	1153.011	1152.315	-526.842	1143.111	1152.310	-540.102
	9000	1143.126	1149.120	-536.900	1153.026	1152.320	-526.812	1143.126	1152.320	-540.102
	10000	1143.117	1149.120	-536.900	1153.017	1152.320	-526.800	1143.117	1152.320	-540.202
200	5000	1559.340	1660.010	-804.920	1564.012	1627.900	-804.200	1562.890	1616.794	-801.122
	6000	1560.210	1660.420	-804.890	1564.002	1625.010	-804.120	1562.320	1616.774	-801.102
	7000	1562.120	1631.840	-804.100	1562.122	1621.840	-804.070	1562.120	1616.744	-801.072
	8000	1562.130	1621.880	-804.070	1560.592	1620.230	-804.070	1562.050	1616.634	-801.072
	9000	1562.130	1621.890	-804.070	1560.532	1620.210	-804.070	1562.043	1616.604	-801.072
	10000	1562.130	1621.900	-804.070	1560.532	1620.200	-804.070	1562.041	1616.604	-801.072

5.4.3 Importance sampling

Importance sampling consists of replacing an original integral over a distribution by sampling another easier-to sample distribution, called the proposal density, and then replace the integral by sampling. Given that $p(\mathbf{y}_i \,|\, \overline{\boldsymbol{\Theta}}) = \int g_i(\mathbf{b}_i)\, d\mathbf{b}_i$ with $g_i(\mathbf{b}_i) = p(\mathbf{y}_i \,|\, \overline{\boldsymbol{\Theta}}, \mathbf{b}_i)\, p(\mathbf{b}_i \,|\, \overline{\boldsymbol{\Theta}})$ is replaced by $p(\mathbf{y}_i \,|\, \overline{\boldsymbol{\Theta}}) = \int [g_i(\mathbf{b}_i)/q_i(\mathbf{b}_i)]\, q_i(\mathbf{b}_i) d\mathbf{b}_i$, with $q_i(\mathbf{b})$ an appropriate proposal density. Then $p(\mathbf{y}_i \,|\, \overline{\boldsymbol{\Theta}}, \mathbf{b}_i)\, p(\mathbf{b}_i \,|\, \overline{\boldsymbol{\Theta}})$ is proportional to $p(\mathbf{b}_i \,|\, \mathbf{y}_i, \overline{\boldsymbol{\Theta}})$, the mean and the variance of $g_i(\mathbf{b}_i)$ can be estimated from an additional MCMC run fixing the parameters to $\boldsymbol{\Theta} = \overline{\boldsymbol{\Theta}}$.
We used this approach to evaluate $p(\mathbf{y}_i \,|\, \overline{\boldsymbol{\Theta}})$ for each observation unit components needed for the marginal criteria.

As pointed out by Quintero and Lesaffre (2018), this posterior distribution is approximately normal for large sized observation units under regularity conditions, so it is adequate to select a normal density for $q_i(\mathbf{b}_i)$ with the above mean and variance. This approach is based on the independent Metropolis-Hastings algorithm with proposal density $q_i(\mathbf{b}_i)$. For small sized observation units, the function $g_i(\mathbf{b}_i)$ resembles the latent prior density, so it is appropriate to select $q_i(\mathbf{b}_i) = p(\mathbf{b}_i \,|\, \overline{\boldsymbol{\Theta}})$ (Quintero and Lesaffre, 2018). Then, after sampling $\widetilde{\mathbf{b}}_i^m$ from $q_i(\mathbf{b})$ for $m = 1, \ldots, M$, different components for the marginal criteria are computed based on the plug-in deviance given by

$$\widehat{p}(\mathbf{y}_i \,|\, \overline{\boldsymbol{\Theta}}) = 1/M \sum_{m=1}^{M} [p(\mathbf{y}_i \,|\, \overline{\boldsymbol{\Theta}}, \widetilde{\mathbf{b}}_i^m) p(\widetilde{\mathbf{b}}_i^m \,|\, \overline{\boldsymbol{\Theta}})/q_i(\widetilde{\mathbf{b}}_i^m)],$$

and the mean deviance where $\boldsymbol{\Theta}^m$ is substituted with deviance for each iteration. Hence, the mean deviance is given by

$$\widehat{p}(\mathbf{y}_i \,|\, \boldsymbol{\Theta}^m) = 1/M \sum_{m=1}^{M} \left[1/L \sum_{l=1}^{L} [p(\mathbf{y}_i \,|\, \boldsymbol{\Theta}^m, \widetilde{\mathbf{b}}_i^{m,l}) p(\widetilde{\mathbf{b}}_i^{m,l} \,|\, \boldsymbol{\Theta}^m)/q_i(\widetilde{\mathbf{b}}_i^{m,l})] \right].$$

Thus, to compute the marginal criteria components we use importance sampling based on MCMC for large-sized observation units, but for small-sized observation units, independent sampling method can be used. This strategy for importance sampling simplifies and generalises the replication method in Chan and Grant (2016a).

5.5 Extensions of the Poisson-mixed model

We will illustrate below the performance of the marginal and conditional model selection criteria on selecting the appropriate fixed effects. However, with count data there is always the possibility of overdispersion and occasionally of underdispersion. Overdispersion occurs when the data display more variability than is predicted by the assumed model. For counts, we usually start with a Poisson model that assumes that the mean and variance of the counts are equal. When the variance is larger (smaller) than the mean, we speak of overdispersion (underdispersion) compared to the Poisson model. Most often counts encountered in medical data do not satisfy the Poisson assumption. As such, ignoring over/underdispersion may influence the model estimates and therefore the (statistical) conclusions. Indeed, it is well-known that when overdispersion in the data is ignored, many of the regressors will show to be wrongly 'significant'. On the other hand, Fitzmaurice (1997) evaluated the performance of the classical frequentist model selection criteria AIC and BIC, but also of his proposed modified likelihood ratio statistics. The author observed that the considered selection criteria often prefer overdispersion models even when there is no overdispersion in the data set. Obviously, this can lead to a wrong interpretation of the model parameters. We refer to Lambert (1992) for more background on the issues of overdispersion and its impact on the conclusions of a statistical analysis.

Therefore, we wished to evaluate the performance of the above three model selection criteria when overdispersion is present or absent in the data set. We ignore here the case of underdispersion, since this occurs less frequently in practice. But in order to detect such deviation from the Poisson model, we need to have statistical models for repeated count data that allow for overdispersion. Without clustering, some models have been suggested to model overdispersion. A popular choice is the negative binomial distribution, which arises as a continuous mixture of Poisson distributions with means that have a gamma distribution. If overdispersion is due to an excess of zeros, one could model the data with a zero-inflated Poisson distribution or a zero-inflated negative binomial distribution, both are mixtures of the basic (Poisson/negative binomial) distribution with a degenerate distribution at zero. Also for longitudinal count data, models have been suggested that deal with overdispersion, see e.g. Booth et al. (2003); Aregay et al. (2013, 2015); Molenberghs et al. (2007). Here, we focus on the extensions suggested by Molenberghs et al. (2007, 2010), which we briefly describe below.

5.5.1 The Poisson-type models for count data with overdispersion

A natural extension of the random effects Poisson model is to make use of the generalisations suggested for a Poisson model. That is, to allow for overdispersion by assuming a Poisson-gamma model or a zero-inflated Poisson/negative binomial model given the random effects. The first proposal was suggested in Molenberghs et al. (2007, 2010). More specifically, these authors suggest

$$\begin{aligned} Y_{ij}|\,\mathbf{b}_i &\sim Poi(\lambda_{ij}\,|\,\mathbf{b}_i),\ (i=1,\ldots,n; j=1,\ldots,m_i) \\ \lambda_{ij} &= \theta_{ij}\exp\left(\mathbf{x}_{ij}^T\boldsymbol{\beta}+\mathbf{z}_{ij}^T\mathbf{b}_i\right), \\ \mathbf{b}_i &\sim N(\mathbf{0},\mathbf{D}), \\ E(\boldsymbol{\theta}_i) &= E[(\theta_{i1},\ldots,\theta_{im_i})^T], \\ Var(\boldsymbol{\theta}_i) &= \Sigma_i, \end{aligned} \tag{5.13}$$

whereby θ_{ij} measures the overdispersion in the outcome for the ith subject at the jth occasion. When θ_{ij} has a Gamma(α_1,α_2) distribution, we call it a Poisson-gamma mixed effects model (PGMM). Alternatively, one could assume that the θ_{ij} has a lognormal distribution. In that case we speak of a Poisson-lognormal mixed effects model (PLMM). Molenberghs et al. (2007) provide the expressions for the mean vector, the variance-covariance matrix and the joint marginal probability. Here we focus on the PGMM model.

5.5.2 Zero-inflated GLMM

In the same spirit one could suggest a zero-inflated Poisson mixed-effects model (ZIPMM) or a zero-inflated negative binomial mixed-effects model (ZINBM). That is, given the random effects one could assume a zero-inflated Poisson/negative binomial. More specifically, the ZIPMM model for Y_{ij} is given by:

$$\begin{aligned} & Y_{ij}|\,\mathbf{b}_i \sim ZIP(p_{0,ij},\lambda_{ij}\,|\,\mathbf{b}_i),\ (i=1,\ldots,n; j=1,\ldots,m_i) \\ & \quad\text{with} \\ & Y_{ij}|\,\mathbf{b}_i \sim \begin{cases} 0, & \text{with probability} \quad p_{0,ij} \\ Poi(\lambda_{ij}), & \text{with probability} \quad (1-p_{0,ij}), \end{cases} \end{aligned} \tag{5.14}$$

where $\lambda_{ij}=\exp(\mathbf{x}_{ij}^T\boldsymbol{\beta}+\mathbf{z}_{ij}^T\mathbf{b}_i)$. A ZIPMM will reflect the data accurately when overdispersion is caused by an excess of zeros (Adrion and Mansmann, 2012).

The use of the ZIPMM is necessary when the nature of the source of zeros is not certain. Conversely, if overdispersion is attributed to factors beyond the inflation of zeros, a ZINBM is more appropriate (Yau et al., 2003). It is important to note that the rate of zero-inflation and the nature of the source of zeros may change over time, but such considerations will be ignored here.

5.5.3 Sampling methods for extended GLMM

With an extra random effect θ_{ij} in the model, the question is how to apply the sampling techniques. One approach is to apply the sampling techniques of Section 5.4 on both the θ_{ij}'s and $\mathbf{b}_i$'s jointly. Alternatively one could integrate first θ_{ij}'s from the likelihood, and then apply the sampling techniques on $\mathbf{b}_i$. Given $\mathbf{b}_i$, the Poisson-gamma distribution averaged over θ_{ij} yields a conditional (on $\mathbf{b}_i$) negative binomial distribution. In Molenberghs et al. (2007) it is shown that

$$Y_{ij} \mid \mathbf{b}_i \sim \mathsf{NB}(\alpha_1, \gamma_{ij}),$$

with $\gamma_{ij} = 1/(1 + \lambda_{ij}\alpha_2)$, where α_1 and α_2 are the parameters of the gamma distribution and $\lambda_{ij} = \exp(\mathbf{x}_{ij}^T \boldsymbol{\beta} + \mathbf{z}_{ij}^T \mathbf{b}_i)$. Using this marginalised model (over θ_{ij}'s) allows to use the above considered sampling techniques to compute the marginal model selection criteria.

5.6 Application section: Analysis of the epilepsy data set

We consider the analysis of data obtained from 89 epileptic patients that were randomised to a novel anti-epileptic drug (AED) in combination with one or two other AEDs (44 patients) or to placebo (45 patients) (Faught et al., 1996). A 12 weeks baseline period served as a stabilisation period. They were then measured weekly for 16 weeks, after which they were entered into a long-term open extension study. Some patients were lost to follow up after 27 weeks. The primary variable (an outcome of interest) is the number of times a patient experienced seizure during the last week. Booth et al. (2003) used this data set as an illustrating example when modelling longitudinal counts data with overdispersion. Others have also used this data set to illustrate their proposed statistical models, see e.g. Aregay et al. (2013); Rakhmawati et al. (2016); Faught et al. (1996); Molenberghs et al. (2007). Figure 5.2 shows the individual curves for both the treatment groups and this figure reveals substantial variability in counts between subjects. The graph also reveals the presence of extreme values. The presence of overdispersion in counts is seen in Table 5.2 where the sample mean and variance of the counts at each week for the treatment and placebo groups are shown. Overdispersion is also seen in Figure 5.3 where the means and variances at each week are shown for each of the treatment groups.

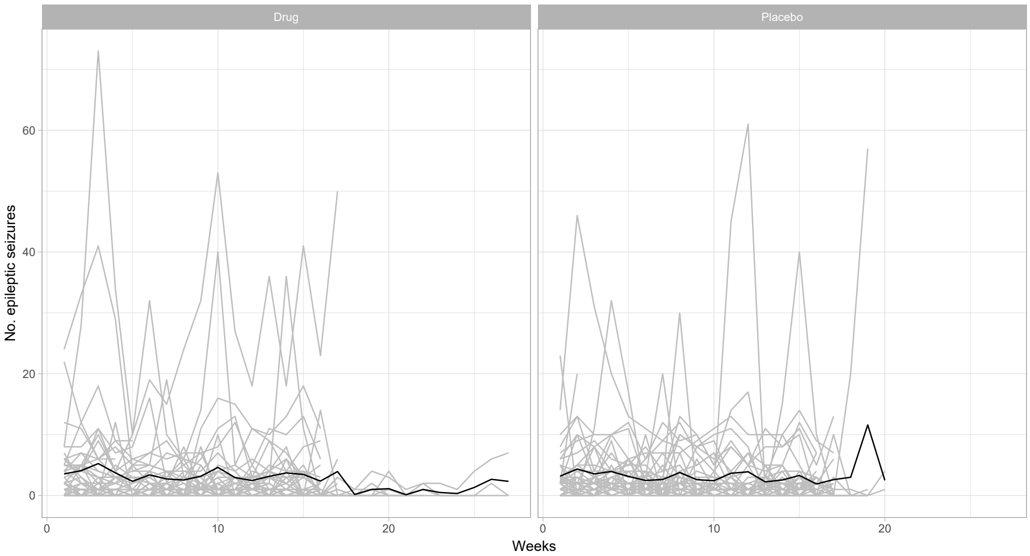

Figure 5.2: *Epilepsy data: Individual profiles for both treatment groups. Left panel: AED, Right panel: placebo.*

Breslow and Clayton (1993) analysed the epilepsy data by considering the following covariates: logarithm of baseline seizure (base) count, treatment (trt), logarithm of age, visit, and the treatment by log(base) interaction. We fitted the model

$$\begin{aligned}
Y_{ij}|\, b_{0i} &\sim \text{Poisson}(\lambda_{ij}\,|\, b_{0i}),\\
\eta_{ij} &= \log(\lambda_{ij}) = \beta_1 + \beta_2 trt_i + \beta_3 \log(base_i) + \beta_4 visit_{ij} + \beta_5 \log(age_i)\\
&\quad + \beta_6 trt_i \times \log(base_i) + b_{0i},
\end{aligned} \tag{5.15}$$

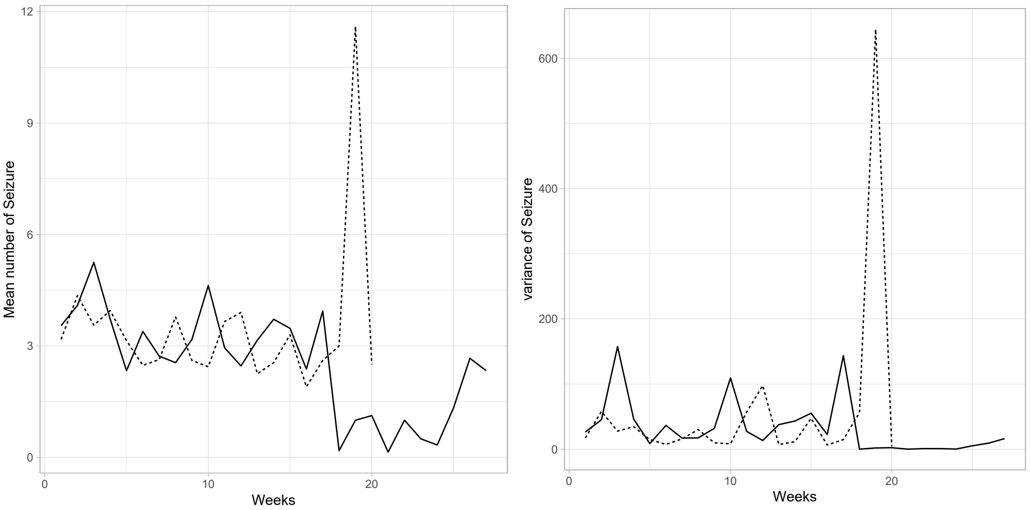

Figure 5.3: *Epilepsy data: Evolution over time of mean count (left panel) and variance of counts (right panel) for both treatment groups (AED: solid line, placebo: dashed line).*

with $b_{0i} \sim N(0, \sigma^2_{b_0})$. We fitted the Poisson mixed effects model and extensions discussed in Section 5.5 to the data using *rjags*. Here, 60,000 iterations were sampled after discarding the initial 20,000 iterations as a burn-in. The thinning factor was set as 10. For all models considered, convergence was assessed using trace plots and the Brooks, Gelman and Rubin's (BGR) diagnostic (Gelman et al., 1992). The following vague priors were chosen, for the fixed-effect parameters: $\beta_k \sim N(0, 10^2)\,(k = 0, \ldots, 6)$. For models with a random intercept and slope, we considered a separation prior for their covariance matrix i.e we have taken a half standard-Cauchy prior (Gelman, 2006) for the standard deviation of the random effects and Unif$(-1, 1)$ for the correlation parameter. For the overdispersion parameter, we assumed $\theta_{ij} \sim$ Gamma$(\alpha, 1/\alpha)\,(i = 1, \ldots, n; j = 1, \ldots, m_i)$ where $\alpha \sim$ Unif$(0, 100)$. Finally, the zero-inflated probability is assigned $p_{0,ij} \sim$ Beta$(0.5, 0.5)\,(i = 1, \ldots, n; j = 1, \ldots, m_i)$.
The results in Table 5.3 show that all marginal criteria prefer the zero-inflated

Table 5.2: *Epilepsy study: Sample mean (sample variance) at selected time-points for each of the two treatment groups.*

Week	Mean (variance)	
	Placebo	Treatment
2	4.35 (58.00)	4.09 (45.24)
4	3.95 (34.53)	3.72 (46.24)
8	3.78 (30.22)	2.55 (17.43)
10	2.44 (8.30)	4.63 (109.37)
12	3.90 (97.84)	2.95 (27.49)
14	2.55 (11.64)	3.71 (43.31)
16	1.90 (6.55)	2.39 (22.63)
18	3.00 (56.33)	0.18 (0.16)
20	2.50 (4.50)	1.13 (2.41)
27	-	2.33 (16.33)

PGMM (ZIPGMM), which is in agreement with what was obtained by Warton (2005). For the conditional criteria the best two models (PGMM and ZIPGMM) are the same as for the marginal criteria, but they rank models PMM and ZIPMM in the opposite way compared to the marginal criteria. So, it seems that there is not much difference between the solution offered by the conditional and the marginal criteria in this case. The simulations in next section check whether this is a general finding.

Table 5.3: *Epilepsy study: The value of both versions of the Bayesian selection criteria for each of the considered models for the epilepsy data sets.*

Criteria	PMM	PGMM	ZIPMM	ZIPGMM
cDIC	6045.68	4840.36	5331.78	4764.96
cWAIC	5966.37	4291.05	5215.23	4264.97
cLPML	-2132.48	-2607.61	-2145.52	-2983.19
mDIC	6203.09	6047.12	6383.50	6013.77
mWAIC	6213.70	6025.52	6472.96	6006.03
mLPML	-3016.79	-3049.28	-3085.34	-3087.93

5.7 Simulation Studies

Three simulation studies illustrate the performance of the conditional and marginal selection criteria in identifying the true data-generating model. These simulations are based on the Poisson model (5.15) discussed above under varying conditions and settings. The simulation studies mimic the data set described in Section 5.6. Using the R procedure *glmer* procedure, we obtained the maximum likelihood estimates : $\widehat{\beta}_1 = -3.96715$, $\widehat{\beta}_2 = -2.12053$, $\widehat{\beta}_3 = 0.94952$, $\widehat{\beta}_4 = -0.05872$, $\widehat{\beta}_5 = 0.89705$, $\widehat{\beta}_6 = 0.56223$, and $\widehat{\sigma}^2_{b_0} = 2.36045$. Unless specified, these values will be used as true parameters in the simulation studies described below. For the purpose of these simulations studies, we varied the number of subjects as well as number of observations per subject. Two scenarios were considered: (1) Random-effects: we assumed that the random effects structure is known but that the considered models differ from the true model in the fixed effects part; (2) Fixed-effects: we assumed that the fixed effects part is known but the random effects part is unknown.

We have taken the same settings, i.e. distribution of the random effects and the priors of the model parameters as in the previous section. For each of the two scenarios, both the conditional and marginal version all three model selection criteria were computed and recorded. Each time the model with the smallest selection criterion was selected. The simulation studies aim to confirm the superiority of the marginal criteria over conditional criteria for repeated count data as shown for linear mixed effects models (Ariyo et al., 2019b,a). Here, we also check the performance of these criteria when overdispersion is of concern. More specifically, we are interested in exploring the performance of the conditional and marginal versions of the three model selection criteria:

- to select the correct data-generating model when: (i) the random effects structure is known and correctly specified, but the fixed effects part is unknown, (ii) the fixed effects structure is known and correctly specified, but the random effects structure is unknown;

- in the absence and presence of overdispersion;

- when the number of covariates is more than the number of subjects.

In addition, we aim to:

- evaluate the performance of the two sampling methods: replication method & importance sampling in calculating the marginal criteria;

- measure the impact of the number of subjects and the number of observations per subject on the performance of the conditional and marginal criteria and the two sampling methods.

5.7.1 Simulation study 1

Here, we generated 300 data sets using the settings described in Section 5.7. We illustrate the performances of both versions of DIC, PSBF and WAIC by fitting three alternative models: (i) the true model ($\mathcal{M}_1$) give by equation (5.15), (ii) an under-specified model ($\mathcal{M}_0$) and (iii) over-specified model ($\mathcal{M}_2$). For the random-effect scenario, $\mathcal{M}_0$ is given by (5.15) without $trt_i \times \log(base_i)$ interactions and $\mathcal{M}_2$ is given by (5.15) with additional covariate $trt_i \times \log(age_i)$. Likewise, for fixed-effect scenario, $\mathcal{M}_0$ is given by (5.15) without the random intercept while $\mathcal{M}_2$ is given by (5.15) with the random intercept and slope. For these two scenarios, $\mathcal{M}_1$ is the true model.

Here, we illustrate the performance of the conditional and the marginal selection criteria in identifying the true model. Additionally, the effects of the number of subjects in the data were also evaluated in this simulation under these two scenarios. The number of observations per subject may influence the performance of replications method considered. Hence, we evaluate the performance of Bayesian model selection criteria for a moderately large number of subjects $N = 50$ and a varying number of observations per subject.

Table 5.4 presents the number of times the conditional and marginal criteria select the data-generating model for different number of observation per subject when $N = 50$. As seen in the table, the number of cluster sizes significantly influences the performance of both criteria, however, the marginal criteria often select model $\mathcal{M}_1$ while the conditional criteria select the wrong model, especially with random effects scenario. Conversely, for fixed effects scenario, the impact of the number of observations per subject on the performance of the marginal criteria is inconspicuous. Overall, the marginal criteria outperform the conditional criteria. We also evaluate the performance of the Bayesian selection criteria under the scenarios discussed above with the assumption that the overdispersion in the data set is ignored. Here, we introduced an extra parameter θ_{ij} to simulate data with overdispersion. For the extra parameter, $\theta_{ij} \sim$ Gamma$(\alpha, 1/\alpha)$ was assumed. High, moderate and low overdispersion level was induced by setting α to be $0.25; 1; 5$ respectively. We evaluate the model selection procedures when overdispersion is ignored in the data set. Here we used three wrong models: (ignoring overdispersion), models A, B and C which are the same with $\mathcal{M}_0$, $\mathcal{M}_1$ and $\mathcal{M}_2$ described above. Where model B is the closest to the data-generating

Table 5.4*: Simulation 1: The number of times the selection criteria selects the data-generating model when varying the number of observations per subject for* $n = 50$.

		Number of observations/subject								
		2			4			8		
Scenario	Criteria	$\mathcal{M}_0$	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_0$	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_0$	$\mathcal{M}_1$	$\mathcal{M}_2$
Fixed-Effects	cDIC	43	204	53	45	210	45	54	209	37
	cWAIC	41	203	56	78	168	54	81	170	49
	cLPML	50	202	48	57	204	39	39	217	44
	mDIC	49	248	3	42	253	5	39	252	9
	mWAIC	52	243	5	47	240	3	40	259	1
	mLPML	53	240	7	46	248	6	39	259	2
Random Effects										
	cDIC	41	52	207	38	47	215	27	49	224
	cWAIC	40	54	206	38	16	146	1	12	187
	cLPML	21	66	215	21	27	252	21	31	248
	mDIC	45	198	51	32	210	58	25	215	60
	mWAIC	46	196	58	34	208	53	26	217	57
	mLPML	45	198	57	37	211	52	36	214	50

model without overdispersion, the number of times each model gave the least value for model selection criteria is presented in Table 5.5.

As expected, both criteria performed poorly when overdispersion was ignored. As the overdispersion increases in the data set, the conditional criteria select the model with extra fixed and random effects parameters while the marginal criteria select the model without extra parameter (model B), the closest model to the data-generating model that ignores overdispersion. These results are similar to the conclusion in Fitzmaurice (1997) that when overdispersion is ignored, model selection tends to select a model with too many parameters and can thus lead to the over the interpretation of the parameters. In the Bayesian context, Millar (2009) advocated the use of the marginalized version of DIC and Bayes' factors as the use of the conditional DIC was misleading in the hierarchical modelling for overdispersed count data.

Table 5.5: *Simulation 1: The number of times three model specifications have the least value when overdispersion in the data set is ignored. For Low (L), Medium (M) and High (H) overdispersion.*

		L			M			H		
Scenario	Criteria	A	B	C	A	B	C	A	B	C
Fixed-Effects	cDIC	33	82	185	31	72	197	29	70	205
	cWAIC	37	83	180	38	70	192	31	69	200
	cLPML	36	83	181	41	69	190	32	67	201
	mDIC	46	229	25	79	190	31	89	186	25
	mWAIC	49	221	30	73	199	28	87	180	33
	mLPML	40	230	30	77	192	31	84	183	33
Random Effects										
	cDIC	111	37	152	70	25	195	55	35	210
	cWAIC	109	39	152	67	37	196	67	35	198
	cLPML	114	46	140	58	42	200	67	36	197
	mDIC	54	167	79	91	141	68	114	133	53
	mWAIC	90	149	61	97	142	61	103	136	61
	mLPML	41	137	122	73	137	90	110	130	60

5.7.2 Simulation study 2

From each of the PMM, PGMM and ZIPMM described above, we generate 300 data sets. Data were simulated based on equation (5.15) together with parameter estimates and the extra parameters for PGMM are based on $y_{ij} \sim \text{Poi}(\lambda_{ij}\theta_{ij})$, with y_{ij} and λ_{ij} as defined above and $\theta_{ij} \sim \text{Gamma}(\alpha, 1/\alpha)$, where α takes the values $0.25, 1,$ and 5. To ensure balanced data sets, we simulated datasets with an equal number of observations per subject. Using an appropriate number of replications as discussed in Section 5.4.2, we compute both the marginal and conditional criteria for all the three models via a self-written R code.

Different data settings were considered where data are generated from a particular model (true fit) and the model is evaluated with other alternative models. For instance, we generated data from the PMM (true fit) and considered PGMM, and ZIPMM as alternative models. Likewise, we generate data from the PGMM (true fit) and estimated with PMM and ZIPMM as alternative models and so on. The number of times when selection criteria have a lower value for an alternative model as against the true model were recorded and the percentage of misselection was calculated.

Figure 5.4 shows the histograms of the differences in selection criteria between the true model (PMM) and the alternative model (PGMM). The figure shows that the conditional criteria have a wider range of values as compared with the marginal criteria. Additionally, the marginal criteria have lower values for the PGMM model than the data-generating Poisson model in only about 2% of the times. This reflects the small penalty for the inclusion of an extra parameter in the PGMM model. Conversely, the fit of PGMM to PMM data has smaller values of the conditional criteria in 34.2%, 28.0% and 26.6% times respectively. This reflects the tendency of the conditional criteria to select a model with an extra parameter.

Similarly, Figure 5.5 shows that the conditional criteria prefer the ZIPMM model (model with extra parameter) often as against the Poisson data-generating model. In fact, the percentages of times the wrong model was selected increase from 34.2% to 52.0% for cDIC. Notwithstanding the narrow differences as shown in Figure 5.5. The marginal criteria, on the other hand, show superior performance, preferring the ZIPMM model to the data-generating Poisson model less often.

When the PMM and ZIPMM models are fitted to the PGMM data, the results give much larger values for the marginal criteria, that the marginal criteria perform better by preferring data-generating PGMM data to their conditional counterparts (see Figures 5.4 and 5.5). These results show the superior performance of marginal criteria in identifying the true data-generating model in count data sets. This is similar to the results earlier obtained for LMM (Ariyo et al., 2019b,a), and for GLMM (Millar, 2009; Quintero and Lesaffre, 2018).

5.7.3 Simulation study 3

Here, we evaluated the performance of the two sampling techniques: replication and importance sampling. Following the simulation study described in Section 5.7.1, we generated 300 data sets from Equation (5.15) under different number of subjects and observations/subject. The performance (in %) of the marginal criteria for both sampling methods in selecting the correct data-generating model is recorded in Table 5.6. The advantage of importance sampling is shown when the number of subjects and/or observations/subject is large as the replication method becomes impracticable for a large number of subjects and/or observations/subject due to convergence problems. This is reflected because the variance due to replication is larger than 0.5. This affirms the results in Quintero and Lesaffre (2018).

Table 5.6*: Simulation 3: The performance (in %) of the marginal criteria in selecting the data-generating model ($\mathcal{M}_1$) when varying the number of subjects and the number of observations/subject using two sampling methods: the replication method (Rep) and importance sampling (IS).*

		5		10		30		60	
Criteria	# of subjects	Rep	IS	Rep	IS	Rep	IS	Rep	IS
mDIC	50	78.0	76.7	78.7	79.0	36.7	76.7	28.3	76.0
	100	80.7	82.0	80.0	80.0	31.7	78.7	24.7	74.7
	200	49.0	49.0	47.7	47.0	20.3	20.3	-	24.7
	500	14.3	45.0	1.0	45.3	-	45.7	-	46.3
mWAIC	50	77.7	77.7	78.0	80.7	35.3	75.3	26.7	71.0
	100	81.3	81.7	81.7	80.3	31.0	78.3	24.0	71.3
	200	50.7	50.3	46.3	48.3	19.7	20.7	-	29.7
	500	14.7	44.7	2.7	46.7	-	45.3	-	50.3
mLMPL	50	77.3	77.0	79.0	81.7	34.0	77.0	21.3	75.0
	100	80.0	81.7	82.3	79.0	36.3	78.7	20.7	70.7
	200	42.7	51.3	47.0	48.7	19.0	23.7	-	23.7
	500	15.0	45.7	2.3	46.3	-	45.0	-	42.3

5.7.4 Simulation study 4

When only a few events happen (say when dealing with a rare disease) parameter estimates may be biased or unreliable (Chen and Wehrly, 2016). We evaluated the impact of the setting where the number of parameters p is larger than the number of subjects N on the performance of the marginal and conditional criteria. To this end, we generated 300 data sets from the Poisson model (5.15) with three additional covariates: $\log(age_i^2)$, $trt_i \times \log(age_i)$, $trt_i \times \log(age_i^2)$. We evaluated the performance when (1) $N \sim= p$ and (2) $N < p$. Table 5.7 shows that few numbers of subjects affects the performance of both versions of the criteria but with less impact on the marginal criteria. And again, the marginal criteria outperformed the conditional criteria in this setting.

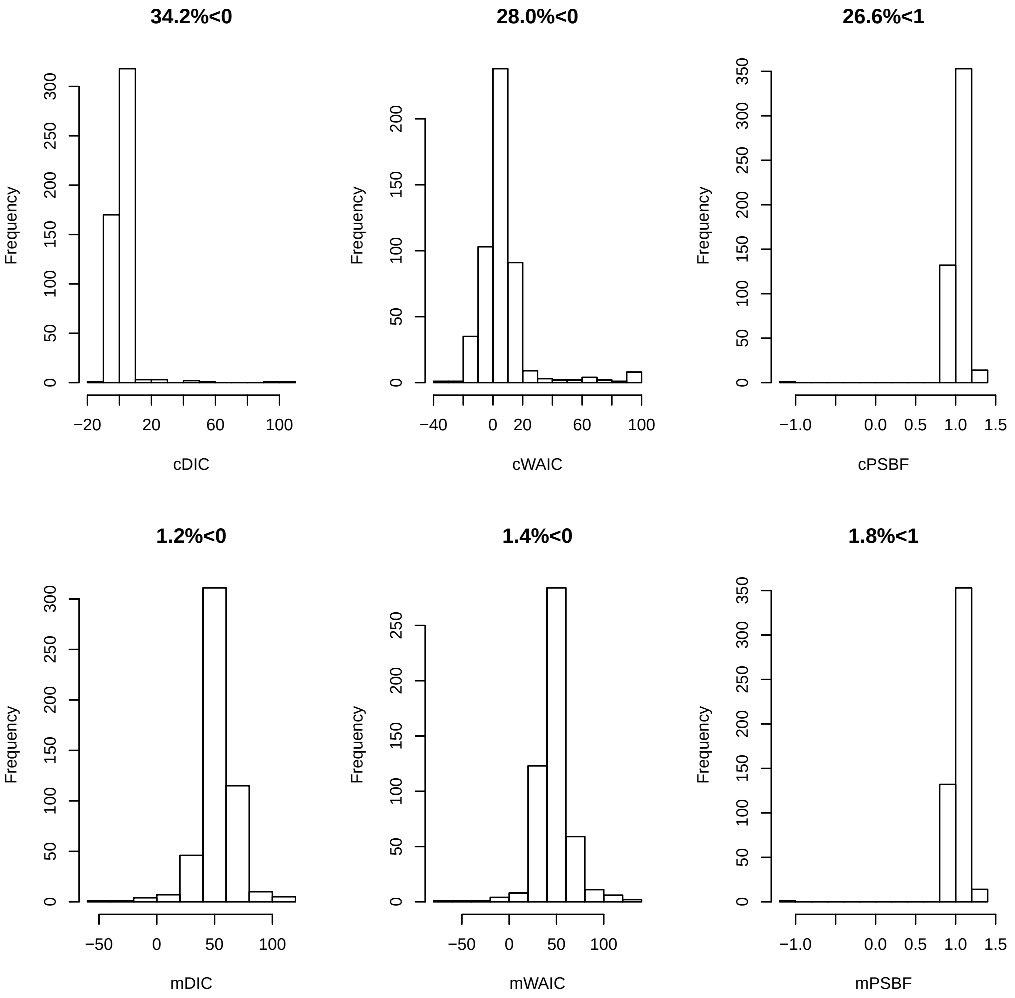

Figure 5.4: *Simulation 2: Bayesian selection criteria (top: conditional, bottom: marginal) under the PGMM (fitted model) minus the selection criteria under the true model from 300 simulated PMM data sets*

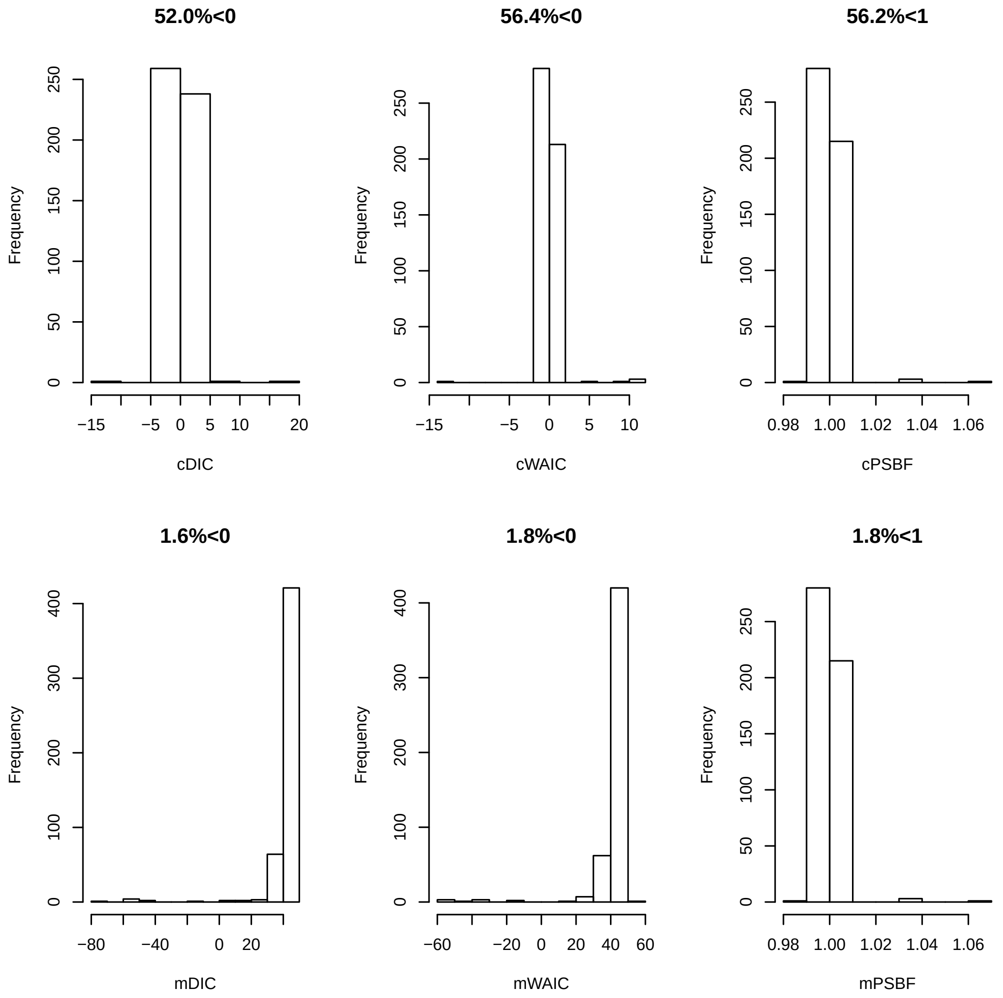

Figure 5.5: *Simulation 2: Bayesian selection criteria (top: conditional, bottom: marginal) under the ZIPMM (fitted model) minus the selection criteria under the true model from 300 simulated PMM data sets*

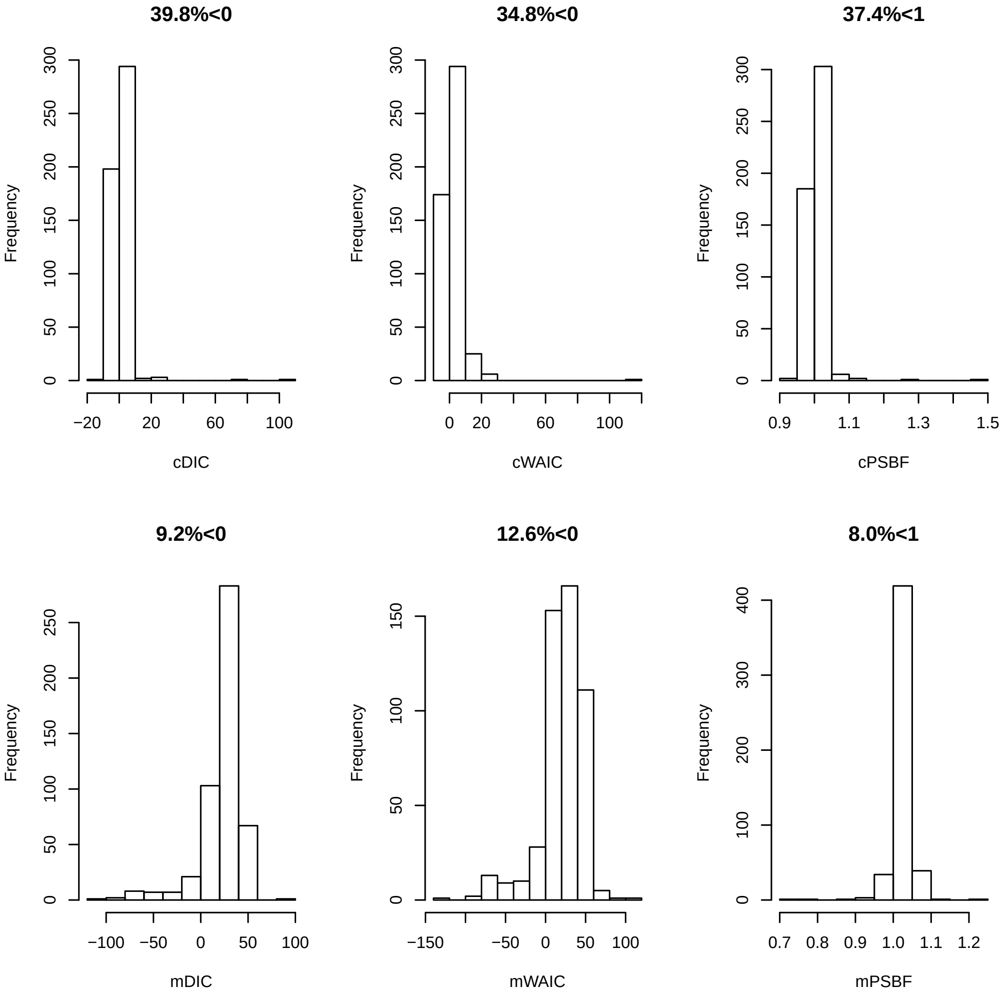

Figure 5.6: *Simulation 2: Bayesian selection criteria (top: conditional, bottom: marginal) under the Poisson model minus the selection criteria under the true model from 300 simulated PGMM data sets*

Table 5.7: *Simulation 4: Performance of the Bayesian model selection criteria when the number of parameters (p) is more than or equal to the number of subjects*

	$N < p$						$N = p$		
	$N = 4$			$N = 6$			$N = 9$		
Criteria	PGMM	PMM	ZIPMM	PGMM	PMM	ZIPMM	PGMM	PMM	ZIPMM
cDIC	70	178	52	85	191	64	67	195	38
cLPML	9	170	105	21	200	79	31	209	60
cWAIC	72	171	57	41	236	23	12	205	83
mDIC	27	181	92	4	288	8	8	291	1
mLPML	35	176	89	1	240	59	24	267	9
mWAIC	24	189	87	0	284	6	10	287	3

5.8 Conclusion

The study evaluates the performance of the marginal criteria in GLMM, especially with over-dispersed count data. Contrary to LMM, the closed-form likelihood of the GLMM is not available. Hence, there is a need for sampling techniques to compute the marginal criteria for non-closed form likelihoods. We set up simulation studies that compare the performance of two versions of the criteria for GLMM with longitudinal count data. We explored the ability of the criteria to select the true data-generating model in different scenarios and settings. One of such settings corresponds to the practical situation when the number of parameters is larger than the subjects. The results affirmed the superiority of the marginal criteria over its conditional counterparts.

This study also evaluates the advantages of the two sampling techniques used to compute the marginal criteria in GLMMs. The results show that importance sampling is advantageous for computing marginal criteria for a large number of subjects, number of observations/subject or both while the replication method is recommended for a smaller number of subjects (less than 200) and/or number of observations/subject (less than 10) for low computational burden.

Since the reliability of the replication methods is based on the number of replications needed to achieve convergence, we recommend a sensitivity analysis to obtain the best value of replications needed to ensure the trade-off between computational time and the deviance accuracy. To promote the usage of the marginal criteria in this context, we designed an R function that computes the marginal criteria using the replication method and importance sampling.

Chapter 6

Bayesian model selection for multilevel mediation models

This chapter has been submitted as:

Ariyo,O., Lesaffre, E., Verbeke, G., Huisman, M., Rijnhart, J.,Heymans, M., and Twisk, J. (2020). **Bayesian model selection for multilevel mediation models**. *Journal of Statistical Computation and Simulation* 0 (0), 1-16.

Abstract

Mediation analysis is a method often used for exploring the complex relationship between two variables through a third mediating variable. Several studies have focused on various multilevel-mediation (MLM) model designs and as such there is a need to select the most appropriate model for the specific research question. This paper aims to illustrate the performance of the deviance information criterion, the pseudo-Bayes factor, and the widely applicable information criterion in multilevel mediation models. Our focus will be on the comparison between the conditional criteria (given random effects) versus the marginal criteria (averaged over random effects) in selecting an appropriate data-generating model. Most of the previous works on the multilevel mediation models rely on the default software based on a lack of awareness of the differences between the marginal and conditional criteria. We demonstrate the superiority of the marginal version of the selection criteria over their conditional counterparts in the mediated longitudinal settings through an extensive simulation study and application to data from the LASA (Longitudinal Aging Study of the Amsterdam) study. In addition, we show that our self-written R function performs well for multilevel mediation models.

6.1 Introduction

Mediation analysis enables researchers to investigate the complex relationship between two variables through a third "mediating" variable. This indirect pathway through a mediating variable (or mediator) helps to explain how exposure affects an outcome (MacKinnon, 2008). The concept of mediation being used in the estimation of single-level mediation for independent subjects from random sampling (Hayes, 2017; MacKinnon, 2008) and multilevel mediation (McNeish, 2017; Zigler and Ye, 2019) has broad applications in both biomedical and social science research.

Multilevel data is usually encountered in medicine where patients are nested within hospitals, or repeated measurements are nested within patient. However, this type of multilevel data violates the assumption of independence necessary for traditional regression methods (Zigler and Ye, 2019). Hence, several authors have examined the use of mediation in multilevel data using multilevel modelling (MLM)(Krull and MacKinnon, 1999; Bauer et al., 2006; MacKinnon, 2008; Preacher et al., 2010; Rusá et al., 2018) and the multilevel structural equation model (MSEM) (Preacher et al., 2010; Lee, 2007; Yanuar et al., 2013; McNeish, 2017; Zigler and Ye, 2019).

Krull and MacKinnon (2001) established the terminology for a multilevel mediation design. They suggested the Predictor-Mediator-Outcome format, wherein a number indicates the level/amount of data where each variable is located. For example, a 1-1-1 means all three variables are measured at level-1. That is, level 1 represents the lowest measurement level, e.g., repeated measurements within a person, and that level 2 represents the cluster level, e.g., the subjects. In a longitudinal study this means that all variables are time-dependent. Other designs include the 2-1-1 design where the predictor is at level 2, and the outcome and mediator have been measured as level 1. This is for instance true in an RCT (randomized controlled trial) with more than one follow-up measurement. The intervention variable is not time-independent, i.e. measured at level 2. Figure 6.1 displays a 1-1-1 design for an MLM where the associations between the variables are split into between-cluster and within-cluster effects. For further details regarding this terminology, see Krull and MacKinnon (2001).

Several MLMs have been proposed in the literature; therefore evaluating the most appropriate model that fits the data best would be a useful exercise. However, in most mediation analyses, MLM models are typically evaluated with respect to bias, coverage probability, and power (Gao and Albert, 2018; Blood and Cheng, 2011; Wang et al., 2019; Zigler and Ye, 2019). Few authors make use of model selection criteria using a frequentist (Wu et al., 2018) or the Bayesian (Rusá et al., 2018) approach. Rusá et al. (2018) used the Widely-Applicable Information criteria (WAIC) to compare their flexible, moderated mediation model with competing models. However, the authors did not consider the marginal version of the WAIC. Here, we compare the performance of the conditional and marginal Bayesian model selection criteria to select the most appropriate MLM model for the data at hand.

The model selection criteria can be based on the conditional likelihood (given the random effects) resulting in conditional criteria, or it can be based on the marginal likelihood (integrating out the random effects), resulting in the marginal criteria. These marginal criteria measure the predictiveness of the model for all subjects from the same population in a future study while the conditional criteria measure the predictiveness of the model for the subjects included in the current study. In this paper, we will compare the performance of conditional and marginal versions of three popular Bayesian model selection criteria: the deviance information criteria (DIC), the Pseudo-Bayes factor (PSBF), and WAIC. While the DIC is popular in practice because it can be easily obtained with the use of traditionally popular Bayesian software packages (especially the BUGS software), PSBF and WAIC are becoming more popular albeit they are still unavailable in most classical Bayesian packages, except for WAIC which is provided by Stan (Carpenter et al., 2017).

The conditional versions of the three model selection criteria discussed above are the most popular and easiest to compute from the generated Markov Chain Monte Carlo (MCMC) samples. Consequently, the marginal versions of the selection criteria are never reported. However, it has been demonstrated theoretically and via simulation studies, that the use of conditional criteria often selects the wrong data-generating model (see, e.g. Chan and Grant, 2016a; Merkle et al., 2018; Ariyo et al., 2019b,a).

In practice, especially in medical and social science research, the researchers often rely on the default software, presumably because of a lack of awareness of the differences between the marginal and conditional criteria. In fact, in the analysis of mediation models, we are not aware of any previous work that distinguishes between the marginal and conditional criteria in the context of a multilevel mediation model. Hence, this paper illustrates the superiority of the marginal version of the selection criteria over their conditional counterparts in mediated longitudinal settings. As such, we have aimed to conduct simulation studies to illustrate the performance of the conditional and marginal criteria in selecting the true data-generating models under different scenarios: (i) when the mediation paths are allowed to vary randomly across clusters (see, e.g. Raudenbush and Bryk, 2002; Zigler and Ye, 2019), (ii) when the mediation paths are fixed across clusters (see e.g. McNeish, 2017), (iii) when the mediation path is zero, with "no mediation effects" (see, e.g. Zigler and Ye, 2019) and (iv) when the distributions of the mediation paths are misspecified.

The outline of the paper is as follows. In Section 6.2, we introduce the LASA data set. The basic concepts of mediation, moderation, and the combination thereof are summarised in Section 6.3. We briefly discuss the model selection criteria in Section 6.4. Different simulation settings and scenarios are presented in Section 6.5, while we illustrate the comparison between the conditional and marginal criteria on LASA data in Section 6.6. Finally, concluding remarks are given in Section 6.7.

6.2 The Longitudinal Aging Study Amsterdam (LASA)

LASA is a prospective cohort study intended to determine the predictors and consequences of ageing, more specifically of physical, cognitive, emotional, and social functioning in older adults (aged 55 to 85) in the Netherlands. The participants were sampled from the registries of urban as well as rural municipalities in different parts of the country. The baseline measurement took place in 1992/1993 and follow-up measurements have been conducted since then about every three years. The data collection consists of the main interview, a self-reported questionnaire, and a medical interview. The example in this paper was originally analyzed and published by Robitaile et al. (2013), who examined processing speed (M) as a mediator between age (X) and cognitive abilities (Y). Processing speed was based on a coding task adapted from the Alphabet Coding Task-15, Reasoning was based on the adapted version of the Raven Colored Progressive Matrices test, and the cognitive ability was based on the 15 Word Test. The cognition has been based on three different measures: 1) immediate recall, 2) delayed recall, and 3) reasoning as the outcome variable. More information can be found in Robitaile et al. (2013).

Each model was based on data from the first five waves of the LASA study. Respondents were excluded from the analyses if they had a score of 23 or lower on the Mini-Mental State Examination during any of the five waves (n = 798), or if they had missing education information ($n = 3$). The analytical cohort consisted of 2,306 respondents in the first wave, of which 1,883 also participated in the second wave (81.7%), with a further 1,562 in the third wave (83.0%), 1,300 in the fourth wave (83.2%), and 1,021 in the fifth wave (78.5%). Detailed information on the LASA can be found in Hoogendijk et al. (2016); Huisman et al. (2011); Robitaile et al. (2013).

Robitaile et al. (2013) applied a lower level mediation 1-1-1 model since all variables M, X, and Y were measured at level-1. Here, we aim to illustrate the performance of the conditional and marginal criteria in the context of a multilevel mediation model using a 1-1-1 MLM model.

6.3 Multilevel mediation model (MLM)

Before discussing the 1-1-1 multilevel mediation model in detail, we explain the basic framework of a multilevel mediation model. When an outcome Y and a predictor X are mediated by a mediator M, it means that M is correlated with X and explains the effect of X on Y. With a continuous outcome Y and a mediator M, a single-level mediation equation is given as

$$\begin{aligned} M &= \beta_1 + \alpha X + e_M \\ Y &= \beta_2 + \beta M + \tau' X + e_Y, \end{aligned} \tag{6.1}$$

where β_1, β_2 denote intercept for mediator and outcome, respectively and e_M, e_Y denote error terms in the equations. The direct effect of X on Y is denoted as $\tau^{'}$ and the indirect effect of $\mathbf{X}$ on $\mathbf{Y}$ through the mediator $\mathbf{M}$ is expressed as the product of α and β i.e $\alpha\beta$. Given that all variables are measured at level 1, the estimate of model parameters is straightforward using standard OLS-regression or maximum likelihood methods (Song, 2018; Preacher and Selig, 2012). However, the direct application of such techniques to multilevel data (such as mediation data) thereby ignoring the nested structure of the data will statistically bias (see also Raudenbush and Bryk, 2002; Tom et al., 2012) the estimates and the conclusions of the analysis. Hence, we consider a set of standard MLM equations predicting Y from X including a random effects structure (random intercept and slope(s)). Following the notation in Yuan and MacKinnon (2009), one can write a two-level lower mediation model (as given in Figure 6.1) with level 1 equations as:

$$\begin{aligned} M_{ij} &= \beta_{1j} + \alpha_j X_{ij} + e_{M_{ij}} \\ Y_{ij} &= \beta_{2j} + \beta_j M_{ij} + \tau_j^{'} X_{ij} + e_{Y_{ij}}, \end{aligned} \tag{6.2}$$

and level 2 is given as

$$\begin{aligned} \beta_{1j} &= \beta_3 + u_{1j} \\ a_j &= \alpha + u_{2j} \\ \beta_{2j} &= \beta_4 + u_{3j} \\ \beta_j &= \beta + u_{4j} \\ \tau_j^{'} &= \tau^{'} + u_{5j}, \end{aligned}$$

where $e_{M_{ij}}$ and $e_{Y_{ij}}$ are level 1 error terms for M and Y respectively; the parameters β_{1j} and β_{2j} are random intercepts, and α_j, β_j and $\tau_j^{'}$ are random slopes. The parameters β_2 and β_3 are population (or average) effects. For multilevel modelling, the first-level residuals $e_{M_{ij}}$ and $e_{Y_{ij}}$ are assumed to be independent and follow normal distribution, that is $e_{M_{ij}} \sim N(0, \sigma^2_{e_{M_{ij}}})$ and $e_{Y_{ij}} \sim N(0, \sigma^2_{e_{Y_{ij}}})$ and the second-level residuals $\mathbf{u}_j = (u_{1j}, u_{2j}, u_{3j}, u_{4j}, u_{5j})^T$ follow a multivariate normal distribution $\mathbf{u}_j \sim N(\mathbf{0}, \mathbf{D})$ where $\mathbf{D}$ is 5×5 covariance matrix.

In multilevel mediation, the average indirect effects in the population are often of primary interest. Yuan and MacKinnon (2009) gave the average indirect effects (applies to models with only random slopes) formula to be

$$ab = E(\alpha_j \beta_j) = \alpha\beta + \sigma_{\alpha_j \beta_j}, \tag{6.3}$$

where $\sigma_{\alpha_j \beta_j}$ denotes the covariance between α_j and β_j.

MacKinnon (2008) and Kenny et al. (2003) also showed that the total effect in a fully random, lower mediated multilevel model is

$$c = \tau^{'} + \alpha\beta + \sigma_{\alpha_j \beta_j} \tag{6.4}$$

and the relative average indirect effect can be expressed as

$$ab/c = \frac{\alpha\beta + \sigma_{\alpha_j \beta_j}}{\tau' + \alpha\beta + \sigma_{\alpha_j \beta_j}}. \tag{6.5}$$

Equation (6.5) is often referred to as the proportion mediated in the mediation analysis literature (Ditlevsen et al., 2005; Ananth, 2019). This statistic has some important disadvantages. For example, the proportion mediated effect cannot be used when the mediation model is inconsistent (i.e., the direct and indirect effect have a different sign), which is actually the case in the LASA data example (see Robitaile et al., 2013). In these situations, the proportion mediated can exceed 1 and can be below 0 and the interpretation may become meaningless (as the limits of a proportion are 0 and 1).

In practice, the heterogeneity in the causal effects across level 2 units may be of scientific interest. For example, in Section 6.6, we analyse the LASA dataset to investigate whether the processing speed mediates between age and cognitive abilities in older adults. The importance of random effects in the lower level mediation (1-1-1 model in particular) was pointed out first in Kenny et al. (2003). For model represented in Equation (6.2) to be estimable and ensure that the mediational effects are unbiased, some assumptions are required as given below:

1. The predictors X_{ij} must be uncorrelated with the random intercepts and slopes and with β_{2j}, β_j, $\tau_j^{'}$ and $e_{Y_{ij}}$.

2. The residuals e_{M_j} and $e_{Y_{ij}}$ are normally distributed with mean zero and uncorrelated with each other.

3. The level 1 residuals are uncorrelated with the random effects i.e e_{M_j} is uncorrelated with β_{1j}, a_j, β_{2j}, β_j and $\tau_j^{'}$.

It is important to note that some of these assumptions may not hold in practice. In this paper, we considered the performance of the conditional and marginal selection criteria when these assumptions are violated.

6.4 Bayesian model selection

We considered three different Bayesian model selection criteria for evaluating a multilevel mediation model: PSBF, DIC and WAIC. We further distinguished

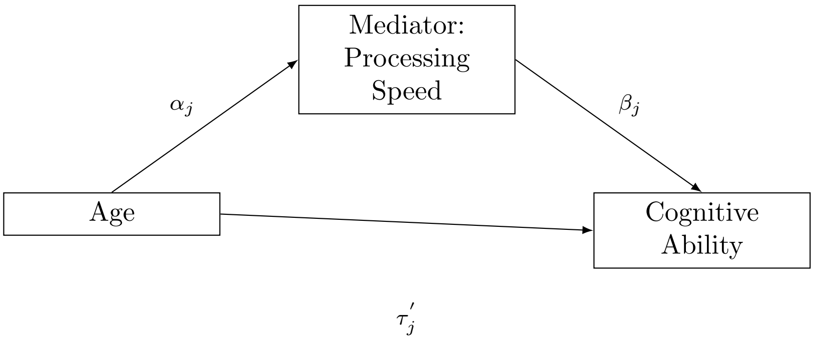

Figure 6.1: *LASA data example: Diagram for a two-level mediation model in which the effect of Age on Cognitive Ability is partially mediated by Processing Speed: Age affect Processing Speed (path α_j), Processing Speed affect Cognitive Ability (path β_j), and Age affect Cognitive Ability (path $\tau_j^{'}$)*

between the marginal and conditional version of these criteria. For MLM, let $\mathbf{\Theta}$ represent all the model parameters, the distinction is that the marginal MLM includes the fixed effects (i.e the intercept for mediator and outcome, when assume fixed) and the parameters making up the covariance matrix of the random effects. Conversely, the conditional MLM includes the random effects (i.e the direct and indirect pathway of the model) in the $\mathbf{\Theta}$.
Further, we denote the collected (longitudinal) mediated outcomes by $\mathbf{y}$ and the obtained covariate values by the matrix $\mathbf{X}$ moderated by $\mathbf{M}$. The posterior distribution is $p(\boldsymbol{\theta} \mid \mathbf{y}, \mathbf{X}, \mathbf{M}) = p(\mathbf{y} \mid \mathbf{\Theta}, \mathbf{X}) p(\mathbf{\Theta}) / p(\mathbf{y} \mid \mathbf{X}, \mathbf{M})$. When the closed-form of this posterior distribution does not exit then it is appropriate to use MCMC methods. Namely, K (dependent) values $\mathbf{\Theta}^1, \ldots, \mathbf{\Theta}^K$ are sampled from the posterior distribution. The true posterior summary measures can then be approximated by their sampled versions.
Recently, some authors have compared the performance of these criteria in different application studies (see e.g Dey et al., 2019; Millar, 2018). However, there is still no consensus about the best criterion for model selection in a Bayesian context. For the distinction between the performance of the marginal against the conditional criteria, other authors have shown that marginal criteria outperform the conditional criteria in most settings for LMMs with some extensions (Ariyo et al., 2019a,b) and GLMMs (Quintero and Lesaffre, 2018; Millar, 2018; Ariyo et al., 2020). This is also true for an item response model (Li et al., 2016; Millar, 2018; Merkle et al., 2018). However, to our knowledge, the superiority of the marginal criteria over the conditional criteria has not been demonstrated in the mediation analysis literature. For reasons of completeness, we will (briefly) discuss the three Bayesian model selection criteria in the subsequent sections.

6.4.1 The pseudo Bayes factor

Model comparison using Bayes' factors requires the computation of the marginal likelihood of two competing models. Given a model $\mathcal{M}$ and model parameters $\theta_{\mathcal{M}}$, we assume that the data $y_1, \ldots, y_n$ are conditionally independent. The marginal likelihood is given by:

$$p(\mathbf{y}|\mathcal{M}) = \int_{\theta_{\mathcal{M}}} \prod_{i=1}^{n} p(y_i|\theta_{\mathcal{M}}, \mathcal{M}) p(\theta_{\mathcal{M}}) d\theta_{\mathcal{M}}. \tag{6.6}$$

However, equation (6.6) is not analytically available in general. Therefore, Geisser and Eddy (1979) suggested replacing (6.6) by the pseudo marginal likelihood

$$\hat{p}(\mathbf{y}|\mathcal{M}) = \prod_{i=1}^{n} p(y_i|y_{-i}, \mathcal{M}), \tag{6.7}$$

where $\prod_{i=1}^{n} p(y_i|y_{-i}, \mathcal{M})$ is the i^{th} conditional predictive ordinate (CPO_i) and the predictive density calculated at the observed y_i given y_{-i}, which is the set of all data except the i^{th} observation. The pseudo-Bayes factor is then obtained by taking the ratio $\widehat{p}(\mathbf{y}\,|\,\mathcal{M}_1)/\widehat{p}(\mathbf{y}\,|\,\mathcal{M}_2)$ to evaluate the preference of model $\mathcal{M}_1$ over model $\mathcal{M}_2$. Low value of this ratio reflect preference of model $\mathcal{M}_2$ based on the current data. In practice, one often evaluates the logarithm of expression (6.7), leading to the log pseudo marginal likelihood for model $\mathcal{M}_r$ equal to $\text{LPML}_r = \sum_{i=1}^{n} \log(\text{CPO}_{i,r})$ where

$$\text{CPO}_{r,\ell} \approx \left[\frac{1}{K} \sum_{k=1}^{K} \frac{1}{p(\mathbf{y}_i \mid \boldsymbol{\theta}_r^k, \mathcal{M}_r)} \right]^{-1}$$

and $\boldsymbol{\theta}_r^k$ represents the model parameters for model $\mathcal{M}_r$.

6.4.2 The deviance information criterion

Deviance is defined as $D(\boldsymbol{\Theta}) = -2\log p(\mathbf{y}|\boldsymbol{\theta})$. The deviance informative criterion (DIC) is then defined as $\text{DIC} = -2\log p(\mathbf{y}|\bar{\boldsymbol{\theta}}) + 2p_{DIC}$, where p_{DIC} corresponds to the effective number of parameters, given by

$$p_{DIC} = -2\ E_{\boldsymbol{\theta}|\mathbf{y}}[\log p(\mathbf{y}|\boldsymbol{\theta})] + 2\log[p(\mathbf{y}|\bar{\boldsymbol{\theta}})]. \tag{6.8}$$

Two versions of p_{DIC} are generally used (Gelman et al., 2014; Spiegelhalter et al., 2014): (i) p_{DIC} in (6.8) which is considered to be numerically stable, and

(ii) $p_{DIC_2} = 2\ Var_{\boldsymbol{\theta}|\mathbf{y}}[(\mathbf{y}|\bar{\boldsymbol{\theta}})]$ which has the advantage of being always positive (Gelman et al., 2014). Consequently, Celeux et al. (2006) suggested several forms for the DIC that can be used for different hierarchical models and Ariyo et al. (2019b) have compared the performance of the marginal and conditional versions of these DIC versions and have shown that there are inconsistencies in the performance of the conditional versions of different forms of DIC whereas the marginal versions perform similarly. Despite its popularity and availability in Bayesian software, DIC has been criticised, see Spiegelhalter et al. (2014) for details. For instance, DIC is not invariant to non-linear transformations of $\boldsymbol{\theta}$ and negative value for p_{DIC} can occur in some cases.

6.4.3 Watanabe-Akaike information criterion

The Watanabe-Akaike information criterion (WAIC) is a Bayesian version of the AIC (Watanabe, 2010) and a worthy successor of DIC (Spiegelhalter et al., 2014) as it uses the posterior predictive distribution of the data to estimate the out-of-sample predictive accuracy of the model. The WAIC is then defined as

$$\text{WAIC} = -2\widehat{\text{lppd}} + 2p_{WAIC},$$

where p_{WAIC} corresponds to an estimate of the effective number of parameters given by

$$p_{WAIC} = 2\ \sum_{i=1}^{n} \left[\log\left(\frac{1}{K} \sum_{k=1}^{K} p(\mathbf{y}_i|\boldsymbol{\theta}^k) \right) - \frac{1}{K} \sum_{k=1}^{K} \log\ p(\mathbf{y}_i|\boldsymbol{\theta}^k) \right].$$

and $\widehat{\text{lppd}}$ which can be estimated using an MCMC sample from the posterior distribution as

$$\widehat{\text{lppd}} = \sum_{i=1}^{n} \log \left[\frac{1}{K} \sum_{k=1}^{K} p(\mathbf{y}_i|\boldsymbol{\theta}^k) \right].$$

Similar to DIC, a model with smaller WAIC is preferred. One advantage of WAIC is its invariability to the scale of the model parameters, which implies that WAIC does not change when $\boldsymbol{\theta}$ is replaced by $\boldsymbol{\psi} = h(\boldsymbol{\theta})$, where h a strictly monotone function. For all the three model selection criteria, the model with the smallest value is preferred.

6.5 Simulation studies

We performed simulation studies in order to illustrate the performance of the conditional and marginal criteria in multilevel mediation models (MLMs). We opted to use 1-1-1 mediation models with random slopes for our simulation studies because this kind of model has been a model of interest in the fundamental work behind multilevel mediation (see Preacher et al., 2011) and has been commonly used in empirical studies (McNeish, 2017). Notwithstanding its simplicity, a 1-1-1 design allows for the modelling of both the between and within components of the indirect effects and pathways (Zhang et al., 2009). Additionally, we were motivated by previous investigations using LASA data (see Robitaile et al., 2013).

Our aim is to study the performance of the conditional and marginal versions of DIC, WAIC and PSBF in MLMs. We further evaluated the impact of cluster sizes and the impact of different number of observations/subject points. Furthermore, we provided an R function to compute the marginal criteria for MLMs using the MCMC output from any Bayesian package. Additionally, this function computes the conditional criteria and is available as supporting material for the paper

6.5.1 Simulation study 1

We first conducted a simple simulation study based on the example in Kenny et al. (2003) and subsequently used in Bauer et al. (2006). Based on the example in Kenny et al. (2003), we generated 500 datasets based on model (6.2) with the following parameters: the random intercept for the mediator β_{2j} had a mean 0 and variance of 0.6, i.e $\beta_{2j} \sim N(0, 0.6)$, the random intercept for the response $\beta_{1j} \sim N(0, 0.4)$. These two random intercept β_{1j}, β_{2j} were normally distributed. The level 1 variance of response (σ^2_{eY}) and the mediator (σ^2_{eM}) were set to 0.45 and 0.65 respectively, the α_j and β_j paths were normally distributed with a mean of 0.6 and a variance of 0.16, while $\tau^{'}_j \sim N(0.2, 0.4)$. The covariance between α_j and β_j was 0.113 i.e $\sigma_{\alpha_j\beta_j} = 0.113$, yielding a correlation of 0.706. We assume that neither α_j and β_j was correlated with $\tau^{'}_j$. We further simulated a predictor X_{ij} from $X_{ij} = \bar{X}_i + e_{X_{ij}}$, where $\bar{X}_i \sim N(0, 1)$ and $e_{X_{ij}} \sim N(0, 1)$.

In addition to Equation (6.2) as the true model, we fitted two alternative models: (i) Equation (6.2) without mediation effects $Y_{ij} = \beta_{2j} + \tau^{'}_j X_{ij} + e_{Y_{ij}}$, (ii) Equation (6.2) without direct effect component (i.e $\tau^{'}_j = 0$), to evaluate the ability of the conditional and marginal criteria to select the data-generating model. We further varied the number of clusters under study from 10 to 100 (10, 25, 50,100), as this number of clusters is similar to those previously used in the literature (Bauer et al., 2006; McNeish, 2017). We also set the number of observations per level 2 units to $m_j = 4, 8, 16, \text{and}, 36,$ which is consistent with those used by Krull and MacKinnon (2001) and Bauer et al. (2006). All models in this simulation study were estimated based on three chains of 10,000 iterations (discarding the first 5,000) and with a thinning equal to 10. The convergence of the MCMC samples was assessed using the Brooks-Gelman-Rubin (BGR) diagnostic, and in cases where the BGR was larger than 1.1, a new MCMC sample was selected with 10,000 extra iterations until convergence was obtained.

Table 6.1 displays the performance of the conditional and marginal criteria for a 1-1-1 multilevel model in identifying the correct model when the mediation pathway was allowed to vary in a normal, random fashion across clusters. As expected, the performance of both versions of the criteria get better as sample sizes increase. However, this increase is less obvious when the sample size is higher than 25. Similarly, as the number of observations increases from 8 to 32, the performance of both versions of the criteria are less obvious. An increase in observation units of more than 8 has a minimal impact on the performance of both versions of the criteria. As such, these results are in agreement with the results previously described in the literature. For instance, McNeish (2017) concluded that, if the right precautions have been taken, only few clusters and observations are needed to provide reliable results.

Additionally, McNeish and Stapleton (2016) suggested that 20 clusters with five or more observations per cluster might be sufficient if the model is estimated with restrictive maximum likelihood. Overall, the marginal criteria outperformed the conditional ones, which is in line with other results previously obtained in the literature (see Ariyo et al., 2019b,a; Chan and Grant, 2016a; Quintero and Lesaffre, 2018; Merkle et al., 2018).

6.5.2 Simulation study 2

We evaluated the performance of the conditional and marginal criteria when the distribution of the mediation paths was misspecified. As such, we generated 500 data sets based on Equation (6.2) with the modification that the distributions of the random slopes α_j and β_j are generated from $\chi^2(3)$ distribution, which has skewness 1.63 and kurtosis 4, which closely resembles the skew-normal distribution (see e.g Wang et al., 2004). In addition to the data-generating model, we fit two alternative models: A model with mediation paths that are (i) assumed normal and (ii) assumed skew-normal.

The percentage of the times that each criterion selected a data generating model is displayed in Table 6.2. The results demonstrate that when the mediation paths are assumed to be skewed, the performance of both criteria is better than if normality is assumed for the mediation pathways. These results show that the assumption that the sampling distribution of the average mediation paths (especially the indirect effects) follows a normal distribution might be unrealistic. This is why several authors warn researchers against making these assumptions for hypothesis testing (Song, 2018; MacKinnon et al., 2004; Hayes and Scharkow, 2013; Preacher and Selig, 2012) and recommend approaches that relax the normality assumption, such as using bootstrap confidence interval (Efron and Tibshirani, 1994; MacKinnon et al., 2004, 2007) and Monte Carlo (MC) simulations (Preacher and Selig, 2012) among others. Regardless of the assumptions for the mediation pathways, the marginal criteria display superior performance compared to the conditional criteria.

6.5.3 Simulation study 3

In the 1-1-1 model, the mediation paths can be estimated as fixed effects. However, when the paths are not allowed to randomly vary across clusters, a large number of clusters may lead to convergence problems while also diminishing the quality of the estimates (McNeish, 2017). Here, we have illustrated the performance of the conditional and marginal DIC, WAIC and PSBF when the mediation pathways are not allowed to be random. Furthermore, we evaluated this condition under a variety of clusters and a different number of observations per level 2 units, as described in Section 6.5.1. As such, we generated 500 datasets based on Equation (6.1) with the following value for each parameters: (i) indirect effect component $\alpha = \beta = 0.40$, and (ii) direct effect component $\tau^{'} = 0.40$. Additionally, $\beta_1 = 0.45$ and $\beta_2 = 0.45$.

We fitted three alternative models: (i) Model (6.1) (ii) Model (6.1) without mediation effect (i.e $\beta = 0$), (iii) Model (6.1) without direct effect component (i.e., $\tau^{'} = 0$). The performance of the marginal and conditional criteria in identifying the correct data-generating model when the distribution of the mediation paths have been fixed has been presented in Table 6.3. The results show that regardless of the sample sizes, there is no difference between the performance of the conditional and marginal criteria.

6.6 Analysis of LASA data

Here, we illustrate the performance of the conditional and marginal criteria using the LASA data described in Section 6.2. We fitted three different models: (i) Model "A" based on equation (6.2), (ii) Model "B" based on Equation (6.9), (iii) Model "C" based on Equation (6.10) and (iv) Model "D" based on Equation (6.2) without mediation effects.

$$\begin{aligned} M_{ij} &= \beta_{1j} + \alpha_j X_{ij} + e_{M_{ij}} \\ Y_{ij} &= \beta_{2j} + \beta_j M_{ij} + e_{Y_{ij}}. \end{aligned} \tag{6.9}$$

$$\begin{aligned} M_{ij} &= \beta_{1j} + \alpha_j X_{ij} + e_{M_{ij}} \\ Y_{ij} &= \beta_{2j} + \beta_j M_{ij} + \tau_j^{'} X_{ij} + e_{Y_{ij}}, \end{aligned} \tag{6.10}$$

where $\binom{a_j}{b_j} \sim (\mathbf{0}, \mathbf{D})$ and $\mathbf{D}$ is a 2×2 covariance matrix.
For each of the models, we calculate the level-specific indirect effects (Equation(6.3)) and the level-specific total effects (Equation (6.4)).
The priors used has been described in Section 6.5.1. We used 10,000 iterations, which after discarding the first 5,000 as burn-in and thinning was set to 10. The convergence of the MCMC samples was assessed using the BGR criteria. In addition to the estimates of model parameters, we compute the conditional and marginal criteria for each model. The results are displayed in Table 6.4.
It can be observed that the conditional criteria support Model "C" and "D". These models assume that the level two parameters β_{1j} and β_{2j} are fixed. This seems to be an incorrect model since the 1-1-1 model assumed these parameters to be random across subjects. In contrast, the marginal criteria favor model "A". This appears therefore to be the most appropriate 1-1-1 mediation model. This underscores the superior performance of the marginal criteria in selecting of the most appropriate model.

Table 6.1: *The percentage of times the conditional and marginal criteria select the true model when mediation path are random vary across clusters under different sample sizes and number of observation per units.*

	Sample size					
No of observation/units	Criteria	10	25	50	100	200
4	cDIC	65.4	71.4	77.2	77.8	78.2
	cWAIC	62.0	68.6	73.4	74.6	75.6
	cPSBF	62.0	67.6	73.6	73.4	78.2
	mDIC	71.4	79.8	84.4	86.0	89.2
	mWAIC	72.2	80.2	84.8	85.8	88.6
	mPSBF	71.6	80.4	82.6	82.8	89.0
8	cDIC	69.4	76.8	78.2	78.2	79.8
	cWAIC	67.6	77.4	79.6	80.0	83.6
	cPSBF	67.4	77.6	78.8	80.2	82.6
	mDIC	79.4	80.4	84.4	86.2	88.0
	mWAIC	73.6	84.8	86.8	86.4	89.2
	mPSBF	73.8	88.6	88.8	89.0	89.8
16	cDIC	76.0	79.4	80.0	80.2	81.4
	cWAIC	75.6	78.2	80.2	80.0	81.4
	cPSBF	73.6	74.8	79.2	80.0	80.8
	mDIC	86.6	89.4	89.8	89.8	89.8
	mWAIC	89.6	89.8	89.6	90.0	90.6
	mPSBF	86.0	89.8	89.8	90.2	90.8
32	cDIC	76.2	78.8	80.4	82.4	84.2
	cWAIC	74.8	78.6	81.2	84.0	84.6
	cPSBF	73.0	78.0	80.6	83.6	84.8
	mDIC	86.2	89.0	90.0	90.8	93.0
	mWAIC	87.8	89.2	90.2	93.8	93.4
	mPSBF	85.8	89.0	90.0	94.2	95.0

Table 6.2: *The percentage of times the conditional and marginal criteria select true model when mediation are not normally distributed across clusters under different sample sizes*

Path distribution	Sample	10	25	50	100	200
	Criteria					
Skew-normal	cDIC	69.4	76.8	78.2	78.2	79.8
	cWAIC	67.6	77.4	79.6	80.0	83.6
	cPSBF	67.4	77.6	78.8	80.2	82.6
	mDIC	79.4	80.4	84.4	86.2	88.0
	mWAIC	73.6	84.8	86.8	86.4	89.2
	mPSBF	73.8	88.6	88.8	89.0	89.8
Normal	cDIC	45.4	54.4	67.2	72.8	78.2
	cWAIC	42.0	55.6	73.4	72.6	75.6
	cPSBF	42.0	57.6	63.6	70.4	78.2
	mDIC	51.4	68.8	76.4	81.0	87.2
	mWAIC	52.2	76.2	73.8	83.8	85.6
	mPSBF	51.6	78.0	70.6	82.8	88.0

Table 6.3: *The percentage of time selection criteria select true model when the mediation pathways are fixed.*

Sample size					
Criteria	10	25	50	100	200
cDIC	86.3	93.6	98.6	99.2	100.0
cWAIC	83.0	94.4	98.6	99.0	100.0
cPSBF	83.6	94.0	98.6	99.6	100.0
mDIC	86.3	93.6	98.6	99.2	100.0
mWAIC	83.0	94.4	98.6	99.0	100.0
mPSBF	83.6	94.0	98.6	99.6	100.0

Table 6.4: *LASA data set: Posterior mean (estimates of a lower level mediation model), 95% probability interval and the conditional and marginal criteria under four fitted models.*

	Model A			Model B			Model C			Model D		
Effects	Estimates	25%	95%	Estimates	25%	95%	Estimates	25%	95%	Estimates	25%	95%
$\alpha\beta$	0.4306	0.4160	0.4437	0.3771	0.3742	0.8021	0.2013	0.1980	0.2046	0.2690	0.2571	0.2778
τ	0.3655	0.3612	0.3696	0.3771	0.3742	0.8021	0.2013	0.1980	0.2046	0.1570	0.1477	0.1699
α	0.6804	0.6778	0.6830	0.6809	0.6786	0.6832	0.3895	0.3825	0.3962	0.3900	0.3831	0.3969
β	0.6323	0.6115	0.6519	0.5534	0.5489	0.5581	0.5145	0.5098	0.5190	0.6877	0.6620	0.7041
$\tau^{'}$	-0.0623	-0.0783	-0.0508	-	-	-	-0.0130	-0.1980	-0.2046	-0,1120	-0,1477	-0,1699
$\sigma^2_{\alpha_j}$	0.0004	0.0004	0.0005	0.0004	0.0003	0.0005	0.0037	0.0032	0.0043	0.0037	0.0032	0.0010
$\sigma^2_{\beta_j}$	0.0017	0.0015	0.0020	0.0061	0.0002	0.0002	0.0015	0.0012	0.0018	0.0012	0.0010	0.0014
$\sigma^2_{\alpha_j\beta_j}$	0.0003	0.0002	0.0004	0.0003	0.0003	0.0004	0.0009	0.0008	0.0010	0.0008	0.0006	0.0011
$\sigma^2_{\tau^{'}_j}$	-0.0561	-0.0783	-0.0508	-	-	-	-	-	-	-0.0034	-0.0031	-0.0042
cDIC		-32226.53			-32226.89			-32234.27			-32229.60	
cWAIC		-32429.84			-32430.11			-32436.01			-32432.23	
cLPPD		-16032.55			-16029.91			-16033.89			-16032,05	
mDIC		-2907.24			-2626.53			-2892.17			-2569.05	
mWAIC		-2916.62			-2614.84			-2887.56			-2567.78	
mLPPD		-1283.78			-1246.82			-1243.78			-1284.47	

6.7 Conclusion

We compared three Bayesian selection criteria in the context of multilevel mediation models. Our focus was on illustrating the performance of the conditional and marginal criteria in selecting the true, data-generating model under different distributional assumptions for mediation pathways. The results of the simulation studies demonstrated the superior performance of the marginal criteria when the mediation pathways are assumed to be random. The conditional criteria often select over-specified mediation pathways, while the marginal criteria select the correct model often. Conversely, when the mediation pathways are assumed to be fixed for the 1-1-1 mediation model, the performance of the marginal and conditional criteria is essentially the same. Both the conditional and marginal criteria prefer (often) the correct model when the mediation pathways are fixed. However, the motivation for assuming fixed or random pathways should be based on the research question. The result from the analysis of the LASA dataset also confirmed the results obtained in the simulation studies as the results of the marginal criteria were consistent with the summary statistics.

These results confirm the results of Ariyo et al. (2019b,a) for LMMs, Chan and Grant (2016a) for volatility models, Millar (2018); Li et al. (2016) for item response models, Merkle et al. (2018) for latent variable model and Quintero and Lesaffre (2018); Ariyo et al. (2020) for GLMM. To encourage the applied researchers to use the marginal criteria, we provide an R function that computes not only the marginal criteria but also the conditional criteria with minimal computational effort. The function is available at https://ibiostat.be/online-resources/bayesian.

The choice of a 1-1-1 multilevel mediation model with three variables was motivated by its popularity in the literature and motivated by the LASA dataset. We believe that the results are valid for other multilevel mediation models as well. When more variables are involved in mediation analysis, the clusters' effects are likely to impact the performance of model selection criteria. Hence, further studies could derive the likelihood for more complex mediation models to evaluate the performance of Bayesian model selection in more complex settings.

Chapter 7

General Discussion

7.1 General Discussion and Future Research

In this thesis, we compared the performance of the conditional and marginal versions of DIC, WAIC, and PSBF when it comes to identifying the accurate data-generating models. We have demonstrated the superiority of the marginal criteria over their conditional counterparts in a variety of settings and scenarios. As such, we conclude here with a summary of the main research contributions of the previous chapters, the limitations of our research, and the possibilities for future studies.

7.2 General Discussion

We have shown the advantages of using the marginal criteria for the most essential LMM. In Chapter 3, we compared the performance of the conditional and marginal versions of DIC, WAIC, and PSBF in LMM with the extension to skew-normal and skew-t distributions for either (or both) random effects and measurement error. The simulation studies were explored via three scenarios. Scenario 1: Assume that the random effects structure is correct and considered the models that differ in their fixed part. In Scenario 2, assume that the fixed structure is correct and consider a model that changes in the random effects structure. For Scenario 3, assume that the true fixed and random effects are to be selected jointly. Further in Scenario 3, the following were assumed: (i) different scales for the covariates, (ii) distributional assumptions not satisfied by either (or both) random-effects and measurement error, (iii) the nature of measurement error (homoscedastic or heteroscedastic), and (iv) the incorrect random effects structure.

Furthermore, we examined two selection strategies: minimum value and absolute difference strategies. Additionally, we evaluated the model selection performance for alternative versions of DIC and WAIC. We further examined the performance of the marginal criteria with two longitudinal clinical data sets (Potthoff & Roy growth data and Jimma infant growth study), while also illustrating the use of these methods on non-clinical data. The results show the superior performance of the marginal criteria over the conditional criteria in all scenarios and settings. Despite the superiority of the marginal criteria, a large number of applied statisticians have continually neglected to utilise them. Therefore, we were prompted to write an R function that makes the marginal selection criteria easier to compute.

Building upon the previous results, in Chapter 4, we examined the effect of vague priors on the Bayesian model selection criteria. In this chapter, we showed that vague priors influence the performance of both versions of the criteria, but only have a minimal impact on the marginal version. We also showed that the conditional criteria performed inconsistently and often selected over-specified models. In contrast, the marginal versions were less sensitive to the choice of priors and most often prefer the correct data-generating model. For the prior of the covariance matrix of the random effects, we evaluated the conjugate Inverse-Wishart prior, the hierarchical version of this prior, the joint prior and some selected separation priors. We showed that the joint prior, the hierarchical prior and the separation priors outperform the conjugate Inverse-Wishart for the longitudinal mixed model involving two or more random effects. As such, we recommend these priors when the use of the conditional version is essential.

In Chapter 5, we extended the settings those as mentioned above into a GLMM where the likelihood is not analytically available. We employed two sampling methods (replication method and importance sampling) to compute the marginal criteria for the GLMM. For the method of replication, we expressed the marginal likelihood component as an expectation of the conditional distribution. Replication method enables the marginalisation of the likelihood by generating replicate samples from the density of the latent variables, which ought to be integrated out to estimate such expectation. We addressed the major setback of this approach in this chapter using the importance sampling proposed in Quintero and Lesaffre (2018) to reduce the computational complexity for large-sample situations. These approaches can be easily implemented from the MCMC output of any Bayesian package. The results affirmed the superior performance of the marginal criteria in this context again, and additionally, we also provided an R function to makes this computation attractive.

Finally, in Chapter 6, we explored the selection criteria in the context of multilevel mediation models. We illustrated the performance of the conditional and marginal criteria in selecting real data-generating models under different distribution assumptions and for mediation pathways. Motivated by its popularity and the LASA dataset, we explored 1-1-1 multilevel mediation models. However, we believe that our conclusion is valid for other types of multilevel mediation models. Using the LASA dataset, we investigated the relationship between age, processing speed, and cognitive functioning. Furthermore, we also evaluated the performance of different approaches to the estimation of mediational effects using Bayesian model selection tools. The results affirmed the superiority of the marginal criteria over the conditional criteria.

The simulation studies in this thesis were motivated by four sets of longitudinal medical data and one set of longitudinal non-medical data. We implemented all of the procedures in JAGS running interactively with R using the rjags package. The relevant codes can be found online at https://ibiostat.be/online-resources/bayesian and in the appendix of this thesis.

7.3 Limitations and Future Research

Although the simulation studies, we considered provided an excellent overview of the performance of the conditional and marginal criteria, nevertheless there is room for improvement. As such, we present below different possibilities that could be regarded as future research for each of the topics that we considered in this thesis.

In evaluating the performance of the conditional and marginal criteria in Chapter 3, we used an LMM with some extensions. From the results for the simulation studies and the application data, we can discern that our conclusion is valid for longitudinal data with missing values in the response and or covariate; however, the main assumption here is that the missing data mechanism is ignorable. Although this assumption may be valid for numerous applications, it is not possible to empirically evaluate the tenability of the assumption. The performance of both versions of the criteria can be assessed by regressing the missing data indicator on the random coefficients of the longitudinal model.

Also, extensions of our R function to joint modelling of the longitudinal measurements and survival processes in the shared parameter framework where a set of random-effects is assumed to induce their interdependence (or non-normal random effects) would be a useful exercise. Further studies may be carried out in evaluating the performance of both versions of the criteria using the multiple linear mixed-effects models with autoregressive (AR) random errors within the subjects. This can be further extended to other covariate structures for within-subjects random error.

In Chapter 4, we focused on the effects of vague priors on the variance of the random effects, as the choice of the prior for the residual variances is considerably less important than the prior for the variances of latent variables (Polson et al., 2012). For the intercept and regression parameters, evaluating the impact of the alternative priors such as the t-distribution, which has been proposed as a prior for logistic models by Gelman et al. (2008), and as an error distribution (e.g., robust growth curve models; Zhang et al., 2013) could be a useful exercise to check the superiority of the marginal criteria over the conditional counterparts. The idea of using the joint prior of Chapter 4.5.2 can be extended to a wider class of random effects. A further approach to the choice of the prior distribution is to use priors with uniform shrinkage (Spiegelhalter, 2001; Daniels and Kass, 1999; Natarajan and Kass, 2000).

As for the model selection when the likelihood is not analytically available, as discussed in Chapter 5, a further study wherein the marginal and conditional versions of DIC, WAIC and PSBF are based on the subject-level joint likelihood is recommended. This joint likelihood does not appear to have an immediate interpretation via leave-one-out-cross-validation because the data has been modelled jointly within an unknown parameter. Further study in this area would be very useful.

Finally, in Chapter 6, we illustrated the superiority of the marginal criteria in the context of a 1-1-1 multilevel mediation model with three variables. The results can be further extended to other multilevel mediation models. Further studies utilizing numerous additional variables might illustrate that the effects of the cluster are likely to affect the performance of the Bayesian model selection criteria. Additionally, the impact of several simultaneous mediation models in the performance of the Bayesian model selection criteria would be most useful. Further studies could derive the likelihood for more complex mediation models to evaluate the performance of the selection criteria in this context.

Scientific acknowledgement

Financial support from the Tertiary Education Trust Fund (TETFund)-AS&D of the Federal Government of Nigeria through the Federal University of Agriculture Abeokuta Nigeria is greatly acknowledged.

The computational resources and services used in this thesis were provided by the VSC (Flemish Supercomputer Center), funded by the Research Foundation - Flanders (FWO) and the Flemish Government department EWI.

Finally, we thank the editors and referees of the different articles for their constructive comments to improve the manuscripts during the peer-review process.

Conflict of Interest

There is no conflict of interest.

Personal Contribution

Below we list the personal contribution for each of the research articles.
Chapter 3: derived the closed form likelihood for the distributions for the random effects used in the article, wrote the R functions which computed the marginal criteria for the models, performed the simulation studies, conducted the analysis of the Nigerian Indigenous Chicken data set in JAGS, and interpretation of the results of the analysis.
Chapter 4: overview of different prior proposals in the literature for variances of the random effects and measurement errors, performed the simulation studies, conducted the analysis of the Nigerian Indigenous Chicken data set in JAGS, and interpretation of the results of the analysis.
Chapter 5: wrote the non-closed form expression of the GLMM especially count data with overdispersion, wrote the R functions that compute the marginal criteria (using both replication ans importance sampling methods) for the models in this context, performed the simulation studies, and responsible for the analysis of the Epilepsy data sets in JAGS.
Chapter 6: Simulation studies, analysis of LASA (Longitudinal Ageing Study of Amsterdam) in JAGS, interpretation of the results of the analysis.

Bibliography

Adeleke, M., Peters, S., Ozoje, M., Ikeobi, C., Bamgbose, A., and Adebambo, O. A. (2011). Genetic parameter estimates for body weight and linear body measurements in pure and crossbred progenies of Nigerian indigenous chickens. *Livestock Research for Rural Development*, 23(1):1–7.

Adrion, C. and Mansmann, U. (2012). Bayesian model selection techniques as decision support for shaping a statistical analysis plan of a clinical trial: an example from a vertigo phase III study with longitudinal count data as primary endpoint. *BMC Medical Research Methodology*, 12(1):137.

Al-Rawwash, M. and Pourahmadi, M. (2013). Gaussian estimation of regression and correlation parameters in longitudinal data. *Journal of the Association of Arab Universities for Basic and Applied Sciences*, 13(1):28–34.

Alvarez, I., Niemi, J., and Simpson, M. (2014). Bayesian inference for a covariance matrix. *arXiv preprint arXiv:1408.4050*.

Ananth, C. V. (2019). Proportion mediated in a causal mediation analysis: how useful is this measure? *BJOG: An International Journal of Obstetrics & Gynaecology*, 126(8):983–983.

Anderson, S. J. (2018). Longitudinal Study Designs. *Handbook of Research Methods in Health Social Sciences*, pages 1–20.

Aregay, M., Shkedy, Z., and Molenberghs, G. (2013). A hierarchical Bayesian approach for the analysis of longitudinal count data with overdispersion: a simulation study. *Computational Statistics & Data Analysis*, 57(1):233–245.

Aregay, M., Shkedy, Z., and Molenberghs, G. (2015). Comparison of additive and multiplicative Bayesian models for longitudinal count data with overdispersion parameters: a simulation study. *Communications in Statistics-Simulation and Computation*, 44(2):454–473.

Arellano-Valle, R., Bolfarine, H., and Lachos, V. (2007). Bayesian inference for skew-normal linear mixed models. *Journal of Applied Statistics*, 34(6):663–682.

Arellano-Valle, R. B. and Genton, M. G. (2005). On fundamental skew distributions. *Journal of Multivariate Analysis*, 96(1):93–116.

Arellano-Valle, R. B. and Genton, M. G. (2010). Multivariate extended skew-t distributions and related families. *Metron*, 68(3):201–234.

Ariyo, O., Lesaffre, E., Verbeke, G., and Quintero, A. (2019a). Model selection for Bayesian linear mixed models with longitudinal data: Sensitivity to the choice of priors. *Communications in Statistics - Simulation and Computation*, 0(0):1–25.

Ariyo, O., Lesaffre, E., Verbeke, G., and Quintero, A. (2020). Bayesian model selection for longitudinal count data. *Communications for Statistical Applications and Methods*, 0(0):1–24.

Ariyo, O., Quintero, A., Muñoz, J., Verbeke, G., and Lesaffre, E. (2019b). Bayesian model selection in linear mixed models for longitudinal data. *Journal of Applied Statistics*, 47(5):890–913.

Baey, C., Cournède, P.-H., and Kuhn, E. (2017). Likelihood ratio test for variance components in nonlinear mixed effects models. *arXiv preprint arXiv:1712.08567*.

Barnard, J., McCulloch, R., and Meng, X.-L. (2000). Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statistica Sinica*, pages 1281–1311.

Bauer, D. J., Preacher, K. J., and Gil, K. M. (2006). Conceptualizing and testing random indirect effects and moderated mediation in multilevel models: new procedures and recommendations. *Psychological Methods*, 11(2):142.

Berchialla, P., Baldi, I., Notaro, V., Barone-Monfrin, S., Bassi, F., and Gregori, D. (2009). Flexibility of Bayesian generalized linear mixed models for oral health research. *Statistics in Medicine*, 28(28):3509–3522.

Bernardo, J. M. (1980). A Bayesian analysis of classical hypothesis testing. *Trabajos de Estadistica Y de Investigacion Operativa*, 31(1):605–647.

Blood, E. A., Cabral, H., Heeren, T., and Cheng, D. M. (2010). Performance of mixed effects models in the analysis of mediated longitudinal data. *BMC Medical Research Methodology*, 10(1):16.

Blood, E. A. and Cheng, D. M. (2011). The use of mixed models for the analysis of mediated data with time-dependent predictors. *Journal of Environmental and Public Health*, 2011.

Booth, J. G., Casella, G., Friedl, H., and Hobert, J. P. (2003). Negative binomial loglinear mixed models. *Statistical Modelling*, 3(3):179–191.

Breslow, N. and Clayton, D. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88(421):9–25.

Breslow, N. E. (1984). Extra-Poisson variation in log-linear models. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 33(1):38–44.

Brooks, S. P. and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7(4):434–455.

Burton, P. R., Tiller, K. J., Gurrin, L. C., Cookson, W. O., Musk, A. W., and Palmer, L. J. (1999). Genetic variance components analysis for binary phenotypes using generalized linear mixed models (GLMMs) and Gibbs sampling. *Genetic Epidemiology*, 17(2):118–140.

Cai, B. and Dunson, D. B. (2006). Bayesian covariance selection in generalized linear mixed models. *Biometrics*, 62(2):446–457.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1).

Celeux, G., Forbes, F., Robert, C., and Titterington, D. (2006). Deviance information criteria for missing data models. *Bayesian Analysis*, 1(4):651–673.

Chan, J. and Grant, A. (2016a). On the observed-data deviance information criterion for volatility modeling. *Journal of Financial Econometrics*, 14(4):772–802.

Chan, J. C. and Grant, A. L. (2016b). Fast computation of the deviance information criterion for latent variable models. *Computational Statistics & Data Analysis*, 100:847–859.

Chen, H.-C. and Wehrly, T. E. (2016). Approximate uniform shrinkage prior for a multivariate generalized linear mixed model. *Journal of Multivariate Analysis*, 145:148–161.

Chen, Z. and Dunson, D. B. (2003). Random effects selection in linear mixed models. *Biometrics*, 59(4):762–769.

Christensen, F. G. W. (2017). *New Approaches to Model Selection in Bayesian Mixed Modeling*. PhD thesis, UC Irvine.

Christiansen, C. L. and Morris, C. N. (1997). Hierarchical Poisson regression modeling. *Journal of the American Statistical Association*, 92(438):618–632.

Coakley, E. S. and Rokhlin, V. (2013). A fast divide-and-conquer algorithm for computing the spectra of real symmetric tridiagonal matrices. *Applied and Computational Harmonic Analysis*, 34(3):379–414.

Congdon, P. (2005). *Bayesian Models for Categorical Data*. John Wiley & Sons.

Daniels, M. J. and Kass, R. E. (1999). Nonconjugate Bayesian Estimation of Covariance Matrices and Its Use in Hierarchical Models. *Journal of the American Statistical Association*, 94(448):1254–1263.

Daniels, M. J. and Pourahmadi, M. (2002). Bayesian analysis of covariance matrices and dynamic models for longitudinal data. *Biometrika*, 89(3):553–566.

De Santis, F. and Spezzaferri, F. (1997). Alternative Bayes factors for model selection. *Canadian Journal of Statistics*, 25(4):503–515.

Dean, C. and Nielsen, J. D. (2007). Generalized linear mixed models: a review and some extensions. *Lifetime Data Analysis*, 13(4):497–512.

Demirhan, H. and Hamurkaroglu, C. (2011). On a multivariate log-gamma distribution and the use of the distribution in the Bayesian analysis. *Journal of Statistical Planning and Inference*, 141(3):1141–1152.

Demirhan, H. and Kalaylioglu, Z. (2015). Joint prior distributions for variance parameters in Bayesian analysis of normal hierarchical models. *Journal of Multivariate Analysis*, 135:163–174.

Dey, D. K., Chen, M.-H., and Chang, H. (1997). Bayesian approach for nonlinear random effects models. *Biometrics*, 53(4):1239–1252.

Dey, S., Delampady, M., and Gopalaswamy, A. M. (2019). Bayesian model selection for spatial capture–recapture models. *Ecology and Evolution*, 9(20):11569–11583.

Ditlevsen, S., Christensen, U., Lynch, J., Damsgaard, M. T., and Keiding, N. (2005). The mediation proportion: a structural equation approach for estimating the proportion of exposure effect on outcome explained by an intermediate variable. *Epidemiology*, pages 114–120.

Efron, B. and Tibshirani, R. J. (1994). *An Introduction to the Bootstrap*. CRC press.

Engel, B. and Keen, A. (1994). A simple approach for the analysis of generalized linear mixed models. *Statistica Neerlandica*, 48(1):1–22.

Fan, T.-H., Wang, Y.-F., and Zhang, Y.-C. (2014). Bayesian model selection in linear mixed effects models with autoregressive (p) errors using mixture priors. *Journal of Applied Statistics*, 41(8):1814–1829.

Faught, E., Wilder, B., Ramsay, R., Reife, R., Kramer, L., Pledger, G., and Karim, R. (1996). Topiramate placebo-controlled dose-ranging trial in refractory partial epilepsy using 200-, 400-, and 600-mg daily dosages. *Neurology*, 46(6):1684–1690.

Fitzmaurice, G. M. (1997). Model selection with overdispersed data. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 46(1):81–91.

Funatogawa, I. (2017). *Longitudinal Data Analysis: Autoregressive Linear Mixed Effects Models*. Springer.

Gao, T. and Albert, J. M. (2018). Bayesian causal mediation analysis with multiple ordered mediators. *Statistical Modelling*, pages 1471082–18798067.

Gayle, V. and Lambert, P. (2018). *What is Quantitative Longitudinal Data Analysis?* Bloomsbury Publishing.

Geisser, S. and Eddy, W. F. (1979). A predictive approach to model selection. *Journal of the American Statistical Association*, 74(365):153–160.

Gelfand, A. and Dey, D. (1994). Bayesian model choice: Asymptotics and exact calculations. *Journal of the Royal Statistical Society - Series B*, 56(3):501–514.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85(410):398–409.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis*, 1(3):515–534.

Gelman, A., Carlin, J., Stern, H., and Rubin, D. (2004a). *Bayesian Data Analysis*. Chapman and Hall.

Gelman, A. and Hill, J. (2007). *Data Analysis using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

Gelman, A., Hwang, J., and Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, 24(6):997–1016.

Gelman, A., Rubin, D. B., et al. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–472.

Gelman, A., Stern, H. S., Carlin, J. B., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2004b). *Bayesian Data Analysis*. Chapman and Hall/CRC.

Gelman, A., Van Dyk, D., Huang, Z., and Boscardin, J. (2008). Using redundant parameterizations to fit hierarchical models. *Journal of Computational and Graphical Statistics*, 17(1):95–122.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *Journal of Applied Statistics*, 20(5-6):25–62.

George, E. I. and McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889.

Gong, L., Flegal, J. M., Spindler, S. R., and Mote, P. L. (2015). Bayesian model selection on linear mixed-effects models for comparisons between multiple treatments and a control. *arXiv preprint arXiv:1509.07510*.

Hartigan, J. et al. (1996). Locally uniform prior distributions. *The Annals of Statistics*, 24(1):160–173.

Hayes, A. F. (2017). *Introduction to Mediation, Moderation, and Conditional Process Analysis: A Regression-based Approach*. Guilford Publications.

Hayes, A. F. and Scharkow, M. (2013). The relative trustworthiness of inferential tests of the indirect effect in statistical mediation analysis: Does method really matter? *Psychological Science*, 24(10):1918–1927.

Hinde, J. and Demétrio, C. G. (1998). Overdispersion: models and estimation. *Computational Statistics & Data Analysis*, 27(2):151–170.

Hoffman, L. (2015). *Longitudinal Analysis: Modeling within-person Fluctuation and Change*. Routledge.

Hoogendijk, E. O., Deeg, D. J., Poppelaars, J., van der Horst, M., van Groenou, M. I. B., Comijs, H. C., Pasman, H. R. W., van Schoor, N. M., Suanet, B., Thomése, F., et al. (2016). The longitudinal aging study amsterdam: cohort update 2016 and major findings. *European Journal of Epidemiology*, 31(9):927–945.

Howe, E. J., Buckland, S. T., Després-Einspenner, M.-L., and Kühl, H. S. (2019). Model selection with overdispersed distance sampling data. *Methods in Ecology and Evolution*, 10(1):38–47.

Huang, A., Wand, M. P., et al. (2013). Simple marginally noninformative prior distributions for covariance matrices. *Bayesian Analysis*, 8(2):439–452.

Huang, Y. and Dagne, G. (2011). A Bayesian approach to joint mixed-effects models with a skew-normal distribution and measurement errors in covariates. *Biometrics*, 67:260–269.

Huang, Y. and Dagne, G. (2012). Bayesian semiparametric nonlinear mixed-effects joint models for data with skewness, missing responses, and measurement errors in covariates. *Biometrics*, 68(3):943–953.

Huisman, M., Poppelaars, J., Horst, M., Beekman, A., Brug, J., Tilburg, T. G., and Deeg, D. (2011). Cohort profile: The Longitudinal Aging Study Amsterdam. *International Journal of Epidemiology*, 40:868–876.

Hurtado Rúa, S. M., Mazumdar, M., and Strawderman, R. L. (2015). The choice of prior distribution for a covariance matrix in multivariate meta-analysis: a simulation study. *Statistics in Medicine*, 34(30):4083–4104.

Iddi, S. and Doku-Amponsah, K. (2016). Statistical model for overdispersed count outcome with many zeros: an approach for marginal inference. *South African Statistical Journal*, 50(2):313–337.

Ivanova, A., Molenberghs, G., and Verbeke, G. (2014). A model for overdispersed hierarchical ordinal data. *Statistical Modelling*, 14(5):399–415.

Ivanova, A., Molenberghs, G., and Verbeke, G. (2016). Mixed models approaches for joint modeling of different types of responses. *Journal of Biopharmaceutical Statistics*, 26(4):601–618.

Ivanova, A., Molenberghs, G., and Verbeke, G. (2017). Mechanism for missing data incorporated in joint modelling of ordinal responses. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 66(5):1049–1064.

Jeffreys, H. (1961). Theory of probability.,(Oxford University Press: Oxford, UK).

Joe, H. (2006). Generating random correlation matrices based on partial correlations. *Journal of Multivariate Analysis*, 97(10):2177–2189.

Jones, R. H. (2011). Bayesian information criterion for longitudinal and clustered data. *Statistics in Medicine*, 30(25):3050–3056.

Kalaylioglu, Z. and Demirhan, H. (2017). A joint Bayesian approach for the analysis of response measured at a primary endpoint and longitudinal measurements. *Statistical Methods in Medical Research*, 26(6):2885–2896.

Kass, R. E., Natarajan, R., et al. (2006). A default conjugate prior for variance components in generalized linear mixed models (comment on article by Browne and Draper). *Bayesian Analysis*, 1(3):535–542.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795.

Kenny, D. A., Korchmaros, J. D., and Bolger, N. (2003). Lower level mediation in multilevel models. *Psychological Methods*, 8(2):115.

Krull, J. L. and MacKinnon, D. P. (1999). Multilevel mediation modeling in group-based intervention studies. *Evaluation Review*, 23(4):418–444.

Krull, J. L. and MacKinnon, D. P. (2001). Multivarite modeling of individual and group level mediated effects. *Multivariate Behavioral Research*, 36:249–277.

Lachos, V. H., Castro, L. M., and Dey, D. K. (2013). Bayesian inference in nonlinear mixed-effects models using normal independent distributions. *Computational Statistics & Data Analysis*, 64:237–252.

Lachos, V. H., Ghosh, P., and Arellano-Valle, R. B. (2010). Likelihood based inference for skew-normal independent linear mixed models. *Statistica Sinica*, 20(1):303–322.

Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, pages 963–974.

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, 34(1):1–14.

Lambert, P. C., Sutton, A. J., Burton, P. R., Abrams, K. R., and Jones, D. R. (2005). How vague is vague? A simulation study of the impact of the use of vague prior distributions in MCMC using WinBUGS. *Statistics in Medicine*, 24(15):2401–2428.

Lawless, J. F. (1987). Negative binomial and mixed Poisson regression. *Canadian Journal of Statistics*, 15(3):209–225.

Lee, S.-Y. (2007). *Structural Equation Modeling: A Bayesian approach*, volume 711. John Wiley & Sons.

Lesaffre, E., Asefa, M., and Verbeke, G. (1999). Assessing the goodness-of-fit of the Laird and Ware model an example: the Jimma Infant Survival Differential Longitudinal Study. *Statistics in Medicine*, 18(7):835–854.

Lesaffre, E. and Lawson, A. (2012). *Bayesian Biostatistics (Statistics in Practice)*. Wiley: Chichester.

Lesaffre, E., Todem, D., and Verbeke, G. (2000). Flexible modelling of the covariance matrix in a linear random effects model. *Biometrical Journal*, 42(7):807–822.

Lewandowski, D., Kurowicka, D., and Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis*, 100(9):1989–2001.

Li, B., Bruyneel, L., and Lesaffre, E. (2013a). A multivariate multilevel gaussian model with a mixed effects structure in the mean and covariance part. *Statistics in Medicine*, 33:1877–1899.

Li, L., Qiu, S., Zhang, B., and Feng, C. X. (2016). Approximating cross-validatory predictive evaluation in Bayesian latent variable models with integrated IS and WAIC. *Statistics and Computing*, 26(4):881–897.

Li, Y., Zeng, T., and Yu, J. (2013b). Robust deviance information criterion for latent variable models. *CAFE Research Paper No. 13.19*, Available at https://papers.ssrn.com/sol3/papers.cfm?abstract$_id = 2316341.$

Littell, R. C., Milliken, G. A., Stroup, W. W., Wolfinger, R. D., and Schabenberger, O. (2007). *SAS for mixed models*. SAS institute.

Lu, G. and Ades, A. (2009). Modeling between-trial variance structure in mixed treatment comparisons. *Biostatistics*, 10(4):792–805.

Lunn, D. J., Thomas, A., Best, N., and Spiegelhalter, D. (2000). WinBUGS-A Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*, 10(4):325–337.

MacKinnon, D. (2008). *Introduction to Statistical Mediation Analysis*. Routledge.

MacKinnon, D. P., Fritz, M. S., Williams, J., and Lockwood, C. M. (2007). Distribution of the product confidence limits for the indirect effect: Program PRODCLIN. *Behavior Research Methods*, 39(3):384–389.

MacKinnon, D. P., Lockwood, C. M., and Williams, J. (2004). Confidence limits for the indirect effect: Distribution of the product and resampling methods. *Multivariate Behavioral R esearch*, 39(1):99–128.

Mason, A., Richardson, S., and Best, N. (2012). Two-pronged strategy for using DIC to compare selection models with non-ignorable missing responses. *Bayesian Analysis*, 7(1):109–146.

McArdle, J. J. and Nesselroade, J. R. (2014). *Longitudinal data analysis using structural equation models.* American Psychological Association.

McCullagh, P. (1989). *Generalized linear models*. Routledge.

McCulloch, C., Searle, S., and Neuhaus (2008). *Generalized, Linear and Mixed Models*. John Wiley & Sons.

McNeish, D. (2017). Multilevel mediation with small samples: A cautionary note on the multilevel structural equation modeling framework. *Structural Equation Modeling: A Multidisciplinary Journal*, 24(4):609–625.

McNeish, D. M. and Stapleton, L. M. (2016). The effect of small sample size on two-level model estimates: A review and illustration. *Educational Psychology Review*, 28(2):295–314.

Merkle, E., Furr, D., and Rabe-Hesketh, S. (2018). Bayesian model assessment: Use of conditional vs marginal likelihoods. *arXiv preprint arXiv:1802.04452*.

Millar, R. (2009). Comparison of hierarchical Bayesian models for overdispersed count data using DIC and Bayes' factors. *Biometrics*, 65(3):962–969.

Millar, R. B. (2018). Conditional vs marginal estimation of the predictive loss of hierarchical models using WAIC and cross-validation. *Statistics and Computing*, 28(2):375–385.

Molenberghs, G., Renard, D., and Verbeke, G. (2002). A review of generalized linear mixed models. *Journal de la Société Française de Statistique*, 143(1-2):53–78.

Molenberghs, G., Thijs, H., Jansen, I., Beunckens, C., Kenward, M. G., Mallinckrodt, C., and Carroll, R. J. (2004). Analyzing incomplete longitudinal clinical trial data. *Biostatistics*, 5(3):445–464.

Molenberghs, G. and Verbeke, G. (2005). *Model for Discrete Longitudinal Data*. Springer Science+Business Media, Inc. New York.

Molenberghs, G., Verbeke, G., and Demétrio, C. G. (2007). An extended random-effects approach to modeling repeated, overdispersed count data. *Lifetime Data Analysis*, 13(4):513–531.

Molenberghs, G., Verbeke, G., Demétrio, C. G., Vieira, A. M., et al. (2010). A family of generalized linear models for repeated measures with normal and conjugate random effects. *Statistical Science*, 25(3):325–347.

Mukerjee, R. and Ghosh, M. (1997). Second-order probability matching priors. *Biometrika*, 84(4):970–975.

Müller, S., Scealy, J. L., Welsh, A. H., et al. (2013). Model selection in linear mixed models. *Statistical Science*, 28(2):135–167.

Natarajan, R. and Kass, R. E. (2000). Reference Bayesian methods for generalized linear mixed models. *Journal of the American Statistical Association*, 95(449):227–237.

O'Malley, A. J. and Zaslavsky, A. M. (2008). Domain-level covariance analysis for multilevel survey data with structured nonresponse. *Journal of the American Statistical Association*, 103(484):1405–1418.

Pinheiro, J. C. and Bates, D. M. (1996). Unconstrained parametrizations for variance-covariance matrices. *Statistics and Computing*, 6(3):289–296.

Plummer, M. (2011). Cannot invert matrix. [Online; posted 11-November-2011].

Polson, N. G., Scott, J. G., et al. (2012). On the half-Cauchy prior for a global scale parameter. *Bayesian Analysis*, 7(4):887–902.

Potthoff, R. and Roy, S. (1964). A generalized multivariate analysis of variance model useful especially for growth curve problems. *Biometrika*, 5:313–326.

Pourahmadi, M. (1999). Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation. *Biometrika*, 83(3):677–690.

Preacher, K. J. and Selig, J. P. (2012). Advantages of Monte Carlo confidence intervals for indirect effects. *Communication Methods and Measures*, 6(2):77–98.

Preacher, K. J., Zhang, Z., and Zyphur, M. J. (2011). Alternative methods for assessing mediation in multilevel data: The advantages of multilevel SEM. *Structural Equation Modeling*, 18(2):161–182.

Preacher, K. J., Zyphur, M. J., and Zhang, Z. (2010). A general multilevel SEM framework for assessing multilevel mediation. *Psychological Methods*, 15(3):209.

Quintero, A. and Lesaffre, E. (2017). Multilevel covariance regression with correlated random effects in the mean and variance structure. *Biometrical Journal*, 59(5):1047–1066.

Quintero, A. and Lesaffre, E. (2018). Comparing hierarchical models via the marginalized deviance information criterion. *Statistics in Medicine*, 37(16):2440–2454.

Raftery, A. E., Newton, M. A., Satagopan, J. M., and Krivitsky, P. N. (2007). Estimating the integrated likelihood via posterior simulation using the harmonic mean identity. *Bayesian Statistics*, 8:1–45.

Raiffa, H. and Schlaifer, R. (1961). Applied statistical decision theory.

Rakhmawati, T. W., Molenberghs, G., Verbeke, G., and Faes, C. (2016). Local influence diagnostics for hierarchical count data models with overdispersion and excess zeros. *Biometrical Journal*, 58(6):1390–1408.

Raudenbush, S. W. and Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and data analysis methods*, volume 1. Sage.

Rizopoulos, D. (2011). Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data. *Biometrics*, 67(3):819–829.

Rizopoulos, D. (2012). *Joint Models for Longitudinal and Time-to-Event Data: With applications in R*. Chapman and Hall/CRC.

Robitaile, A., Piccinin, A., Muniz-Terrera, G, ., Hoffman, L., Johansson, B., Deeg, D., J, M., Aartsen, H., Comijs, C., and Hofer, S. (2013). Simple conditions for the convergence of the Gibbs sampler and Metropolis-Hastings algorithms. *Psychology and Aging*, 28(4):887–901.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the Royal Statistical Society - Series B*, 71:319–392.

Rusá, Š., Komárek, A., Lesaffre, E., and Bruyneel, L. (2018). Multilevel moderated mediation model with ordinal outcome. *Statistics in Medicine*, 37(10):1650–1670.

Säfken, B., Rügamer, D., Kneib, T., and Greven, S. (2018). Conditional model selection in mixed-effects models with cAIC4. *arXiv preprint arXiv:1803.05664*.

Sahu, S. K., Dey, D. K., and Branco, M. D. (2003). A new class of multivariate skew distributions with applications to Bayesian regression models. *Canadian Journal of Statistics*, 31(2):129–150.

Schervish, M. J. (2012). *Theory of Statistics*. Springer Science & Business Media.

Schnell, P. M., Tang, Q., Offen, W. W., and Carlin, B. P. (2016). A Bayesian credible subgroups approach to identifying patient subgroups with positive treatment effects. *Biometrics*, 72(4):1026–1036.

Schuurman, N. K., Grasman, R. P. P. P., and Hamaker, E. L. (2016). A comparison of Inverse-Wishart prior specifications for covariance matrices in multilevel autoregressive models. *Multivariate Behavioral Research*, 51(2-3):185–206. PMID: 27028576.

Scurrah, K. J., Palmer, L. J., and Burton, P. R. (2000). Variance components analysis for pedigree-based censored survival data using generalized linear mixed models (GLMMs) and Gibbs sampling in BUGS. *Genetic Epidemiology*, 19(2):127–148.

Song, H. (2018). A primer on multilevel mediation models for egocentric social network data. *Communication Methods and Measures*, 12(1):1–24.

Spiegelhalter, D., Best, N., Carlin, N., and van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society - Series B*, 64(4):583–639.

Spiegelhalter, D., Best, N., Carlin, N., and van der Linde, A. (2014). The deviance information criterion: 12 years on. *Journal of the Royal Statistical Society - Series B*, 76(3):485–493.

Spiegelhalter, D., Thomas, A., Best, N., and Lunn, D. (2003). *WinBUGS User Manual*, 1.4 edition.

Spiegelhalter, D. J. (2001). Bayesian methods for cluster randomized trials with continuous responses. *Statistics in Medicine*, 20(3):435–452.

Spiegelhalter, D. J., Abrams, K. R., and Myles, J. P. (2004). *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*. John Wiley & Sons.

Srivastava, M. S. and Kubokawa, T. (2010). Conditional information criteria for selecting variables in linear mixed models. *Journal of Multivariate Analysis*, 101(9):1970–1980.

Tokdar, S. T. and Kass, R. E. (2010). Importance sampling: a review. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(1):54–60.

Tokuda, T., Goodrich, B., Van Mechelen, I., Gelman, A., and Tuerlinckx, F. (2011). Visualizing distributions of covariance matrices. *Columbia Univ., New York, USA, Tech. Rep*, pages 18–21.

Tom, A., Bosker, T. A. S. R. J., and Bosker, R. J. (2012). *Multilevel analysis: an introduction to basic and advanced multilevel modeling*. Sage.

Tran, M.-N., Scharth, M., Pitt, M. K., and Kohn, R. (2016). Importance sampling squared for Bayesian inference in latent variable models. *arXiv preprint arXiv:1309.3339*.

Tuerlinckx, F., Rijmen, F., Verbeke, G., and De Boeck, P. (2006). Statistical inference in generalized linear mixed models: A review. *British Journal of Mathematical and Statistical Psychology*, 59(2):225–255.

Vaida, F. and Blanchard, S. (2005). Conditional Akaike information for mixed-effects models. *Biometrika*, 92(2):351–370.

van Smeden, M., de Groot, J. A., Moons, K. G., Collins, G. S., Altman, D. G., Eijkemans, M. J., and Reitsma, J. B. (2016). No rationale for 1 variable per 10 events criterion for binary logistic regression analysis. *BMC Medical Research Methodology*, 16(1):163.

van Smeden, M., Moons, K. G., de Groot, J. A., Collins, G. S., Altman, D. G., Eijkemans, M. J., and Reitsma, J. B. (2019). Sample size for binary logistic prediction models: Beyond events per variable criteria. *Statistical Methods in Medical Research*, 28(8):2455–2474. PMID: 29966490.

Verbeke, G. and Lesaffre, E. (1996). A linear mixed-effects model with heterogeneity in the random-effects population. *Journal of the American Statistical Association*, 91(433):217–221.

Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. Springer Series in Statistics, New York.

Wand, M. P., Ormerod, J. T., Padoan, S. A., Frühwirth, R., et al. (2011). Mean field variational Bayes for elaborate distributions. *Bayesian Analysis*, 6(4):847–900.

Wang, J., Boyer, J., Genton, M. M. G., et al. (2004). A note on an equivalence between chi-square and generalized skew-normal distributions. *Statistics & Probability Letters*, 66(4):395–398.

Wang, Y.-B., Chen, Z., Goldstein, J. M., Buck Louis, G. M., and Gilman, S. E. (2019). A Bayesian regularized mediation analysis with multiple exposures. *Statistics in Medicine*, 38(5):828–843.

Warton, D. I. (2005). Many zeros does not mean zero inflation: comparing the goodness-of-fit of parametric models to multivariate abundance data. *Environmetrics: The Official Journal of the International Environmetrics Society*, 16(3):275–289.

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11:3571–3594.

Watanabe, S. (2013). A wide applicable Bayesian information criterion. *Journal of Machine Learning Reserach*, 14:3571–3594.

Wei, J. and Higgins, J. (2013). Bayesian multivariate meta-analysis with multiple outcomes. *Statistics in Medicine*, 32(17):2911–2934.

Wolfinger, R. and O'Connell, M. (1993). Generalized linear mixed models a pseudo-likelihood approach. *Journal of Statistical Computation and Simulation*, 48(3-4):233–243.

Wu, W., Carroll, I. A., and Chen, P.-Y. (2018). A single-level random-effects cross-lagged panel model for longitudinal mediation analysis. *Behaviour Research Methods*, 50(5):2111–2124.

Yanuar, F., Ibrahim, K., and Jemain, A. A. (2013). Bayesian structural equation modeling for the health index. *Journal of Applied Statistics*, 40(6):1254–1269.

Yau, K. K., Wang, K., and Lee, A. H. (2003). Zero-inflated negative binomial mixed regression modeling of over-dispersed count data with extra zeros. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 45(4):437–452.

Yuan, Y. and MacKinnon, D. P. (2009). Bayesian mediation analysis. *Psychological Methods*, 14(4):301.

Zhang, Z., Lai, K., Lu, Z., and Tong, X. (2013). Bayesian inference and application of robust growth curve models using student's t distribution. *Structural Equation Modeling: A Multidisciplinary Journal*, 20(1):47–78.

Zhang, Z., Zyphur, M. J., and Preacher, K. J. (2009). Testing multilevel mediation using hierarchical linear models: Problems and solutions. *Organizational Research Methods*, 12(4):695–719.

Zigler, C. K. and Ye, F. (2019). A comparison of multilevel mediation modeling methods: Recommendations for applied researchers. *Multivariate Behavioral Research*, 54(3):338–359.

Appendix A

R functions

S.1 JAGS code

Here is the JAGS file for the illustating example.

```
model
{
#for random effects mean and variance
zero[1] <- 0
zero[2] <- 0
#Likelihood
for (k in 1:K){
mu[k] <- beta[1]+beta[2]*breed[k]+
beta[3]*age[k]+b[id[k],1]+b[id[k],2]*age[k]
bw[k] ~ dnorm( mu[k], tau_e )
 }
#Random effects for each subject
for( j in 1:n){
b[j,1:2] ~ dmnorm(zero,tau_b)
}
 # Priors:
 #Prior on regression effects
for(p in 1:3){
beta[p]~dnorm(0,1.0E-06)
}

 #Prior on residual precision
sigma_e<-pow(var_e,0.5)
var_e<-pow(tau_e,-1)
tau_e~dgamma(0.001,0.001)
# Prior on random effect precision
sigma_b0 ~ dunif(0, 100)
sigma_b1 ~ dunif(0, 100)
rho ~ dunif(-1, 1)
sigma_b[1, 1] <- pow(sigma_b0, 2)
sigma_b[2, 2] <- pow(sigma_b1, 2)
sigma_b[1, 2] <- rho * sigma_b0 * sigma_b1
sigma_b[2, 1] <- sigma_b[1, 2]
tau_b[1:2, 1:2] <- inverse(sigma_b)
}
```

S.2 Illustrating Example

This function computes the marginal selection criteria (mDIC,mWAIC and mPSBF) and thier corresponding conditional counterparts.

We show simple example to illustrate how this function works.

```
##Install appropriate libraries###############
library("mvtnorm")
library("rjags")
library(readxl)
library("xlsx")
wd <- getwd()

##Call function that calculates the criteria
source("LMM_bayesselect.R")

###########Read data from excel or any other sources
chickdata <- read_excel("chickdata1.xls")
head(chickdata)
age<-(chickdata$age)
bw<-(chickdata$bw)/1000

file.show("JAGS_Model_LMM1.txt")

##########Data analysis using JAGS#############
# Any model can be analyse from JAGS or other software

jdata <- list(bw=bw, age=age,
              breed=c(chickdata$breed),
              id=c(chickdata$id),
              K=length(bw),
              n=max(chickdata$id))

jags.inits <- function(){list(beta=rnorm(3),
                      tau_e=rchisq(1,10),
                      sigma_b0=runif(1,0.5,1),
                         sigma_b1=runif(1,0.5,1),
                         rho=runif(1,0,1))}

results <-jags.model(file="JAGS_Model_LMM1.txt",
                        data=jdata,
                        inits=jags.inits,
                        n.chains = 3)
#JAGS_Model_LMM1.txt: can be changed depending on model we plan
    to fit
update(results, n.iter=10000,n.burn=5000,n.thin=10)
results_s<- coda.samples(results,
                           c("beta","sigma_b",
                             "sigma_e","var_e","mu"),
                           n.iter=10000, thin=10)

# Use MCMC output for the function to work
### Compute criteria ######
RA <-LMM_Bayesselect(results_s, resp=bw,
                        clust=chickdata$id,
                        X=cbind(1,chickdata$breed,age),
                        Z=cbind(1,age),betas=c("beta[1]",
                        "beta[2]","beta[3]"),
                        var_y="var_e",varZ="sigma_b")
# Show criteria ################
RA
```

S.3 R function for LMMs

```
LMM_Bayesselect <- function(mc_sam, resp, clust, X, Z, betas,
    var_y, varZ)
{
  resp <- as.matrix(resp)
  nobs <- nrow(resp)*ncol(resp)
  Z <- as.matrix(Z)
  if(nrow(Z)==1)
    Z <- as.matrix(rep(1,nobs*ncol(resp)))

  names_mcmc <- dimnames(mc_sam[[1]])[[2]]
  if(length(betas)>1)
  {
    sam_beta <- as.matrix(mc_sam[,betas])
  }else{
    sam_beta <- as.matrix(mc_sam[,grep(betas,names_mcmc)])
  }
   sam_mean <- sam_beta* t(X)
  if(ncol(resp)>1)
    clust <- rep(1:nrow(resp),ncol(resp))

  param_in1 <- names_mcmc[names_mcmc*in*var_y]
  if(length(param_in1)==0)
    print(paste("Please save in MCMC also",var_y))

  sam_vary <- as.matrix(mc_sam[,var_y])

  sam_meanb <- as.matrix(mc_sam[,grep("mu",names_mcmc)])
  param_in2 <- grep(varZ,names_mcmc)
  if(length(param_in2)==0)
    print(paste("Please save in MCMC also",varZ))

  if(ncol(Z)>1)
  {
    sam_varz <- as.matrix(mc_sam[,grep(paste0(varZ,"\\["),names_
        mcmc)])
  }else{
    sam_varz <- as.matrix(mc_sam[,varZ])
  }
  nK <- nrow(sam_beta) #Final length of MCMC with all chains

  clustl <- split(seq(nobs), clust) #cluster ID list

  nclus <- length(clustl
####intitialize vectors#######
#####Deviance over posterior means#########
  sam_varz_p <- colMeans(sam_varz)
  sam_vary_p <- mean(sam_vary)

  logp_yi<- rep(0,nK)
  logp_yib<- rep(0,nK)
  loopi <- function(i,sam_mean,sam_meanb,nK,clustl,Z,resp,sam_
      varz_p,sam_vary_p,sam_varz,sam_vary)
  { #Loop for observations
    mu_yi <- matrix(sam_mean[,unlist(clustl[i])], nrow=nK)
    mub_yi<-matrix(sam_meanb[,unlist(clustl[i])], nrow=nK)
    Z_i <- matrix(Z[unlist(clustl[i]),], ncol=ncol(Z))
    resp_i <- resp[unlist(clustl[i])]
    indav <- which(!is.na(resp_i))
###for posterior means#######
    mu_yi_p <- colMeans(sam_mean)[unlist(clustl[i])]
    mu_yi_pb<- colMeans(sam_meanb)[unlist(clustl[i])]

    marvar_p <- Z_i *matrix(sam_varz_p, ncol=ncol(Z))* t(Z_i) +
        diag(sam_vary_p, nrow(Z_i))
    logL_pm <- dmvnorm(resp_i[indav], mu_yi_p[indav], sigma=as.
        matrix(marvar_p[indav,indav]), log=TRUE)
  logL_pmb <- dmvnorm(resp_i[indav], mu_yi_pb[indav], sigma=as
      .matrix(diag(sam_vary_p, nrow(Z_i))[indav,indav]), log=
      TRUE)

 ## Log likelihood#####
like<-function(k,Z_i,sam_varz,Z,sam_vary,resp_i,mu_yi,mub_yi,
    indav){ #Loop for MCMC iterations
     marvar <- Z_i* matrix(sam_varz[k,], ncol=ncol(Z))*t(Z_i) +
          diag(sam_vary[k], nrow(Z_i))
    p_yi <-dmvnorm(resp_i[indav], mu_yi[k,indav], sigma=as.
        matrix(marvar[indav,indav]))
    p_yib<-dmvnorm(resp_i[indav], mub_yi[k,indav],sigma=as.
        matrix(diag(sam_vary[k], nrow(Z_i))[indav,indav]))
    c(p_yi,p_yib,log(p_yi),log(p_yib),1/p_yi,1/p_yib)
  }

  result <- sapply(c(seq(1,nK,1)),like,Z_i=Z_i,sam_varz=sam_
      varz,Z=Z,sam_vary=sam_vary,resp_i=resp_i,mu_yi=mu_yi,mub_
      yi=mub_yi,indav=indav)
  logp_yi <- apply(cbind(logp_yi,-2*result[3,]),1,sum,na.rm =
      TRUE)
  logp_yib <- apply(cbind(logp_yib,-2*result[4,]),1,sum,na.rm
      = TRUE)
  resultmean <- apply(result,1,mean) # vector with means of c(
      p_yi, p_yib, log(p_yi), log(p_yib), 1/p_yi, 1/p_yib)
  resultvar <- apply(result,1,var)# vector with variances of c
      (p_yi, p_yib, log(p_yi), log(p_yib), 1/p_yi, 1/p_yib)

  logCPOi <- log(1/resultmean[5])
  logCPObi <- log(1/resultmean[6])
  logLm <- resultmean[3]
  logLmb <- resultmean[4]
  log_mLm <- log(resultmean[1])
  log_mLmb <- log(resultmean[2])
  varp_yi <- resultvar[3] #variance log p_yi
  varp_yib <- resultvar[4] #variance log p_yib
  c(logLm,logLmb,log_mLm,log_mLmb,logL_pm,logL_pmb,varp_yi,
      varp_yib,logCPOi,logCPObi,logp_yi,logp_yib)
}

resloopi <- sapply(c(seq(1,nclus,1)),loopi,sam_mean=sam_mean,
    sam_meanb=sam_meanb,nK=nK,clustl=clustl,Z=Z,resp=resp,sam_
    varz_p=sam_varz_p,sam_vary_p=sam_vary_p,sam_varz=sam_varz,
    sam_vary=sam_vary)

sumres <- apply(resloopi,1,sum) # vector with means of c(logLm
    ,logLmb,log_mLm,log_mLmb,logL_pm,logL_pmb,varp_yi,varp_yib,
    logCPOi,logCPObi,logp_yi,logp_yib) where logp_yi and lop_
    yib are vectors of 1 X niteration size, i.e, probabilities
```

```
      for each sample of \theta.

   mid <- (length(sumres)-10)/2
     DICm1 <- Devm + pDm #Spiegelhalter
     pDmb <- Devmb - Dev_pmb
     DICmb1 <- Devmb + pDmb
     WAICm <- Dev_pmW+2*pDW
     WAICmb <- Dev_pmWb+2*pDWb
    PSBF=sumres[10] #sum logCPObi
   #conditional
   lppdb <- -2*sumres[10]
   cplppd <- lppdb-Dev_pmWb
   clppd <- lppdb+cplppd

   return(list(mDIC1=DICm1, mWAIC=WAICm,mPSBFn=-oPSBF,cDIC1=
      DICmb1,cWAIC=WAICmb,cPSBFn=-PSBF,mplppd= mplppd/2))

}
```

S.4 Likelihood for Skew-normal function

```
marvar_p <- Z_i *(matrix(sam_varz_p, ncol=ncol(Z))) * t(Z_i) +
    diag(sam_vary_p, nrow(Z_i))
       logL_pmb <- dmvnorm(resp_i[indav], mu_yi_pb[indav], sigma=
          as.matrix(diag(sam_vary_p, nrow(Z_i))[indav,indav]),
          log=TRUE)

       if(ncol(Z)==1){
         Lam_i<-1/(1/sam_varz_p+t(Z_i)*diag(sam_vary_p,nrow(Z_i))
           *Z_i)
         } else{Lam_i<-inv(inv(as.matrix(sam_varz_p, ncol=ncol(Z)
            ))+t(Z_i)*as.matrix(diag(sam_vary_p,nrow(Z_i)))*Z_i)}

        mu_1i<-Lam_i*t(Z_i)*inv(diag(sam_vary_p,nrow(Z_i)))
        A<-t(as.matrix(rep(sam_deltae_p, nrow(Z_i)),nrow=1))*inv(
            sqrtm(as.matrix(diag(sam_vary_p,nrow(Z_i)))))
        B<- rep(0,nrow(Z_i))
        mu_2i<- rbind(A,B)

       C<- t(as.matrix(rep(sam_deltae_p, nrow(Z_i)),nrow=1))*as.
          matrix(diag(sam_vary_p, nrow(Z_i)))*Z_i

       if(ncol(Z)==1){
        D<-t(matrix(rep(sam_deltab_p, ncol=ncol(Z)),nrow=1))/sam_
            varz_p^0.5
    } else{ D<-t(matrix(rep(sam_deltab_p, ncol=ncol(Z)),nrow
        =1))*inv(sqrtm(matrix(sam_varz_p, ncol=ncol(Z))))}

    LAM_i<-rbind(C,D)
    upper <- as.numeric((mu_2i-LAM_i* mu_1i)*(resp_i[indav]-mu
        _yi_p[indav]))
    sigma<-diag(2)*+LAM_i* Lam_i*t(LAM_i)

      cum<-pmvnorm(lower=rep(-Inf,length(upper)), upper=upper,
          mean=rep(0, length(upper)),sigma=sigma)
    logL_pm <- -0.5*(dmvnorm(resp_i[indav], mu_yi_p[indav],
        sigma=as.matrix(marvar_p[indav,indav]),log=TRUE))+ sum(
        log(cum))
```

Leuven Biostatistics and Statistical Bioinformatics (L-BioStat)
Kapucijnenvoer 35
Block D, bus 7001
3000 LEUVEN, BELGIË
tel. + 32 16 37 33 89
fax + 32 16 33 70 15
www.kuleuven.be