KATHOLIEKE UNIVERSITEIT
LEUVEN

Faculty of Business and Economics

Focused estimation and model averaging for high-dimensional data, an overview

Gerda Claeskens

DEPARTMENT OF DECISION SCIENCES AND INFORMATION MANAGEMENT (KBI)

KBI 1101

Focused estimation and model averaging for high-dimensional data, an overview

Gerda Claeskens
ORSTAT and Leuven Statistics Research Center
K.U.Leuven, Naamsestraat 69
3000 Leuven, Belgium
Gerda.Claeskens@econ.kuleuven.be

December 2010

Abstract

The quest for a good estimator of a certain focus or target is present regardless of the dimensionality of the data. Obtaining such a good estimator with low mean squared error, or a prediction with low prediction error often proceeds via a variable selection or model selection search. Estimators can also be averaged to enlarge the space of possible estimators in an attempt to further lower the mean squared error. While these methods are being studied mostly in situations with the number of variables much smaller than the sample size, this paper concentrates on the additional difficulties and challenges when applying these methods in a context of high-dimensional data, i.e. data with more parameters than observations. Items discussed in the paper include focused model selection with squared error loss and methods addressed towards best squared prediction loss.

Keywords: variable selection, focused information criterion, penalized estimation, lasso, model averaging, efficiency, prediction loss.

1 Introduction

In variable selection problems we can make a distinction between two types of research questions. The first type is where the main emphasis is on *identifying* important variables within a possibly large group of variables. Often this identification proceeds by estimating the effects of these variables and using some type of threshold procedure to decide on which variables to keep and which to omit. One example is to identify genes that are related to a certain characteristic or pathology. In such case one wants to select from a large set of genes those that ‘are expressed’, that have a nonzero effect.

Another group of research questions is more concerned with the quality of a particular *estimator* of a quantity of interest, we call this the focus. The estimated focus depends on the variables that are in the model on which the estimator is based. For these questions there is less emphasis on precisely which variables are used to construct the estimator, rather, of importance is that the estimator is accurate. As a particular example, let us consider a study that investigates the effects of a climate change. A large number of variables are measured (weather related variables such as temperature, precipitation, wind, humidity, measures at the oceans, on land, . . .). Several focus quantities can be phrased: what is the expected change in temperature on earth within ten years from now, within 50 years, within 100 years? What is the probability of a certain type of extreme weather situation (for example a large flood), which might be of interest to insurance companies. For all of these examples it is less important to know whether or not, say, a measure of crop moisture is used in the model to estimate the expected temperature change, of more importance is to get a good estimator of this expected temperature rise. The current paper deals with this type of model selection questions that relate to a focus and to the quality of its estimator.

In particular, we will address the question of selecting a model that gives us a good estimator of the focus (where ‘good’ will be defined later) in a setting with a large number of variables p that might exceed the sample size n.

Since a good estimator is the main goal, computing this estimator in several models and then taking a weighted average, might lead to an even better estimator. This is the concept of model averaging. By averaging estimators obtained in different models, we enlarge the space of possible estimators, which might be beneficial from an accuracy point of view.

In this paper we give an overview of the existing approaches for a more targeted or directed model search and we introduce some focused model selection methods and techniques for model averaging for high dimensional data. A large amount of the literature already deals with the identification of single variables, we only briefly touch upon that issue here and rather concentrate on the focus aspect. Hence, an overview of variable identification methods is beyond the scope of this paper. A recent overview on penalized estimation methods for variable selection is given by Fan and Lv (2010).

The purpose of this paper is to present an overview of related results, rather than precise mathematical statements for each specific case. While more research is ongoing within this area, a detailed study of other interesting questions, such as for example which type of penalty function yields the smallest risk for a focus estimator, or how to determine and to define optimality of weights for model averaging, is not treated here.

2 Simultaneous selection and estimation

For situations where the number of variables exceeds the sample size, traditional maximum likelihood methods fail and penalties or constraints on the parameters are introduced in order to find estimators of the unknowns. Currently, the most often used methods are the lasso (Tibshirani, 1996), the adaptive lasso (Zou, 2006), a smoothly clipped absolute deviation penalty (SCAD, Fan and Li, 2001) and the Dantzig selector (Candes and Tao, 2007). These methods have in common that the estimation problem with more unknowns than there are observations is dealt with by setting a number of parameters to zero and by using shrinkage methods to estimate the remaining parameters. We discuss each of them in turn.

2.1 Penalized estimation methods

Suppose we have independent data $(\widetilde{\boldsymbol{x}}_i, Y_i)$ representing a vector covariate $\widetilde{\boldsymbol{x}}_i$ and a response variable Y_i, for $i = 1, \ldots, n$, which we model by means of a density function $f(y_i \mid \widetilde{\boldsymbol{x}}_i, \boldsymbol{\beta}_n)$, where the unknown parameter vector $\boldsymbol{\beta}_n$ has a length that is of order $o(n^a)$ with potentially $a \geq 1$.

For simplicity of notation, in the remaining part of the paper we will continue to use likelihood models, although the terminology can be extended to other types of estimating equations. Fan et al. (2009) use a quasi-likelihood or a loss function such as the hinge loss or ℓ_1 loss instead of the log-likelihood as their criterion function, while Caner (2009) uses a generalized method of moments objective function with a lasso penalty. To accommodate the situation where the length of $\boldsymbol{\beta}_n$ is larger than the sample size n, a penalty function q_λ is added to the criterion function used for estimating $\boldsymbol{\beta}_n$,

$$\ell_n(\boldsymbol{\beta}_n) = \frac{1}{n} \sum_{i=1}^{n} \log f_i(y_i \mid \widetilde{\boldsymbol{x}}_i, \boldsymbol{\beta}_n) - q_\lambda(\boldsymbol{\beta}_n). \tag{1}$$

A simultaneous selection of the components in $\boldsymbol{\beta}_n$ and estimation of these components is achieved by using a penalty function which satisfies the ‘sparsity’ condition. This means that coefficients are not only shrunk but some of them are effectively set to equal zero. Penalty functions of the form $q_\lambda(\boldsymbol{\beta}_n) = \lambda \|\boldsymbol{\beta}_n\|_\alpha$ with $0 \leq \alpha \leq 1$ and with $\|\cdot\|_\alpha$ denoting the ℓ_α-norm, satisfy the sparsity constraint. Taking $\alpha = 1$ corresponds to the popular ‘lasso’ estimator (Tibshirani, 1996). Variations on this theme exist, for example adding an ℓ_2 penalty to the ℓ_1 penalty corresponds to the elastic net estimator (Zou and Hastie, 2005), weights may be included in the norm such as for the adaptive lasso (Zou, 2006), or variables may be grouped in order to include or exclude them as a group rather than one by one (Yuan and Lin, 2006), e.g. for indicator variables used to model categorical data. For an overview on several ℓ_1 penalized approaches and the lars-algorithm (Efron et al., 2004), see Hesterberg et al. (2008).

An at the outset different looking estimator is the Dantzig selector (Candes and Tao, 2007),

$$\widehat{\boldsymbol{\beta}} = \arg\min \|\boldsymbol{\beta}_n\|_1 \text{ under the constraint } \|\frac{1}{n}\boldsymbol{X}^t(\boldsymbol{Y} - \boldsymbol{X\beta_n})\|_\infty \leq \lambda$$

which shows similarities in theoretical and practical behaviour to the lasso, and has been shown to be equivalent in some situations (Efron et al., 2007; Bickel et al., 2009; James et al., 2009). A study of the Dantzig selector for Cox proportional hazard regression models is presented in Antoniadis et al. (2010).

The minimax concave penalty approach of Zhang (2010) tries to remedy some of the bias of the lasso estimator by using a different penalty. The method requires an additional tuning parameter ν for which holds that when $\nu \to \infty$, the penalty approaches the ℓ_1 penalty, while for $\nu \searrow 0$ the penalty approaches the ℓ_0 form. More precisely, the estimator is obtained by minimizing with respect to $\boldsymbol{\beta}_n$,

$$\frac{1}{2n}\|\boldsymbol{Y} - \boldsymbol{X\beta_n}\|^2 + \lambda \sum_{j=1}^{p} \int_0^{|\beta_{n,j}|} (1 - x/(\nu\lambda))_+ dx,$$

with p the length of $\boldsymbol{\beta}_n$.

The smoothly clipped absolute deviation (SCAD) penalty of Fan and Li (2001) is designed with the goal to achieve sparsity of the estimator in combination with continuity and a reduced bias. The SCAD penalty is defined through the derivative and also requires an additional tuning parameter $\nu > 2$,

$$q'_\lambda(t) = \lambda \left\{ I(t \leq \lambda) + \frac{(\nu\lambda - t)_+}{(\nu - 1)\lambda} I(t > \lambda) \right\}. \tag{2}$$

As a penalty one takes $\sum_{j=1}^{p} q_\lambda(|\beta_{n,j}|)$. This approach has been further studied and applied in various situations. For use in ultra-high dimensions, see Fan and Lv (2008).

2.2 Optimality in which sense?

Since the above penalized approaches combine selection with estimation of the unknown parameter vector $\boldsymbol{\beta}_n$, it is no surprise that it does not for each and every of a wide range of possible uses of the estimator $\widehat{\boldsymbol{\beta}}_n$ lead to optimality properties (in some prespecified sense) .

For the situation of the use of a lasso penalty, Meinshausen and Bühlmann (2006) show, while working with graphs, that the choice of the value of λ in the penalty that is used to achieve the minimal squared prediction error for a new independent observation, is not optimal in terms of variable selection consistency. This means that using the 'optimal' penalty in terms of prediction error leads to including redundant components of $\boldsymbol{\beta}_n$ in the model. Leng et al. (2006) and Zhao and Yu (2006) come to a similar conclusion for linear models.

Kim et al. (2008) start the other way around and search for conditions under which the SCAD estimator is consistent for $\boldsymbol{\beta}^*$, the true parameter value in a linear model. They compare the SCAD estimator to the 'best' estimator in terms of minimizing $\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}_n\|^2$ with respect to $\boldsymbol{\beta}_n$, for the given sample (this is called the oracle estimator in their terminology). Under some conditions it is obtained that the probability that the SCAD estimator is equal to this oracle estimator converges to one. However, when turning to prediction accuracy, they conclude that the SCAD estimator is inferior to the oracle estimator.

It is expected that similar statements hold for the other type of penalized estimators.

In the next section we construct an approach where estimation and model/variable selection is disentangled in order to select that estimator which has the minimal estimated mean squared error for one situation. The construction can be redone when searching for a model which leads to the best prediction at a certain focus point. It cannot be asked to come up with a single estimator that is omnibus and best for all purposes in all respects.

3 Separating model selection from estimation

3.1 Specifying the focus and the loss function

While the penalized estimation methods such as lasso (with its variations), elastic net, SCAD penalties, the Dantzig selector,... all advocate simultaneous selection and estimation of the regression coefficients, it is clear that this might not be optimal from the standpoint of estimating a specific functional of those regression coefficients in terms of a specified loss function. With focused model selection we separate the selection and estimation part with the goal of selecting that model for which the estimated loss of the estimator of the focus, a targeted function of the model coefficients, is the smallest of the considered models.

The quality of a parameter estimator, say μ, may be summarized via its mean squared error (MSE). Different estimators, in general, have different values for the mean squared error. Obviously, the value of the mean squared error depends on the model that is used to construct the estimator, since the model determines the bias and the variance. When irrelevant variables are included in a model, the variance of the estimator will, in general, increase. On the other hand, when important variables are left out of the model, the estimator might be biased. A good estimator has a small mean squared error. One version of the focused information criterion (FIC) estimates the mean squared error of the focus estimator in the different models under consideration. Since the criterion uses directly the estimator of the focus parameter, we really direct the criterion towards searching for the best model in mean squared error sense for this focus. Other versions of the FIC can be constructed for use with other loss functions such as the error rate for predictions of binary variables (Claeskens et al., 2006). The FIC has been introduced by Claeskens and Hjort (2003) (see also Claeskens and Hjort, 2008b, Ch. 6), and further studied in various situations, such as for Cox proportional regression models (Hjort and Claeskens, 2006), linear hazard models (Hjort, 2008), capture-recapture models (Bartolucci and Lupparelli, 2008), volatility forecasting (Brownlees and Gallo, 2008), optimal hedge ratios (Lien and Shrestha, 2005) and autoregressive time series (Claeskens et al., 2007). Minimizing an estimator of averaged mean squared error instead of at a specified focus point is studied by Claeskens and Hjort (2008a).

3.2 Penalized estimation under local misspecification

We study the risk properties of estimators in penalized likelihood-based models for further use in the construction of focused variable selection. Similar to the situation of models with a small number of variables relative to the sample size (Claeskens and Hjort, 2003), the starting assumption is that the true model is in a local neighborhood of a certain fixed model. More precisely, suppose we have independent data $(\widetilde{\boldsymbol{x}}_i, y_i) = (\boldsymbol{x}_i, \boldsymbol{z}_i, y_i)$, with $i = 1, \ldots, n$, where the covariates $\boldsymbol{x}_i$ are part of all models, while the covariates $\boldsymbol{z}_i$ are subject to a variable selection search. Under local misspecification, the true $p = (p_\theta + p_\gamma)$-dimensional parameter vector $\boldsymbol{\beta}_n = (\boldsymbol{\theta}_0, \boldsymbol{\gamma}_0 + n^{-1/2}\boldsymbol{\delta})$, where θ_0 is a vector of length p_θ consisting of the parameters that are common to all considered models. The last p_γ components of $\boldsymbol{\beta}_n$ are parameters which will be searched over. We define $\boldsymbol{\beta}_0 = (\boldsymbol{\theta}_0, \boldsymbol{\gamma}_0)$ the parameter vector of the minimal, or null model. The vector $\boldsymbol{\delta}$ determines the size of the neighborhood, with $\boldsymbol{\delta} = \boldsymbol{0}$ reducing the model to the minimal model.

Throughout this paper we work with a setting where $p_\theta < n$ is not depending on the sample size, while p_γ may be strictly larger than n. To accommodate the setting with $p_\gamma > n$, a penalty is introduced in the estimation of the parameters $\boldsymbol{\beta}_n$. An estimator $\widehat{\boldsymbol{\beta}}_n$ is obtained by maximizing the penalized log-likelihood objective function (again, other than likelihood functions could be used)

$$\ell_n(\boldsymbol{\theta}, \boldsymbol{\gamma}) = \frac{1}{n} \sum_{i=1}^{n} \log f_i(y_i \mid \widetilde{x}_i, \boldsymbol{\theta}, \boldsymbol{\gamma}) - q_\lambda(\boldsymbol{\gamma}). \tag{3}$$

Since the length of $\boldsymbol{\theta}$ is always strictly smaller than n, we do not penalize these parameters in the estimation procedure. Similarly, with lasso-estimation the intercept parameter is usually left unspecified. An interesting extension is to investigate the case when p_θ is also allowed to grow with n, which will require a penalty on these coefficients $\boldsymbol{\theta}$ as well.

Rather than working with an ℓ_1 penalty, as in the lasso approach, for which the mean squared error has no available explicit expression as yet, we here suggest using $\psi(x)$ as an approximation to $|x|$, with for ε a small number such as 10^{-10},

$$\psi(x) = (x^2 + \varepsilon)^{1/2} \text{ and } q_\lambda(\boldsymbol{\gamma}) = \frac{\lambda}{n} \sum_{j=1}^{p_\gamma} \psi(\gamma_j - \gamma_{j0}). \tag{4}$$

The use of this approximation has two advantages: (i) the criterion function in (3) is differentiable and (ii) the expressions for the squared bias and variance of the resulting estimators can be unbiasedly estimated, which is not the case when using an absolute value.

The majority of the current research on lasso-type estimators works under the assumption of a fixed true model. One exception is Knight and Fu (2000) who include local asymptotical results in their study. The results obtained so far do not write the asymptotic distribution in a form that can be used for minimizing the risk since it is only proven that the estimator converges in distribution to the minimizer of another objective function, with no available explicit expressions for the bias and variance of the estimator under local misspecification. Similarly for other type of penalized estimation methods, bounds on the mean squared error (MSE) are known, though not the MSE itself.

3.3 Mean squared error of the focus estimator

Let $(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\gamma}})$ denote the maximizers of (3) with respect to $(\boldsymbol{\theta}, \boldsymbol{\gamma})$ when using (4) as a penalty. The focus is denoted by $\mu(\boldsymbol{\theta}, \boldsymbol{\gamma})$. The variable selection will proceed by estimating the mean squared error of $\widehat{\mu} = \mu(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\gamma}})$ in various models, specified by leaving out some or all of the components of the vector $\boldsymbol{\gamma}$, and by selecting that model for which the estimated MSE is the smallest. The true value of the focus is denoted as $\mu_{\text{true}} = \mu(\boldsymbol{\theta}_0, \boldsymbol{\gamma}_0 + n^{-1/2}\boldsymbol{\delta})$. Let S denote any subset of $\{1, \ldots, p_\gamma\}$. The estimator of the focus in the model containing the common variables $\boldsymbol{\theta}$ and the optional variables in subset S is defined as $\widehat{\mu}_S = \mu(\widehat{\boldsymbol{\theta}}_S, \widehat{\boldsymbol{\gamma}}_S, \boldsymbol{\gamma}_{0,S^c})$ where $\boldsymbol{\gamma}_{0,S^c}$ consists of those $\gamma_{0,j}$ for which $j \notin S$, and $\widehat{\boldsymbol{\gamma}}_S$ is a vector of length $|S|$, the cardinality of S. Note that also the estimator $\widehat{\boldsymbol{\theta}}$ receives a subscript S to indicate that its value might depend on the specific set S, the length of $\boldsymbol{\theta}_S$ is always fixed to p_θ and does not change across the different models. Mathematically, taking a subset of size $|S|$ of a vector v of length p_γ is denoted by premultiplying this vector by a projection matrix π_S with dimension $|S| \times p_\gamma$, to result in $\pi_S \boldsymbol{v} = v_S$, a vector consisting of those v_j for $j \in S$.

We assume that the penalty constant satisfies $\lambda_n/\sqrt{n} \to \lambda_0 > 0$ and that the Fisher information matrix at the full model, evaluated at $\boldsymbol{\beta}_0$, is nonsingular. This matrix is defined as

$$\boldsymbol{J} = \text{Var}\big(\tfrac{\partial}{\partial \boldsymbol{\beta}} \ell_n(\boldsymbol{\beta}_0)\big) = \begin{pmatrix} \boldsymbol{J}_{00} & \boldsymbol{J}_{01} \\ \boldsymbol{J}_{10} & \boldsymbol{J}_{11} \end{pmatrix} \text{ with inverse } \boldsymbol{J}^{-1} = \begin{pmatrix} \boldsymbol{J}^{00} & \boldsymbol{J}^{01} \\ \boldsymbol{J}^{10} & \boldsymbol{J}^{11} \end{pmatrix},$$

where $\boldsymbol{J}^{11} = (\boldsymbol{J}_{11} - \boldsymbol{J}_{10}\boldsymbol{J}_{00}^{-1}\boldsymbol{J}_{01})^{-1}$. We denote by $\boldsymbol{J}_S$ the Fisher information matrix for the submodel S. The proof of the following theorem is placed in the appendix.

Theorem 1. *Let* $f_{\text{true}} = f(Y_i|\widetilde{\boldsymbol{x}}_i; \boldsymbol{\theta}, \boldsymbol{\gamma}_0 + \boldsymbol{\delta}/\sqrt{n})$, f *is two times continuously differentiable in a neighbourhood of* $(\boldsymbol{\theta}_0, \boldsymbol{\gamma}_0)$, *the matrix* $\boldsymbol{J}$ *is nonsingular and* $\frac{\lambda_n}{\sqrt{n}} \to \lambda_0 > 0$ *as* $n \to \infty$. *Define* $\boldsymbol{c}^t = \lambda_0 \psi'(\boldsymbol{\delta})^t = \lambda_0 \left(\frac{\delta_1}{\sqrt{\delta_1^2 + \epsilon}}, \ldots, \frac{\delta_q}{\sqrt{\delta_q^2 + \epsilon}} \right)$. *Then it holds that there is*
(i) convergence of the score vector

$$\begin{pmatrix} \sqrt{n}\frac{\partial}{\partial \boldsymbol{\theta}}\ell_n(\boldsymbol{\beta}_n) \\ \sqrt{n}\frac{\partial}{\partial \boldsymbol{\gamma}}\ell_n(\boldsymbol{\beta}_n) \end{pmatrix} \to_d \begin{pmatrix} \boldsymbol{U} \\ \boldsymbol{V} \end{pmatrix} + \begin{pmatrix} \boldsymbol{J}_{01}\boldsymbol{\delta} \\ \boldsymbol{J}_{11}\boldsymbol{\delta} + \boldsymbol{c} \end{pmatrix},$$

where $(\boldsymbol{U}^t, \boldsymbol{V}^t)^t \sim N_p(\boldsymbol{0}, \boldsymbol{J})$;

(ii) convergence of the parameter estimators

$$\begin{pmatrix} \sqrt{n}(\widehat{\boldsymbol{\theta}}_S - \boldsymbol{\theta}_0) \\ \sqrt{n}(\widehat{\boldsymbol{\gamma}}_S - \boldsymbol{\gamma}_{0,S}) \end{pmatrix} \to_d \begin{pmatrix} \boldsymbol{B}_S \\ \boldsymbol{C}_S \end{pmatrix} = \boldsymbol{J}_S^{-1} \begin{pmatrix} \boldsymbol{J}_{01}\boldsymbol{\delta} + \boldsymbol{U} \\ \boldsymbol{J}_{11,S}\boldsymbol{\delta} + \boldsymbol{V}_S + \boldsymbol{c}_S \end{pmatrix}$$

$$\sim \; N_{p_\theta + |S|}\left(\boldsymbol{J}_S^{-1} \left\{ \begin{pmatrix} \boldsymbol{J}_{01} \\ \boldsymbol{J}_{11,S} \end{pmatrix} \boldsymbol{\delta} \right\} + \begin{pmatrix} \boldsymbol{0}_p \\ \boldsymbol{c}_S \end{pmatrix}, \boldsymbol{J}_S^{-1} \right);$$

(iii) convergence of the focus estimator. For $\widehat{\mu}_S = \mu(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\gamma}_S})$, $\mu_{\text{true}} = \mu(\boldsymbol{\theta}_0, \boldsymbol{\gamma}_0 + \boldsymbol{\delta}/\sqrt{n})$, μ *continuously differentiable with respect to* $(\boldsymbol{\theta}, \boldsymbol{\gamma})$ *in a neighborhood of* $(\boldsymbol{\theta}_0, \boldsymbol{\gamma}_0)$, *and with* $\boldsymbol{\omega} = \boldsymbol{J}_{10}\boldsymbol{J}_{00}^{-1}\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\theta}} - \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\gamma}}$,

$$\begin{aligned} \sqrt{n}(\widehat{\boldsymbol{\mu}}_S - \mu_{\text{true}}) \quad \to_d & \quad \Lambda_S = (\tfrac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\theta}})^t \boldsymbol{B}_S + (\tfrac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\gamma}_S})^t \boldsymbol{C}_S - (\tfrac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\gamma}})^t \boldsymbol{\delta} \\ = & \quad (\tfrac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\theta}})^t \boldsymbol{J}_{00}^{-1}\boldsymbol{U} + \boldsymbol{\omega}^t(\boldsymbol{\delta} - \boldsymbol{G}_S\boldsymbol{D} - \boldsymbol{J}^{11,S,0}\boldsymbol{c}), \end{aligned}$$

where $\boldsymbol{G}_S = \boldsymbol{J}^{11,S,0}(\boldsymbol{J}^{11})^{-1}$, $\boldsymbol{J}^{11,S,0} = \pi_S^t \boldsymbol{J}^{11,S}\pi_S$, $\boldsymbol{D} \sim N_q(\boldsymbol{\delta}, \boldsymbol{J}^{11})$ *and all partial derivatives are evaluated at the null model* $(\boldsymbol{\theta}_0, \boldsymbol{\gamma}_0)$.

The limiting distributions are characterized by two bias components, a first one arising because of the local misspecification framework when $\boldsymbol{\delta}$ is non-zero, and a second bias component due to the penalization during the estimation (with a nonzero $\boldsymbol{c}$ under the assumed condition on λ_n). Under the stronger assumption that $\lambda_n/\sqrt{n} \to 0$, it follows that $\boldsymbol{c} = \boldsymbol{0}$ and the same asymptotic distribution results as for non-penalized estimators.

With a non-random model S the limiting distribution Λ_S is normal $\Lambda_S \sim N\{E(\Lambda_S),\ \mathrm{Var}(\Lambda_S)\}$, with mean and variance

$$\begin{aligned} E(\Lambda_S) &= \boldsymbol{\omega}^t\{(\boldsymbol{I}_q - \boldsymbol{G}_S)\boldsymbol{\delta} - \boldsymbol{J}^{11,S,0}\boldsymbol{c}\} \\ \mathrm{Var}(\Lambda_S) &= (\tfrac{\partial\boldsymbol{\mu}}{\partial\boldsymbol{\theta}})^t\boldsymbol{J}_{00}^{-1}\tfrac{\partial\boldsymbol{\mu}}{\partial\boldsymbol{\theta}} + \boldsymbol{\omega}^t\boldsymbol{J}^{11,S,0}\boldsymbol{\omega} = \tau_0^2 + \boldsymbol{\omega}^t\boldsymbol{J}^{11,S,0}\boldsymbol{\omega}. \end{aligned} \tag{5}$$

A variance–bias tradeoff is clearly visible. Indeed, for larger models ($|S|$ large) the variance of $\widehat{\mu}_S$ will in general be larger than for models with less parameters, where the bias due to model misspecification will be larger.

Adding the squared bias and the variance of Λ_S gives that the mean squared error of $\widehat{\mu}_S$ at model S is given by

$$\mathrm{mse}(S) = \tau_0^2 + \boldsymbol{\omega}^t\boldsymbol{J}^{11,S,0}\boldsymbol{\omega} + \boldsymbol{\omega}^t\{(\boldsymbol{I}_q - \boldsymbol{G}_S)\boldsymbol{\delta} - \boldsymbol{J}^{11,S,0}\boldsymbol{c}\}\{(\boldsymbol{I}_q - \boldsymbol{G}_S)\boldsymbol{\delta} - \boldsymbol{J}^{11,S,0}\boldsymbol{c}\}^t\boldsymbol{\omega}. \tag{6}$$

3.4 Estimation of the mean squared error

The idea of focused model selection (FIC) is to estimate mse(S) in (6) for each of the considered models S and to choose that model which gives the smallest estimated mean squared error $\widehat{\mathrm{mse}}(S)$. Thus, FIC($S$) = $\widehat{\mathrm{mse}}(S)$. We select that estimator $\widehat{\mu}_S$ for which the estimated risk (mean squared error) is the smallest.

The approximation of $|x|$ by $(x^2 + \epsilon)^{1/2}$ in the penalty has not only as an advantage that the criterion function is differentiable, but also that the expressions for the squared bias and the variance can be unbiasedly estimated. Plugging in empirical matrices for population Fisher information matrices, and inserting parameter estimates (e.g. at the biggest model) for unknown parameters, leaves us with the estimation of $\boldsymbol{\delta}\boldsymbol{\delta}^t$ in the squared bias component, the third term in (6). Since $\widehat{\boldsymbol{\delta}} = \sqrt{n}(\widehat{\boldsymbol{\gamma}}_{\mathrm{full}} - \boldsymbol{\gamma}_0) \to D \sim \mathrm{N}_{p_\gamma}(\boldsymbol{\delta}, \boldsymbol{J}^{11})$, we use as an estimator $\widehat{\boldsymbol{\delta}}\widehat{\boldsymbol{\delta}}^t - \widehat{\boldsymbol{J}}^{11}$ when this results in a positive value for the estimated squared bias, otherwise the squared bias is estimated by zero.

We rewrite $\boldsymbol{c} = \boldsymbol{\delta} \odot \boldsymbol{a}$, with $\odot$ denoting the componentwise multiplication. The bias estimation then results in using $\widetilde{\boldsymbol{\delta}}^2 = \max\{0, \mathrm{diag}(\widehat{\boldsymbol{\delta}}\widehat{\boldsymbol{\delta}}^t - \widehat{\boldsymbol{J}}^{11})\}$ as an estimator for $(\delta_1^2, \ldots, \delta_{p_\gamma}^2)$ in the denominator of the vector $\boldsymbol{a}$, leading to the estimator $\widehat{\boldsymbol{a}}$.

The squared bias and variance terms are estimated by (with $\otimes$ denoting the Kronecker product)

$$\begin{aligned} \text{FIC.bias.sq} &= \max\{0, \widehat{\boldsymbol{\omega}}^t \cdot \textbf{bias-comp} \cdot (\widehat{\boldsymbol{\delta}}\widehat{\boldsymbol{\delta}}^t - \widehat{\boldsymbol{J}}^{11}) \cdot \textbf{bias-comp}^t \cdot \widehat{\boldsymbol{\omega}}\} \\ \textbf{bias-comp} &= \{(\boldsymbol{I}_q - \widehat{\boldsymbol{J}}^{11,S,0}(\widehat{\boldsymbol{J}}^{11})^{-1}) - \widehat{\boldsymbol{J}}^{11,S,0} \odot (\mathbf{1}_{p_\gamma} \otimes \widehat{\boldsymbol{a}})\} \\ \text{FIC.var} &= \widehat{\tau}_0^2 + \widehat{\boldsymbol{\omega}}^t \widehat{\boldsymbol{J}}^{11,S,0} \widehat{\boldsymbol{\omega}}. \end{aligned}$$

This results in

$$FIC(S;\lambda) = \widehat{\mathrm{mse}(S;\lambda)} = \text{FIC.var} + \text{FIC.bias.sq}. \qquad (7)$$

The subset S for which the FIC value in (7) is the smallest gives the best index set. By not leaving out constants that are the same for each model, this expression for FIC keeps its interpretation as an estimate of the mean squared error.

To deal with the construction and inversion of a high-dimensional information matrix $\boldsymbol{J}$, we use a Nyström approximation to $\boldsymbol{J}$ (for details, see Belabbas and Wolfe, 2007). This results in a symmetric positive semi-definite matrix of dimension $p_\gamma \times p_\gamma$ for which we find the best rank k-approximation, with $k = \min(n, |S|)$.

The selection of the tuning parameter λ proceeds by minimizing the estimated mean squared error. Thus the selected model S and value λ are chosen to reach

$$\min_\lambda \min_S \text{FIC}(S;\lambda).$$

In practice, a grid search may be performed to find the best λ.

4 Simulation study

In a limited simulation study we compare variable selection by FIC to the simultaneous estimation and selection procedures of the lasso and SCAD, see Section 2.1. For simplicity we assume a linear model $Y_i = \boldsymbol{X}_i\boldsymbol{\beta}_{\text{true}} + \varepsilon_i$ with $\varepsilon_i \sim N(0,1)$. Common to all models are the unknown intercept and error variance (thus $p_\theta = 2$). The data are generated, first with a sample size $n = 25$. With (i) $p_\gamma = 20$, this gives a situation with a large number of parameters though not exceeding the sample size, (ii) $p_\gamma = 100$, which creates a setting where the number of parameters largely exceeds the sample size. For sample size $n = 100$ we create an easier setting with $p_\gamma = 20$.

The true model is generated according to four settings.

Setting 1: None of the generated coefficients is zero and all regression variables are independent. $\boldsymbol{\beta}_{\text{true}} = (1, -1/2, 1/3, -1/4, 1/5, \ldots, \pm 1/p_\gamma)/\sqrt{n}$, for $i = 1, \ldots, n$: $\boldsymbol{X}_i \sim N_{p_\gamma}(\mathbf{0}_{p_\gamma}, \boldsymbol{I}_{p_\gamma})$.

Setting 2: None of the generated coefficients is zero, as in setting 1, though the regression variables are dependent with a covariance matrix $\boldsymbol{\Sigma}$, where $\sigma_{jj} = 1$ and $\sigma_{jk} = 0.5$ when $j \neq k$, for $i = 1, \ldots, n$: $\boldsymbol{X}_i \sim N_{p_\gamma}(\mathbf{0}_{p_\gamma}, \boldsymbol{\Sigma})$

Setting 3: Except for the first 5 coefficients, the remaining coefficients are zero and all regression variables are independent, $\boldsymbol{\beta}_{\text{true}} = (1, -1, 1, -1, 1, 0, \ldots, 0)/\sqrt{n}$, for $i = 1, \ldots, n$: $\boldsymbol{X}_i \sim N_{p_\gamma}(\mathbf{0}_{p_\gamma}, \boldsymbol{I}_{p_\gamma})$.

Setting 4: The coefficients are taken as in setting 3, the regression variables are dependent as in setting 2.

As focus points for which we wish to obtain an estimator with small mean squared error, we take $\mu_j(\boldsymbol{\beta}) = \beta_0 + \boldsymbol{x}_{0j}\boldsymbol{\beta}$, where $\boldsymbol{x}_{01}$ consists of p_γ randomly chosen values in the interval $[-1, 1]$, $\boldsymbol{x}_{02}$ takes the first three components of $\boldsymbol{x}_{01}$ and as last components it takes randomly generated values within the interval $[4, 8]$, while $\boldsymbol{x}_{03}$ has as first three components randomly generated values in the interval $[4, 8]$ and its last $(p_\gamma - 3)$ components correspond to those of $\boldsymbol{x}_{01}$, thus, randomly generated from the interval $[-1, 1]$.

Three methods are used in the comparison (i) FIC with penalty (4) where λ has been chosen to minimize FIC, (ii) lasso where 10-fold cross-validation is applied to determine λ, the estimated coefficients are found via

$$\min_{\beta} \sum_{i=1}^{n} (Y_i - \beta_0 - \boldsymbol{X}_i\boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{p_\gamma} |\beta_j|,$$

Table 1: Averaged squared error of the estimators for three different focus parameters over 1000 simulated data sets using FIC, lasso and the SCAD penalty. The sample size is equal to $n = 25$, the number of variables $p_\gamma = 20$.

Focus:		μ_1	μ_2	μ_3
		$x_{01k} \in [-1, 1]$	$x_{02k(k>3)} \in [4, 8]$	$x_{03k(k\leq 3)} \in [4, 8]$
Setting 1	FIC	0.180	0.084	1.802
	Lasso	0.105	4.353	2.762
	SCAD	0.510	8.351	2.571
Setting 2	FIC	0.222	0.072	1.846
	Lasso	0.108	2.348	2.875
	SCAD	0.528	4.651	3.484
Setting 3	FIC	0.222	0.308	2.231
	Lasso	0.131	6.740	3.803
	SCAD	0.818	12.098	3.747
Setting 4	FIC	0.082	0.126	1.200
	Lasso	0.078	0.323	1.137
	SCAD	0.799	5.225	3.872

and (iii) the SCAD approach with the penalty defined by (2) and 10-fold cross-validation to determine the value of λ. For the lasso, the `R`-library `glmnet` has been used, the SCAD values have been computed via the `R`-library `SIS`.

Tables 1–3 give the averaged squared errors of the estimators $\mu_j(\widehat{\beta}_0+\boldsymbol{x}_{0j}^t\widehat{\boldsymbol{\beta}})$ for $j = 1, 2, 3$ over 1000 simulation runs, computed via the estimates that are obtained by the three methods. Table 1 presents the averaged squared errors of the estimates of the focus parameters over the 1000 simulation runs for the case where $n = 25$ and $p_\gamma = 20$. For all four settings, the best results for FIC are obtained for the second and third focus (μ_2, μ_3) where some of the covariate values are large (within the interval $[4, 8]$). For these two focus points, for all

Table 2: Averaged squared error of the estimators for three different focus parameters over 1000 simulated data sets using FIC, lasso and the SCAD penalty. The sample size is equal to $n = 25$, the number of variables $p_\gamma = 100$.

Focus:		μ_1	μ_2	μ_3
		$x_{01k} \in [-1, 1]$	$x_{02k(k>3)} \in [4, 8]$	$x_{03k(k\leq3)} \in [4, 8]$
Setting 1	FIC	0.083	0.243	0.242
	Lasso	0.113	4.500	0.522
	SCAD	0.235	5.350	0.557
Setting 2	FIC	0.071	0.217	0.267
	Lasso	0.124	2.422	0.637
	SCAD	0.245	2.514	0.542
Setting 3	FIC	0.094	0.344	0.177
	Lasso	0.133	7.212	0.680
	SCAD	0.280	7.883	0.806
Setting 4	FIC	0.088	0.343	0.178
	Lasso	0.149	2.999	0.816
	SCAD	0.298	2.957	0.782

settings, the lasso method ranks second and SCAD third, concerning the averaged squared error. For the first focus μ_1 for which all x_{j0} values are small, within the interval $[-1, 1]$, lasso and FIC yield comparable numbers, while those of SCAD are again larger.

The results of the more interesting setting where the number of variables $p_\gamma = 100 > n = 25$ are summarized in Table 2. For all settings and all three focus parameters, the FIC yields the estimator with the smallest mean squared error, with an obvious difference for μ_2, for all settings. Both for μ_1 and μ_2, lasso gives slightly better MSE values than SCAD, though the results are comparable, and SCAD performs slightly better than lasso in the case of dependent covariates, see setting 2 (μ_3) and setting 4 (μ_2, μ_3).

Table 3: Averaged squared error of the estimators for three different focus parameters over 1000 simulated data sets using FIC, lasso and the SCAD penalty. The sample size is equal to $n = 100$, the number of variables $p_\gamma = 20$.

Focus:		μ_1	μ_2	μ_3
		$x_{01k} \in [-1, 1]$	$x_{02k(k>3)} \in [4, 8]$	$x_{03k(k\leq3)} \in [4, 8]$
Setting 1	FIC	0.030	0.013	0.611
	Lasso	0.030	0.346	0.635
	SCAD	0.131	0.771	0.399
Setting 2	FIC	0.031	0.013	0.610
	Lasso	0.032	0.200	0.625
	SCAD	0.135	0.588	0.434
Setting 3	FIC	0.077	0.125	1.195
	Lasso	0.068	0.538	1.132
	SCAD	0.273	1.277	0.667
Setting 4	FIC	0.082	0.126	1.200
	Lasso	0.078	0.323	1.137
	SCAD	0.291	0.814	0.680

An easier setting from the point of view of estimation is when $n = 100$ and $p_\gamma = 20$, a situation where penalization is not strictly needed (see Table 3). For focus μ_1, FIC gives about the same performance in terms of mean squared error than the lasso, while FIC performs better in the case of μ_2. In this setting, for μ_3, the SCAD now gives the best results, while FIC and lasso again have a similar behavior.

5 Looking for a best prediction

5.1 Minimum risk methods

Another method that aims at unbiased risk estimators is Stein unbiased risk estimation (SURE) where the expected squared prediction error plays the central role. This is in common use in wavelet estimation where the number of parameters (wavelet coefficients) equals the sample size. Donoho and Johnstone (1995) determine the threshold parameter for wavelet estimation via SURE. A study of SURE, generalized cross-validation, Mallow's C_p and AIC in a setting of sparseness, where it is known that only a few coefficients will be non-zero, is performed by Jansen (2010). One of the conclusions of that paper is that a combination of adaptive minimum risk methods with bias-free hard threshold selection outperforms methods with minimax risk properties, such as the false discovery rate (Benjamini and Hochberg, 1995) in practical situations. This research links the 'classical' criteria that are usually only investigated with p much smaller than n to settings with sparsity, which are common in high-dimensional situations.

5.2 The concept of persistency

When proving that a method is best in terms of predictive properties, persistency can be shown. Greenshtein and Ritov (2004) define persistency of an estimator in a random design linear regression model. The data are denoted $(Y_i, X_{1i}, \ldots, X_{pi})$, $i = 1, \ldots, n$ with distribution $F_n \in \mathcal{F}^n$, a set of distribution functions for n i.i.d random vectors of length $p+1$. The number of variables $p = n^a$ with $a > 1$. The quality of the estimator for $\boldsymbol{\beta}$ is measured by the expected squared prediction error:

$$L_{F_n}(\boldsymbol{\beta}) = E_{F_n}[(Y - \sum_{j=1}^{p} \beta_j X_j)^2],$$

where the expectation is with respect to the true finite sample distribution of the random data vector. Denoting $\boldsymbol{\beta}^*$ the value of the parameter vector $\boldsymbol{\beta}$ where L_{F_n} obtains its minimum, an estimator $\widehat{\boldsymbol{\beta}}_n$ is called persistent when $L_{F_n}(\widehat{\boldsymbol{\beta}}_n) - L_{F_n}(\boldsymbol{\beta}^*) \xrightarrow{P} 0$, $n \to \infty$.

Greenshtein (2006) extends the setting towards general nonnegative prediction loss functions instead of squared error and towards predictors that are not necessarily linear functions of the random covariates. Under certain conditions it has been shown that the lasso estimator is persistent when taking the expected squared prediction error (Greenshtein and Ritov, 2004; Greenshtein, 2006; Bartlett et al., 2009). The prediction error of the lasso estimator is also studied by van der Geer (2008). The persistency property should be interpreted as giving results on the squared prediction error, on average. It does not say anything about the quality of the prediction for an individual point.

If we specify the loss function to use within the persistency for a specific focus, such as estimation or prediction at a certain point $\mu(\boldsymbol{\beta}) = \boldsymbol{x}_0^t\boldsymbol{\beta}$, we take $L_F(\boldsymbol{\beta}) = (\boldsymbol{x}_0^t\boldsymbol{\beta} - \boldsymbol{x}_0^t\boldsymbol{\beta}_{\text{true}})^2$. Since the design point where the focus is evaluated is fixed, this is a non-random quantity. The optimal value $\boldsymbol{\beta}^*$ that minimizes this risk is equal to $\boldsymbol{\beta}_{\text{true}}$. Requiring for this example that $L_{F_n}(\widehat{\boldsymbol{\beta}}_n) - L_{F_n}(\boldsymbol{\beta}^*) \xrightarrow{P} 0$ turns out equivalent to asking for consistency of the estimator.

5.3 Efficiency versus persistency

Efficiency concerns the ratio of the squared prediction error at the selected model with that of the optimal model for which the squared prediction error is minimized. With this concept, the model selection or variable selection process itself is investigated, rather than the estimated coefficients. Thus, while in the persistency criterion the estimator $\widehat{\boldsymbol{\beta}}$ is the random variable determining the convergence (or not) in probability, for efficiency it is the randomness in the selected model $\widehat{S}$ that determines the convergence (or not). For efficiency we look at

$$R_n = \frac{\sum_{i=1}^n E[(\sum_{j\in\widehat{S}} \widehat{\beta}_j x_j - Y_{\text{true},i})^2 | \widehat{S}]}{\sum_{i=1}^n E[(\sum_{j\in S^*} \widehat{\beta}_j x_j - Y_{\text{true},i})^2]},$$

where the set of variables S^* is such that it gives the smallest possible denominator. A variable selection mechanism is called efficient if R_n converges in probability to one when the sample size grows to infinity. Note that the numerator is random in the choice of the set $\widehat{S}$, which is determined by the data. There are several differences with the study of persistency. (i) A ratio of risk values is studied instead of a difference. Since risk values are often related to scales, the ratio might have interpretational advantages. (ii) The covariates are taken as fixed, rather than random. (iii) The randomness of the estimator $\widehat{\boldsymbol{\beta}}$ is accounted for with the calculation of the expected value, the randomness in R_n is determined by the randomness in the selection of the set $\widehat{S}$. For persistency one does not perform a separate selection step, there only is estimation where it may happen that certain components are set to zero (e.g. when using an ℓ_1-penalty). (iv) Efficiency has so far only been investigated for regression models where the number of non-zero coefficients is of the order $o(n^a)$ with $0 < a \leq 1$. Persistency has only been investigated for models with a large number of variables $(a > 1)$ where selection and estimation take place with a single action.

For models with a large number of variables it could be of interest to investigate the behaviour of estimation and selection methods conditional on the design, for application to the given design points or to a new value where prediction should take place. Efficiency results for the case where $a > 1$ would be useful.

5.4 Transductive learning

The search for good predictions at *specific* points rather than trying to predict a complete function in all of its possible domain points is referred to as transductive learning in the machine learning literature (see e.g. Vapnik, 1995). For a discussion between semisupervised learning and transduction, see chapter 25 of Chapelle et al. (2006). The idea behind transduction is to solve an as simple as possible problem, if you only need a prediction at one point, do not first construct an estimator for prediction at the whole covariate space, but rather do it only for the single point of interest. In this regard, there is a similarity to the focused selection of Section 3.

While a large part of the literature is involved with classification problems, transduction regression is concerned with the prediction of response values at certain given covariate vectors. Cortes et al. (2008) investigate the stability of several transduction algorithms. Let the prediction of the response Y at a covariate vector $\boldsymbol{x}$ be denoted by $\widehat{Y}_x = \widehat{h}(\boldsymbol{x})$. Denote by $\boldsymbol{x}_{0j}$, $j = 1, \ldots, n_0$ the new covariate vectors at which predictions of the response are to be obtained. The local transductive regression algorithms minimize an objective function of the form

$$\|\widehat{h}(\boldsymbol{x})\|_K^2 + \lambda_1 \frac{1}{n} \sum_{i=1}^{n} \{\widehat{h}(\boldsymbol{x}_i) - Y_i\}^2 + \lambda_2 \frac{1}{n_0} \sum_{j=1}^{n_0} \{\widehat{h}(\boldsymbol{x}_{0j}) - \widetilde{Y}_{0j}\}^2,$$

where $\widetilde{Y}_{0j}$ $(j = 1, \ldots, n_0)$ are called pseudo-target values, since the true response values are unknown. The local transductive algorithms use local weighted regression to determine the $\widetilde{Y}_{0j}$. The norm $\| \cdot \|_K$ is defined by a kernel function K in a kernel reproducing Hilbert space. We could interpret this term as being a cost (or a penalty) for the complexity of the prediction function $\widehat{h}$. This method involves two further tuning parameters λ_1 and λ_2.
Another type of algorithms are called unconstrained regularization methods, which search for the vector h with length $n + n_0$ to minimize the following objective function

$$\boldsymbol{h}^t \boldsymbol{Q} \boldsymbol{h} + (\boldsymbol{h} - \widetilde{\boldsymbol{Y}})^t \boldsymbol{C} (\boldsymbol{h} - \widetilde{\boldsymbol{Y}}),$$

with explicit solution $\boldsymbol{h} = (\boldsymbol{C}^{-1}\boldsymbol{Q} + \boldsymbol{I}_{n+n_0})^{-1}\widetilde{\boldsymbol{Y}}$. The matrix $\boldsymbol{Q}$ is a symmetric regularization matrix, while $\boldsymbol{C}$ contains empirical weights, both matrices have dimension $(n+n_0) \times (n+n_0)$. The vector $\widetilde{\boldsymbol{Y}}$ consists of the observed response values $(Y_1, \ldots, Y_n)$ together with the pseudo-target values $(Y_{01}, \ldots, Y_{0n_0})$. Several algorithms can be written in this form, for an overview and discussion, see Cortes et al. (2008).
The method of transduction is not immediately guided towards risk minimization starting from the risk of the used estimator (or predictor). The statistical properties regarding e.g. minimum prediction risk, efficiency or consistency largely remain to be investigated.

6 Model averaging with large numbers of variables

When an estimator with a small mean squared errors is sought amongst a collection of models in which these estimators are computed, one can potentially obtain an even better estimator in terms of mean squared error, or prediction error, by taking a weighted sum of the estimators in the considered models. The model averaged estimator with a set of weights $\{w_S; S \in \mathcal{S}\}$ is defined as $\widehat{\mu} = \sum_{S \in \mathcal{S}} w_S \widehat{\mu}_S$. In high-dimensional models, each of the estimators $\widehat{\mu}_S$ could be obtained through a penalization approach for estimation. Different sets S could represent different types of penalties, or different subsets of variables in the model.

For models with the number of parameters less than the sample size, a study of model averaged estimators (in a non-Bayesian sense) has been performed by several authors, including Yang (2001), least squares model averaging is studied by Hansen (2007), Magnus et al. (2010) compare Bayesian model averaging to weighted-average least squares (WALS). In a local misspecification setting, Claeskens and Hjort (2003) and Hjort and Claeskens (2006) study asymptotic properties of model averaged estimators for parametric and Cox regression models respectively. Zhang and Liang (2011) study frequentist model averaging (as well as a focused information criterion) for generalized additive partial linear models using polynomial spline estimators. See Claeskens and Carroll (2007) for model averaging estimators in semiparametric models using local linear estimators.

By cleverly chosen weights, it is possible that the averaged estimator has a smaller mean squared error than any of the $\widehat{\mu}_S$ separately.

One difficulty with model averaged estimators is the limiting distribution of the estimators, which is needed to perform tests, to make confidence intervals,. . . . In case the weights are data-driven (and thus random) the limiting distribution is non-trivial. Especially in the case of high-dimensional data, it is not yet well understood how to correctly perform inference with these estimators in a frequentist framework.

This is easier with Bayesian model averaging (BMA) where the posterior probability of a focus μ (now treated as a random variable) is written as a weighted average of the posterior probabilities of μ in each of the considered models $S \in \mathcal{S}$,

$$P(\mu \,|\, \mathrm{Data}) = \sum_{S \in \mathcal{S}} P(\mu \,|\, \mathrm{Data}, S) P(S \,|\, \mathrm{Data}).$$

Annest et al. (2009) use such BMA with high-dimensional micro-array data. By working with the full posterior distribution the model uncertainty is automatically taken into account, which owes to the succes of BMA methods, though comes often with a price of a high computational cost.

7 Conclusions

Building and selecting a good model for use with high-dimensional data is at least as difficult as in the low-dimensional case. Several aspects (both theoretical and practical) have been investigated already, but many more research needs to be done. In this paper we present several approaches that aim for a directed model search, a model that is good for a prespecified purpose. This 'purpose' may indicate a certain loss function, such as the averaged prediction error, or it may indicate a focus point at which prediction or estimation is to be performed (as with the FIC and transductive learning). As already indicated by several authors, one method cannot be optimal for both prediction accuracy at given points and variable consistency. This indicates that one can do better by guiding the model search by prespecifying the focus and the risk function. In other words, adding a penalty does not by itself give estimators with the smallest mean squared error or with the smallest prediction error.

Along with the growth of the literature on this topic, more will be understood about the statistical properties of such estimators, and of estimators that are constructed in such selected models. The concept of model averaging with random weights is closely related. The randomness of the selected model, or with penalization methods such as SCAD and lasso, the randomness of precisely which coefficients are set to zero, is a difficult topic to understand and to come up with solutions for practical use when conducting inference. This paper points to some issues and challenges that need further study, the full picture is as yet to be drawn.

Acknowledgements

The author likes to thank the organizers and participants of the 23rd Nordic Conference on Mathematical Statistics in Voss (Norway) where some of this material was first presented, and the editors for their encouragement to write this paper. Thanks goes also to Johan Van Kerckhoven for help with the FIC programming.

A Appendix

A.1 Proof of Theorem 1

The main steps in the proof follow the same line of thought as in Hjort and Claeskens (2003) and Claeskens and Hjort (2003).

(i) Via a Taylor series expansion, for $j = 1, \ldots, p_\gamma$,

$$\begin{aligned}
\sqrt{n}\tfrac{\partial}{\partial \boldsymbol{\theta}}\ell_n(\boldsymbol{\beta}_n) &= \sqrt{n}\tfrac{\partial}{\partial \boldsymbol{\theta}}\ell_n(\boldsymbol{\beta}_0) + \tfrac{\partial^2}{\partial \boldsymbol{\theta}\partial \boldsymbol{\gamma}^t}\ell_n(\widetilde{\boldsymbol{\beta}})\boldsymbol{\delta} \\
\sqrt{n}\tfrac{\partial}{\partial \gamma_j}\ell_n(\boldsymbol{\beta}_n) &= \sqrt{n}\tfrac{\partial}{\partial \gamma_j}\ell_n(\boldsymbol{\beta}_0) + \tfrac{\partial^2}{\partial \gamma_j\partial \boldsymbol{\gamma}^t}\ell_n(\widetilde{\boldsymbol{\beta}})\boldsymbol{\delta} - \frac{\lambda_n}{\sqrt{n}}\psi'(\tfrac{\delta_j}{\sqrt{n}})
\end{aligned}$$

with $\widetilde{\boldsymbol{\beta}}$ in between $\boldsymbol{\beta}_n$ and $\boldsymbol{\beta}_0$. By the convergence of the average of the negative of the second derivatives to the corresponding part of the matrix $\boldsymbol{J}$, together with the definition of the constant $\boldsymbol{c}$, the convergence of the score vector is proven.

(ii) We start with a Taylor-series expansion and denote the subvector $\boldsymbol{\beta}_S = (\boldsymbol{\theta}, \boldsymbol{\gamma}_S)$, and let S^c denote the index set complementary to S, that is $S^c = \{j : j \notin S\}$,

$$\begin{aligned}
\sqrt{n}\tfrac{\partial}{\partial \boldsymbol{\beta}_S}\ell_n(\boldsymbol{\beta}) &= \sqrt{n}\tfrac{\partial}{\partial \boldsymbol{\beta}_S}\ell_n(\boldsymbol{\beta}_n) + \tfrac{\partial^2}{\partial \boldsymbol{\beta}_S\partial \boldsymbol{\beta}_S^t}\ell_n(\widetilde{\boldsymbol{\beta}}) \begin{pmatrix} \sqrt{n}(\boldsymbol{\theta} - \boldsymbol{\theta}_0) \\ \sqrt{n}(\boldsymbol{\gamma}_S - \boldsymbol{\gamma}_{S,0} - \boldsymbol{\delta}_S/\sqrt{n}) \end{pmatrix} \\
&\quad - \tfrac{\partial^2}{\partial \boldsymbol{\beta}_S\partial \boldsymbol{\gamma}_{S^c}^t}\ell_n(\widetilde{\boldsymbol{\beta}})\boldsymbol{\delta}_{S^c},
\end{aligned}$$

with $\widetilde{\boldsymbol{\beta}}$ in between $\boldsymbol{\beta}_n$ and $\boldsymbol{\beta}_0$. For the second and third terms on the right-hand side we use the convergence condition on the penalty λ_n together with convergence of the negative of the second derivatives of the objective function to the corresponding elements of the matrix $\boldsymbol{J}$. This, together with result (i) allows to obtain the convergence of the parameter estimators.

(iii) This follows via an additional Taylor series expansion of $\mu(\widehat{\boldsymbol{\beta}}_S)$ about $\mu_{\rm true}$, together with the result of (ii). □

A.2 Mean and variance of the limiting distribution Λ_S in (5)

We explicitly work out the expressions of the random variables B_S and C_S,

$$\begin{aligned} \boldsymbol{B}_S &= \boldsymbol{J}_S^{00}(\boldsymbol{U}+\boldsymbol{J}_{01}\boldsymbol{\delta})+\boldsymbol{J}_S^{01}\boldsymbol{\pi}_S(\boldsymbol{c}+\boldsymbol{V}+\boldsymbol{J}_{11}\boldsymbol{\delta}) \\ \boldsymbol{C}_S &= \boldsymbol{J}_S^{10}(\boldsymbol{U}+\boldsymbol{J}_{01}\boldsymbol{\delta})+\boldsymbol{J}_S^{11}\boldsymbol{\pi}_S(\boldsymbol{c}+\boldsymbol{V}+\boldsymbol{J}_{11}\boldsymbol{\delta}). \end{aligned}$$

Inserting expressions of the blocks of the partitioned matrix $\boldsymbol{J}^{-1}$, with $\boldsymbol{J}^{11,S}=(\boldsymbol{\pi}_S(\boldsymbol{J}^{11})^{-1}\boldsymbol{\pi}_S^t)^{-1}$, and superscripts denoting blocks of the inverse matrix, gives that, with $\boldsymbol{W}=\boldsymbol{V}-\boldsymbol{J}_{10}\boldsymbol{J}_{00}^{-1}\boldsymbol{U}$,

$$\begin{aligned} E(\boldsymbol{B}_S) &= \boldsymbol{J}_{00}^{-1}\boldsymbol{J}_{01}\big((\boldsymbol{I}_q-\boldsymbol{\pi}_S^t\boldsymbol{J}^{11,S}\boldsymbol{\pi}_S(\boldsymbol{J}^{11})^{-1})\boldsymbol{\delta}-\boldsymbol{\pi}_S^t\boldsymbol{J}^{11,S}\boldsymbol{\pi}_S\boldsymbol{c}\big) \\ E(\boldsymbol{C}_S) &= \boldsymbol{J}^{11,S}\boldsymbol{\pi}_S((\boldsymbol{J}^{11})^{-1}\boldsymbol{\delta}+\boldsymbol{c}) \\ \mathrm{Var}(\boldsymbol{B}_S) &= \mathrm{Var}(\boldsymbol{J}_{00}^{-1}\boldsymbol{U})+\mathrm{Var}(\boldsymbol{J}_{00}^{-1}\boldsymbol{J}_{01}\boldsymbol{\pi}_S^t\boldsymbol{J}^{11,S}\boldsymbol{\pi}_S\boldsymbol{W}) \\ &= \boldsymbol{J}_{00}^{-1}+\boldsymbol{J}_{00}^{-1}\boldsymbol{J}_{01}\boldsymbol{\pi}_S^t\boldsymbol{J}^{11,S}\boldsymbol{\pi}_S\boldsymbol{J}_{10}\boldsymbol{J}_{00}^{-1}=\boldsymbol{J}_S^{00} \\ \mathrm{Var}(\boldsymbol{C}_S) &= \mathrm{Var}(\boldsymbol{J}^{11,S}\boldsymbol{\pi}_S\boldsymbol{W})=\boldsymbol{J}^{11,S} \\ \mathrm{Cov}(\boldsymbol{B}_S,\boldsymbol{C}_S) &= -\mathrm{Cov}(\boldsymbol{J}_{00}^{-1}\boldsymbol{J}_{01}\boldsymbol{\pi}_S^t\boldsymbol{J}^{11,S}\boldsymbol{\pi}_S\boldsymbol{W},\boldsymbol{J}^{11,S}\boldsymbol{\pi}_S\boldsymbol{W})=-\boldsymbol{J}_{00}^{-1}\boldsymbol{J}_{01}\boldsymbol{\pi}_S^t\boldsymbol{J}^{11,S}. \end{aligned}$$

Using part (iii) of Theorem 1 and the expressions above yields the expressions in (5). □

References

Annest, A., Bumgarner, R., Raftery, A., and Yeung, K. (2009). Iterative Bayesian model averaging: a method for the application of survival analysis to high-dimensional microarray data. *BMC Bioinformatics*, 20:article 72.

Antoniadis, A., Fryzlewicz, P., and Letué, F. (2010). The Dantzig selector in Cox's proportional hazards model. *Scand. J. Statist.*, 37(4):531–552.

Bartlett, P., Mendelson, S., and Neeman, J. (2009). ℓ_1-regularized linear regression: Persistence and oracle inequalities. Technical report, U.C. Berkeley.

Bartolucci, F. and Lupparelli, M. (2008). Focused information criterion for capture-recapture models for closed populations. *Scand. J. Statist.*, 35(4):629–649.

Belabbas, M.-A. and Wolfe, P. J. (2007). Spectral methods in machine learning: New strategies for very large data sets. In *Proceedings of the National Academy of Sciences of the USA (Applied Mathematics)*.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B*, 57(1):289–300.

Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009). Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.*, 37(4):1705–1732.

Brownlees, C. T. and Gallo, G. M. (2008). On variable selection for volatility forecasting: The role of focused selection criteria. *J. Finan. Econometrics*, 6(4):513–539.

Candes, E. and Tao, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n. *Ann. Statist.*, 35:2313–2351.

Caner, M. (2009). Lasso-type GMM estimator. *Econometric Theory*, 25(1):270–290.

Chapelle, O., Schölkopf, B., and Zien, A., editors (2006). *Semi-Supervised Learning*. MIT Press, Cambridge, MA.

Claeskens, G. and Carroll, R. (2007). An asymptotic theory for model selection inference in general semiparametric problems. *Biometrika*, 2:249–265.

Claeskens, G., Croux, C., and Van Kerckhoven, J. (2006). Variable selection for logistic regression using a prediction focussed information criterion. *Biometrics*, 62:972–979.

Claeskens, G., Croux, C., and Van Kerckhoven, J. (2007). Prediction focussed model selection for autoregressive models. *Aust. N. Z. J. Stat.*, 49:359–379.

Claeskens, G. and Hjort, N. L. (2003). The focused information criterion. *J. Amer. Statist. Assoc.*, 98:900–916. With discussion and a rejoinder by the authors.

Claeskens, G. and Hjort, N. L. (2008a). Minimising average risk in regression models. *Econometric Theory*, 24:493–527.

Claeskens, G. and Hjort, N. L. (2008b). *Model Selection and Model Averaging*. Cambridge University Press, Cambridge.

Cortes, C., Mohri, M., Pechyony, D., and Rastogi, A. (2008). Stability of transductive regression algorithms. In *Proceedings of the 25th international conference on Machine learning*, ICML '08, pages 176–183, New York, NY, USA. ACM.

Donoho, D. L. and Johnstone, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *J. Amer. Statist. Assoc.*, 90:1200–1224.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *Ann. Statist.*, 32:407–499. With discussion and a rejoinder by the authors.

Efron, B., Hastie, T., and Tibshiran, R. (2007). Discussion: The Dantzig selector: Statistical estimation when p is much larger than n. *Ann. Statist.*, 35(6):2358–2364.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.*, 96(456):1348–1360.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. Ser. B*, 70(5):849–911.

Fan, J. and Lv, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statist. Sinica*, 20(1):101–148.

Fan, J., Samworth, R., and Wu, Y. (2009). Ultrahigh dimensional feature selection: Beyond the linear model. *J. Mach. Learn. Res.*, 10:2013–2038.

Greenshtein, E. (2006). Best subset selection, persistence in high-dimensional statistical learning and optimization under l_1 constraint. *Ann. Statist.*, 34(5):2367–2386.

Greenshtein, E. and Ritov, Y. (2004). Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli*, 10(6):971–988.

Hansen, B. E. (2007). Least squares model averaging. *Econometrica*, 75:1175–1189.

Hesterberg, T., Choi, N. H., Meier, L., and Fraley, C. (2008). Least angle and l_1 penalized regression: a review. *Stat. Surv.*, 2:61–93.

Hjort, N. L. (2008). Focused information criteria for the linear hazard regression model. In Vonta, F and Nikulin, M and Limnios, N and HuberCarol, C, editor, *Statistical Models and Methods for Biomedical and Technical Systems*, Statistics for Industry and Technology, pages 487–502.

Hjort, N. L. and Claeskens, G. (2003). Frequentist model average estimators. *J. Amer. Statist. Assoc.*, 98:879–899. With discussion and a rejoinder by the authors.

Hjort, N. L. and Claeskens, G. (2006). Focussed information criteria and model averaging for Cox's hazard regression model. *J. Amer. Statist. Assoc.*, 101:1449–1464.

James, G. M., Radchenko, P., and Lv, J. (2009). DASSO: connections between the Dantzig selector and lasso. *J. R. Stat. Soc. Ser. B*, 71(Part 1):127–142.

Jansen, M. (2010). Minimum risk methods in the estimation of unknown sparsity. Technical report, U.L.B.

Kim, Y., Choi, H., and Oh, H.-S. (2008). Smoothly clipped absolute deviation on high dimensions. *J. Amer. Statist. Assoc.*, 103(484):1665–1673.

Knight, K. and Fu, W. (2000). Asymptotics for lasso-type estimators. *Ann. Statist.*, 28(5):1356–1378.

Leng, C., Lin, Y., and Wahba, G. (2006). A note on the Lasso and related procedures in model selection. *Statistica Sinica*, 16(4):1273–1284.

Lien, D. and Shrestha, K. (2005). Estimating the optimal hedge ratio with focus information criterion. *J. Futures Markets*, 25(10):1011–1024.

Magnus, J. R., Powell, O., and Prüfer, P. (2010). A comparison of two model averaging techniques with an application to growth empirics. *J. Econometrics*, 154(2):139–153.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.*, 34(3):1436–1462.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B*, 58:267–288.

van der Geer, S. (2008). High-dimensional generalized linear models and the Lasso. *Ann. Statist.*, 35:614–645.

Vapnik, V. N. (1995). *The nature of statistical learning theory.* Springer-Verlag, New York.

Yang, Y. (2001). Adaptive regression by mixing. *J. Amer. Statist. Assoc.*, 96(454):574–588.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B*, 68(1):49–67.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.*, 38(2):894–942.

Zhang, X. and Liang, H. (2011). Focused information criterion and model averaging for generalized additive partial linear models. *Ann. Statist.*, 39(1):174–200.

Zhao, P. and Yu, B. (2006). On model selection consistency of Lasso. *J. Mach. Learn. Res.*, 7:2541–2563.

Zou, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.*, 101(476):1418–1429.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B*, 67(2):301–320.