KU LEUVEN

FACULTY OF ECONOMICS AND BUSINESS

A high-dimensional focused information criterion

Gueuning T, Claeskens G.

KBI_1706

A High-dimensional Focused Information Criterion

Thomas Gueuning and Gerda Claeskens

ORSTAT and Leuven Statistics Research Center
KU Leuven, Faculty of Economics and Business
Naamsestraat 69, 3000 Leuven, Belgium
thomas.gueuning@kuleuven.be, gerda.claeskens@kuleuven.be

March 21, 2017

Abstract

The focused information criterion for model selection is constructed to select the model that best estimates a particular quantity of interest, the focus, in terms of mean squared error. We extend this focused selection process to the high-dimensional regression setting with potentially a larger number of parameters than the size of the sample. We distinguish two cases: (i) the case where the considered submodel is of low-dimension and (ii) the case where it is of high-dimension. In the former case, we obtain an alternative expression of the low-dimensional focused information criterion that can directly be applied. In the latter case we use a desparsified estimator that allows us to derive the mean squared error of the focus estimator. We illustrate the performance of the high-dimensional focused information criterion with a numerical study and a real dataset.

Keywords: Desparsified estimator; Focused information criterion; High-dimensional data; Variable selection.

Running headline: A high-dimensional FIC

1 Introduction

We extend the theory of the focused information criterion (FIC) for variable selection in parametric models to allow a diverging dimension of the parameter, permitting us to apply the method on high-dimensional data where the number of parameters may exceed the sample size. To do so, we extend the desparsified estimator of van de Geer et al. (2014) to the local misspecification framework. The FIC philosophy puts less emphasis on which variables are in the model but rather on the accuracy of the estimator of a focus, which is a differentiable function of the model parameters. The accuracy of the estimation is assessed via the mean squared error (MSE).

For example in the context of prediction with linear models, the FIC permits to use different variables to make predictions for different new observations of the covariate vector. We illustrate this on a real data set containing 4088 variables and 71 observations that we split in a training set of size 50 and a testing set of size 21. Whereas the usual approach consists in using the same penalized estimator and thus the same covariates to obtain the 21 predictions, the FIC allows us to use different covariates for each of the 21 different predictions. In our example, the mean squared prediction error is improved from 0.235 with a penalized estimator approach to 0.180 with the FIC approach.

The FIC has been introduced by Claeskens and Hjort (2003) for low-dimensional likelihood models, see

also Claeskens and Hjort (2008b, Ch. 6). This approach of focused selection has further been extended to several application areas including panel count data (Wang et al., 2015), graphical models (Pircalabelu et al., 2015) and personalized medicine (Yang et al., 2015). Focused selection for quantile regression has been studied by (Behl et al., 2014), and for weighted composite quantile regression by Xu et al. (2014). Focused selection for causal inference has been obtained by (Vansteelandt et al., 2012). Other model classes where focused selection has been studied include time series models (Rohan and Ramanathan, 2011; Claeskens et al., 2007), partially linear models (Zhang and Liang, 2011) and survival data (Hjort and Claeskens, 2006), without being complete in this overview.

Variable selection and estimation for high-dimensional data is most often performed simultaneously by using penalization methods; for an overview, see Fan and Lv (2010). The use of lasso-type estimators (Tibshirani, 1996) and its variations is currently well known. For theoretical results, see Bühlmann and van de Geer (2011). However, one should realize that also such methods, as do most other variable selection procedures, aim at selecting one 'best' model that one hence is supposed to use to estimate all quantities of interest related to that dataset. In contrast, the focused information criterion (FIC) may select different models for different quantities interest, which we call the focuses.

The introduction of the FIC for a diverging number of parameters is important and has a large application area. Claeskens (2012) gave a FIC formula for penalized estimators but required the dimension of the parameters to be fixed. Thus the small n (sample size) – large p (number of parameters) case is asymptotically not covered by that work. That form of FIC for penalized estimators with a fixed dimension is used by Pircalabelu et al. (2016) for high-dimensional graphical models.

Besides penalization procedures, several other variable selection procedures have been developed for high-dimensional data. In particular, Luo and Chen (2013) establish the consistency of the extended Bayesian information criterion (EBIC) with a diverging number of relevant features but need to restrict to low-dimensional submodels. Kim et al. (2012) obtain the consistency of the generalized information criterion (GIC) and Wang et al. (2009) propose a modified BIC (mBIC) whose consistency is shown for a number of parameters that diverges slower than the sample size.

The paper is organized as follows. In Section 2, we define the general framework and recall the classical FIC formula for fixed dimensions. In Section 3, we introduce the FIC for high-dimensional data when the considered submodel is of low dimension. This also provides an alternative formula in the classical FIC setting. In Section 4 we consider the high-dimensional submodel case in which $p + |S| > n$ and restrict to linear models. In that case the maximum likelihood estimator is not available because the Fisher information (sub)matrix is not invertible. To tackle this problem we use a desparsifying estimator, following the idea of van de Geer et al. (2014), Javanmard and Montanari (2014) and Zhang and Zhang (2014). In Section 5, we give some practical considerations for the computation of the FIC, including information over the estimation of $\delta\delta^\top$ and in Section 6 we give numerical results. In Section 7, we illustrate the FIC procedure on the real data set riboflavin from package *hdi* and compare it to a regular penalization approach. Section 8 provides some insights over the extension of results of Section 4 to the generalized linear models. All proofs are given in Section 9.

2 Model, notations and limitations of the current FIC literature

2.1 Notation and framework

Let $Y_1, \ldots, Y_n$ be independent response values with concomitant covariates $x_1, \ldots, x_n$, such that Y_i has a density $f(y|x_i, \theta_0, \gamma_n)$. The vector γ_n contains all the parameters on which we want to perform variable selection and has length q_n that is allowed to grow with n. The parameter vector θ_0 is of fixed length p and contains all the parameters that we want to include in every considered model. These parameters are protected. For example, for a linear model $Y_i = \beta_0 + x_i^\top \beta + \sigma\epsilon_i$ with $\epsilon_i \sim \mathcal{N}(0,1)$, it is quite common to include the parameters σ and β_0 in every considered submodel. Thus a natural choice would be $\theta_0 = (\sigma, \beta_0)$. It might also be relevant in some cases to include some of the components of β in the protected set (e.g. some variables that are known beforehand to be important). The covariate x_i is of diverging length r_n. Very often r_n and q_n are of the same order but it is not necessary the case. We present two simple examples that illustrate the link between p, q_n and r_n.

First assume that we want to fit a linear model for $(Y_i, x_i), i = 1, \ldots, n$ and that we want to include the first three components of x_i in every considered model. These three components might for instance be the age, the weight and the height of an individual while all other components might consist of blood information or gene expressions. The full model, the largest model under consideration, is then $Y_i = \beta_0 + \sum_{j=1}^{3} x_{i,j}\beta_j + \sum_{j=4}^{r_n} x_{i,j}\beta_j + \sigma\epsilon_i$ and we have $\theta_0 = (\sigma, \beta_0, \beta_1, \beta_2, \beta_3)$ and $\gamma_n = (\beta_4, \ldots, \beta_{r_n})$. Thus $p = 5$ and $q_n = r_n - 3$.

In a second example, assume that we still want to fit a linear model for (Y_i, x_i) but that none of the components of x_i should be protected. Furthermore assume that we want to consider interaction terms as well as first and second order terms in the possible models. Then $p = 2$ (for the error standard deviation level and the intercept) and $q_n = r_n + r_n(r_n + 1)/2$.

As in the earlier studies about FIC (e.g. Claeskens and Hjort, 2003; Claeskens, 2012) we consider the local misspecification framework where $\gamma_n = \gamma_0 + \delta/\sqrt{n}$, with the major difference that the length q_n of δ is diverging. Each component of δ is $O(1)$. This framework allows us to study the mean squared error (MSE) of the estimator of the further-defined focus, with a balance between the squared bias and the variance, without having the bias or the variance dominating the mean squared error expression. We refer to Claeskens and Hjort (2008b, Sec. 5.5) for more details regarding the local misspecification setting.

Taking $\gamma_n = \gamma_0$, a known value, corresponds to working with the simplest model, often called *the narrow model*. In the two examples given hereabove, it is natural to choose $\gamma_{0,j} = 0$ for each j: the simplest model consists in not including the unprotected variables. In other cases, $\gamma_{0,j}$ might be nonzero, see for instance example 5.4 in Claeskens and Hjort (2008b) in which the skewing logistic regression model $p_i = H(x_i^\top \beta + z_i^\top \alpha)^\kappa$ is considered. In that example, κ is an unprotected parameter that takes value 1 in the narrow model.

We denote by $S_{0,n} = \{j : \delta_j \neq 0\}$ the active set of coefficients where we emphasize in the notation the fact that the length of δ is growing with n and we write $s_n = |S_{0,n}|$, the number of elements of $S_{0,n}$. We

consider subsets S of $\{1, \ldots, q_n\}$ and denote by (sub)model S the model containing θ and those parameters γ_j with j belonging to S. This model corresponds to working with a density $f(y|x, \theta, \gamma_S, \gamma_{0,S^c})$ with S^c the complementary set of S. The slight abuse of notation groups the components of γ by whether their index is present or absent in S. When fitting a model with the p protected parameters in θ, and $|S|$ added parameters, in total $p + |S|$ parameters need to be estimated. Let us denote by $(\hat{\theta}_S, \hat{\gamma}_S)$ an estimator of $(\theta_0, \gamma_{n,S})$. See Sections 3 and 4 for more details.

2.2 The focused information criterion

Following the FIC philosophy, we are interested in estimating as accurately as possible (in terms of MSE) a particular quantity of interest $\mu_{\text{true}} = \mu(\theta_0, \gamma_n)$, called *the focus*. A model that is best in terms of MSE for one focus μ, might not be the best for another focus. This leads to a tailored model choice where one first specifies the focus and then searches for the best model for that particular goal. In this sense it should be clear that the FIC is not constructed to aim for selection consistency.

We make the assumption that μ is differentiable with respect to θ and γ such that $\|[(\frac{\partial \mu}{\partial \theta})^\top, (\frac{\partial \mu}{\partial \gamma_S})^\top]\|_\infty = K = O(1)$ in a neighborhood of θ_0, γ_0. Several examples of such quantities of interest are given in Claeskens and Hjort (2008b). The focus might for example be the prediction for a particular subgroup of the population, the estimation of the impact of one particular covariate on the response or a particular quantile for a specific value of the covariates. The goal is to find the submodel S whose corresponding estimator $\hat{\mu}_S = \mu(\hat{\theta}_S, \hat{\gamma}_S, \gamma_{0,S^c})$ of the focus is the best in terms of mean squared error. For a submodel S we are thus interested in the limiting distribution Λ_S of $\sqrt{n}(\hat{\mu}_S - \mu_{\text{true}})$. The focused information criterion estimates the corresponding limiting mean squared error. Thus,

$$\text{FIC}(S) = \widehat{\text{E}(\Lambda_S)^2} + \widehat{\text{Var}(\Lambda_S)}.$$

Different models, say indexed by S and S', might have different values for the bias and variance of the submodel-based estimator of μ, thus Λ_S might be different from $\Lambda_{S'}$. Hence, models can be ranked based on their FIC value. The model S with the smallest FIC(S) value amongst all considered models, is selected as the best one for the purpose of estimating the focus μ.

In the low-dimensional framework with γ_n and δ of fixed dimensions $q \times 1$, Claeskens and Hjort (2003) show that if $(\hat{\theta}_S, \hat{\gamma}_S)$ is the maximum likelihood estimator, the limiting MSE of Λ_S is

$$\text{MSE}(S) = \omega^\top (I_q - G_S)\delta\delta^\top (I_q - G_S)^\top \omega + \left(\frac{\partial \mu}{\partial \theta}\right)^\top J_{00}^{-1} \frac{\partial \mu}{\partial \theta} + \omega^\top G_S Q_S G_S^\top \omega, \tag{1}$$

with the $(p+q) \times (p+q)$ Fisher information matrix J and its inverse matrix denoted by

$$J = \begin{pmatrix} J_{00} & J_{01} \\ J_{10} & J_{11} \end{pmatrix}, \quad J^{-1} = \begin{pmatrix} J^{00} & J^{01} \\ J^{10} & J^{11} \end{pmatrix},$$

where $Q = J^{11}$, $G_S = \pi_S^\top Q_S \pi_S Q^{-1}$, $Q_S = J^{11,S}$, $\omega = J_{10} J_{00}^{-1} \frac{\partial \mu}{\partial \theta} - \frac{\partial \mu}{\partial \gamma}$ and $\pi_S \in \mathbb{R}^{S \times q}$ the projection matrix related to S, obtained by extracting the rows of the $q \times q$ identity matrix for which the row number is in S. Claeskens and Hjort (2008b, Sec. 6.7) show that for linear models the limiting MSE is in fact

the exact MSE when the focus takes the form $\mu = x^\top\beta + z^\top\gamma$. For a vector-valued focus one could use a one-dimensional summary of the corresponding mean squared error matrix to be minimized over the different models, such as the matrix' trace, determinant, a matrix norm, etc.

The current FIC formula (1) can not be applied in our framework for two reasons. First, and most importantly, the theory assumes that the dimension of γ_n is fixed. A diverging number of parameters is not supported by the theory. In this paper, we allow the dimension q_n of γ_n to grow with n and we make the sparsity assumption $s_n = o(n^{1/4})$. Secondly, the current version of the FIC formula is in many cases not available for high-dimensional data, even for low-dimensional submodels. Indeed, it requires to invert the Fisher information matrix J that is in many cases not invertible. For example, for a normal linear model, the Fisher information matrix $J = \sigma^{-2}\text{diag}\{2, n^{-1}X^\top X\}$. When $q_n > n$ the matrix J is by construction not invertible so that the expression (1) is not defined. These considerations motivate us to develop the FIC theory for a diverging number of parameters.

We distinguish two cases in the model selection search: (i) the submodel is low-dimensional such that regular least squares or maximum likelihood estimators can be computed, and (ii) the submodel is high-dimensional, requiring a regularized estimator. In both cases, an adjustment of the existing focused selection approach is needed. These two cases are studied in the next two sections.

We now give some notations. For two random variables A and B, the notation $A \doteq_d B$ means that $A - B \xrightarrow{p} 0$. Furthermore, we write $f_{\text{true}}(y|x) = f(y|x, \theta_0, \gamma_0 + \delta/\sqrt{n})$ the true density function, $f_0(y|x) = f(y|x, \theta_0, \gamma_0)$ the density function in the narrow model and $U(y|x) = \frac{\partial}{\partial\theta}\log f(y|x,\theta,\gamma)_{|(\theta_0,\gamma_0)}$ and $V(y|x) = \frac{\partial}{\partial\gamma}\log f(y|x,\theta,\gamma)_{|(\theta_0,\gamma_0)}$ the derivatives of the log-density evaluated in the narrow model. We define in the regression model's context

$$J(x) = \int f_0(y|x)\begin{pmatrix} U(y|x) \\ V(y|x)\end{pmatrix}\begin{pmatrix} U(y|x) \\ V(y|x)\end{pmatrix}^\top dy, \quad J_n = \frac{1}{n}\sum_{i=1}^n J(x_i) = \begin{pmatrix} J_{n,00} & J_{n,01} \\ J_{n,10} & J_{n,11}\end{pmatrix},$$

the latter matrix is the empirical Fisher information matrix. For a fixed subset S of $\{1, \ldots, q_n\}$, we denote by π_S the $|S|\times q_n$ projection matrix related to S that, when multiplied to a matrix or row vector consisting of q_n rows, selects those rows corresponding to the elements in S. Further, define

$$\pi_S^* = \begin{pmatrix} I_p & 0_{p\times q_n} \\ 0_{|S|\times p} & \pi_S\end{pmatrix}, \quad J_S(x) = \pi_S^* J(x)\pi_S^{*\top} = \int f_0(y|x)\begin{pmatrix} U(y|x) \\ V_S(y|x)\end{pmatrix}\begin{pmatrix} U(y|x) \\ V_S(y|x)\end{pmatrix}^\top dy$$

and write $J_{n,S} = \frac{1}{n}\sum_{i=1}^n J_S(x_i)$ the empirical Fisher matrix in model S and $J_S = \lim_{n\to\infty} J_{n,S}$. Note that $J_{n,S}$ is of fixed dimension $(p+|S|)\times(p+|S|)$ while J_n is of diverging dimension $(p+q_n)\times(p+q_n)$. As a consequence, for an unbounded sequence $\{q_n, n\to\infty\}$, J_n does not converge to a fixed quantity J.

3 FIC for a low-dimensional submodel

We consider the local misspecification framework of Section 2. Let S be a fixed subset of $\{1, \ldots, q_n\}$ such that the number of parameters $p + |S|$ to estimate in the submodel S is smaller than the sample size n.

Let us consider the maximum likelihood estimator for $(\theta_0, \gamma_{n,S})$

$$(\hat{\theta}_S, \hat{\gamma}_S) = \underset{\theta\in\mathbb{R}^p \gamma_S\in\mathbb{R}^{|S|}}{\arg\max} \frac{1}{n}\sum_{i=1}^{n} \log\ f(y_i|x_i, \theta, \gamma_S, \gamma_{0,S^c}) \tag{2}$$

and define the estimator of the focus in this model by

$$\hat{\mu}_S = \mu(\hat{\theta}_S, \hat{\gamma}_S, \gamma_{0,S^c}). \tag{3}$$

Before presenting our theoretical result for the limiting distribution of $\sqrt{n}(\hat{\mu}_S - \mu_{\text{true}})$ we give the corresponding conditions. A Taylor expansion of $f_{\text{true}}(y|x)$ gives

$$f_{\text{true}}(y|x) = f_0(y|x)\left\{1 + V(y|x)^\top \delta/\sqrt{n} + R(y|x, \delta/\sqrt{n})\right\}. \tag{4}$$

We make the following conditions, that are similar to Hjort and Claeskens (2003), phrased here for the regression setting.

(C1) The two integrals $\int f_0(y|x)U(y|x)R(y,t)dy$ and $\int f_0(y|x)V(y|x)R(y,t)dy$ are both $O(\|t\|_1^2)$, with R defined in (4).

(C2) The variables $|U_l(y|x)^2 V_k(y|x)|$ and $|V_j(y|x)^2 V_k(y|x)|$ have finite mean under $f_0(y|x)$ for each $1 \leq l \leq p$ and $j, k \in S$.

(C3) The two integrals $\int f_0(y|x)\left\|U(y|x)\right\|^2 R(y,t)dy$ and $\int f_0(y|x)\left\|V_S(y|x)\right\|^2 R(y,t)dy$ are both $o(1)$.

(C4) The log density has three continuous derivatives w.r.t the $p+|S|$ parameters (θ, γ_S) in a neighbourhood around (θ_0, γ_0), and they are dominated by functions with finite means under f_0.

(C5) $s_n = o(n^{1/4})$.

Conditions (C1) to (C4) are similar to those of Hjort and Claeskens (2003) in the low-dimensional case, while condition (C5) is a sparsity condition to deal with high-dimensional vectors.

Lemma 1. *Under (C1), (C2), (C3) and (C5), we have*

$$\begin{pmatrix} \sqrt{n}\bar{U}_n \\ \sqrt{n}\bar{V}_{n,S} \end{pmatrix} - \begin{pmatrix} J_{n,01}\delta \\ \pi_S J_{n,11}\delta \end{pmatrix} \xrightarrow{d} \mathcal{N}_{p+|S|}(0, J_S)$$

with $\bar{U}_n = \frac{1}{n}\sum_{i=1}^n U(y_i|x_i)$, $\bar{V}_{n,S} = \frac{1}{n}\sum_{i=1}^n V_S(y_i|x_i)$.

Lemma 2. *Under (C1), (C2), (C3), (C4) and (C5), we have*

$$\begin{pmatrix} \sqrt{n}(\hat{\theta}_S - \theta_0) \\ \sqrt{n}(\hat{\gamma}_S - \gamma_{0,S}) \end{pmatrix} - J_S^{-1}\begin{pmatrix} J_{n,01}\delta \\ \pi_S J_{n,11}\delta \end{pmatrix} \xrightarrow{d} \mathcal{N}_{p+|S|}(0, J_S^{-1}).$$

The following theoretical result is an extension of Theorem 6.1 of Claeskens and Hjort (2008b) to the diverging number of parameters case. It covers the important $p + q > n$ case and can thus be applied on high-dimensional data. A proof is given in Section 9.

Theorem 1. *Under conditions (C1) to (C5) it holds for the estimator (3) of the focus in model S that*

$$\sqrt{n}(\hat{\mu}_S - \mu_{\text{true}}) \ \doteq_d \ \Lambda_{n,S}$$

with

$$\Lambda_{n,S} = \left(\frac{\partial \mu}{\partial \theta}\right)^\top C_S + \left(\frac{\partial \mu}{\partial \gamma_S}\right)^\top D_S - \left(\frac{\partial \mu}{\partial \gamma}\right)^\top \delta \tag{5}$$

$$= \begin{pmatrix} \frac{\partial \mu}{\partial \theta} \\ \frac{\partial \mu}{\partial \gamma} \end{pmatrix}^\top \left(B_S \delta + \pi_S^{*\top} J_S^{-1} \begin{pmatrix} U \\ V_S \end{pmatrix} \right) \tag{6}$$

where the partial derivatives are evaluated at the null point (θ_0, γ_0) and where

$$\begin{pmatrix} C_S \\ D_S \end{pmatrix} = J_S^{-1} \begin{pmatrix} J_{n,01}\delta + U \\ \pi_S J_{n,11}\delta + V_S \end{pmatrix} \ \textit{with} \ \begin{pmatrix} U \\ V_S \end{pmatrix} \sim \mathcal{N}_{p+|S|}(0, J_S),$$

and

$$B_S = \pi_S^{*\top} J_S^{-1} \begin{pmatrix} J_{n,01} \\ \pi_S J_{n,11} \end{pmatrix} - \begin{pmatrix} 0_{p \times q_n} \\ I_{q_n} \end{pmatrix}.$$

The sparsity condition $s_n = o(n^{1/4})$ is crucial in this high-dimensional framework. Note that $\Lambda_{n,S}$ depends on n through $\frac{\partial \mu}{\partial \gamma}$, B_S and δ. While (5) leads to the original FIC formula, (6) turns out to be more useful in the high-dimensional case. From Theorem 1, for a model S, the limiting distribution of $\sqrt{n}(\hat{\mu}_S - \mu_{\text{true}})$ is the same as the one of $\Lambda_{n,S}$ which is normally distributed with mean and variance given by

$$\text{E}(\Lambda_{n,S}) = \begin{pmatrix} \frac{\partial \mu}{\partial \theta} \\ \frac{\partial \mu}{\partial \gamma} \end{pmatrix}^\top B_S \delta \ \text{ and } \ \text{Var}(\Lambda_{n,S}) = \begin{pmatrix} \frac{\partial \mu}{\partial \theta} \\ \frac{\partial \mu}{\partial \gamma} \end{pmatrix}^\top \pi_S^{*\top} J_S^{-1} \pi_S^* \begin{pmatrix} \frac{\partial \mu}{\partial \theta} \\ \frac{\partial \mu}{\partial \gamma} \end{pmatrix}.$$

We thus have

$$\text{MSE}(S, \delta) = \begin{pmatrix} \frac{\partial \mu}{\partial \theta} \\ \frac{\partial \mu}{\partial \gamma} \end{pmatrix}^\top \left(B_S \delta \delta^\top B_S^\top + \pi_S^{*\top} J_S^{-1} \pi_S^* \right) \begin{pmatrix} \frac{\partial \mu}{\partial \theta} \\ \frac{\partial \mu}{\partial \gamma} \end{pmatrix}, \tag{7}$$

and $\text{FIC}(S, \delta) = \widehat{\text{MSE}(S, \delta)}$ which defines the FIC in the high-dimensional setting for a low-dimensional submodel S. Interestingly, this formulation does not require the inverse of the Fisher matrix in the full model but only in the submodel S. Thus this expression may be used in the high-dimensional setting with $p + q_n > n$ if the considered submodel S is of low dimension, that is if $p + |S| \leq n$.

In fact, the formula (7) could also be used to compute the FIC in the classical fixed low dimensional case. Indeed, it is possible to show that for fixed q with $p + q < n$ the expressions (1) and (7) are equal, with $J_{n,01}$ and $J_{n,11}$ replaced by their limiting versions J_{01} and J_{11}, this is that

$$\omega^\top (I_q - G_S)\delta\delta^\top (I_q - G_S)^\top \omega + \left(\frac{\partial \mu}{\partial \theta}\right)^\top J_{00}^{-1} \frac{\partial \mu}{\partial \theta} + \omega^\top G_S Q_S G_S^\top \omega$$

is equal to

$$\begin{pmatrix} \frac{\partial \mu}{\partial \theta} \\ \frac{\partial \mu}{\partial \gamma} \end{pmatrix}^\top \left(B_S \delta \delta^\top B_S^\top + \pi_S^{*\top} J_S^{-1} \pi_S^* \right) \begin{pmatrix} \frac{\partial \mu}{\partial \theta} \\ \frac{\partial \mu}{\partial \gamma} \end{pmatrix}.$$

This can be obtained using $Q^{-1} = J_{11} - J_{10}J_{00}^{-1}J_{01}$ and $G_S Q_S G_S^\top = \pi_S^\top Q_S \pi_S$. The main novel contribution of the low-dimensional submodel case, though, is that the theory takes the presence of the high-dimensional vector δ into account.

To conclude this section, we note that if we wish to not protect any variable in the model selection procedure, our theory is still valid. In that case the Fisher information matrix is a $q_n \times q_n$ matrix and Theorem 1 and expression (7) are still valid with slight adjustments. The partial derivative $\frac{\partial \mu}{\partial \theta}$ disappears, B_S becomes $\pi_S^\top J_S \pi_S J_n - I_q$ and π_S^* becomes π_S. This remark also holds for the high-dimensional submodel case in the next section.

4 FIC for a high-dimensional submodel of a linear model

Let S be a subset of $\{1, \ldots, q_n\}$ of size larger than $n - p$. The maximum likelihood estimator (or least-squares estimator) is not available anymore and the results from Section 3 are not applicable. We propose to first use a ℓ_1-penalized estimator and then to desparsify it to obtain an estimator of $(\theta_0, \gamma_{n,S})$ whose distribution can be tracked. The idea to desparsify a penalized estimator has been introduced by several authors, including van de Geer et al. (2014), Javanmard and Montanari (2014) and Zhang and Zhang (2014). In this section, we restrict to linear models but extensions to generalized linear models and convex loss functions are expected to be feasible.

The desparsification is needed because the ℓ_1-based penalties have the property of setting some of the coefficients exactly equal to zero, one can show asymptotic consistency of such selection under some conditions. The remaining non-zero coefficients are estimated by an estimator which can asymptotically be normally distributed. This is the case for the adaptive Lasso (see Zou, 2006) and the SCAD (see Fan and Li, 2001). Since the focus might be a function of both types of coefficients, those that will be estimated by zero and those that will not, the asymptotic distribution of the focus estimator is not tractable due to this mixture containing a point-mass at zero.

Let us assume that for $i = 1, \ldots, n$, the response Y_i is generated by a linear model

$$Y_i = x_{\beta,i}^\top \beta_0 + x_{\gamma,i}^\top \gamma_n + \sigma \epsilon_i \tag{8}$$

with $\epsilon_i \sim \mathcal{N}(0,1)$, where $\beta_0 \in \mathbb{R}^p$ corresponds to the protected variables, $x_{\beta,i}$ is a $p \times 1$ vector of protected covariates, $\gamma_n \in \mathbb{R}^{q_n}$ corresponds to the unprotected parameters with corresponding covariate vector $x_{\gamma,i}$ on which variable selection is performed.

As in most of the high-dimensional literature, we assume that the noise variance σ^2 is known. Reid et al. (2016) describe strategies for estimating σ^2 and their empirical comparison suggests that using the estimator based on the residual sum of squares of cross-validated Lasso solution might yield a good estimator. For theoretical properties we refer to this paper. With σ^2 assumed to be known, the protected parameter θ_0 is thus β_0 and we note that for a linear model, $\gamma_0 = 0_{q_n}$ so that we have $\gamma_n = \delta/\sqrt{n}$ in this

section. We write

$$X_\beta = \begin{bmatrix} x_{\beta,1}^\top \\ \vdots \\ x_{\beta,n}^\top \end{bmatrix} \in \mathbb{R}^{n\times p},\ X_\gamma = \begin{bmatrix} x_{\gamma,1}^\top \\ \vdots \\ x_{\gamma,n}^\top \end{bmatrix} \in \mathbb{R}^{n\times q_n},\ X_{\gamma,S} = X_\gamma \pi_S^\top \text{ and } X_S^* = [X_\beta, X_{\gamma,S}] \in \mathbb{R}^{n\times(p+|S|)}.$$

The matrix X_S^* corresponds to the design matrix in the submodel S. Denoting by Y the vector of responses and ϵ the vector of the errors, we have $Y = X_\beta\beta_0 + X_\gamma\gamma_n + \epsilon$.

This section proceeds as follows. First, in section 4.1, we derive a desparsified estimator that can be interpreted as a generalization of the ordinary least-sqaures estimator. In section 4.2, we describe how to construct a relaxed inverse of the sample covariance matrix. Next, in section 4.3, we consider the case that a submodel S contains the true active set and derive theoretical results. In section 4.4, we derive a FIC formula for a general submodel S.

4.1 Desparsified estimator

Let us consider the following Lasso estimator (Tibshirani, 1996) where we do not penalize the intercept parameter (or take a model without intercept by centering the variables),

$$(\hat\beta_S^{\text{Lasso}}, \hat\gamma_S^{\text{Lasso}}) = \underset{\beta\in\mathbb{R}^p, \gamma_S\in\mathbb{R}^{|S|}}{\arg\min} \frac{1}{2n}\left\| Y - X_S^* \begin{pmatrix}\beta \\ \gamma_S\end{pmatrix}\right\|_2^2 + \lambda \left\| \begin{pmatrix}\beta \\ \gamma_S\end{pmatrix}\right\|_1. \tag{9}$$

We describe how to construct a desparsified estimator. The derivation presented herebelow is based on van de Geer et al. (2014). We write the Karush-Kuhn-Tucker condition

$$\frac{1}{n} X_S^{*\top}\left(Y - X_S^* \begin{pmatrix}\hat\beta_S^{\text{Lasso}} \\ \hat\gamma_S^{\text{Lasso}}\end{pmatrix}\right) = \lambda\hat\kappa_S; \quad \text{with } \hat\kappa_{S,j} = \text{sign}\left(\begin{pmatrix}\hat\beta_S^{\text{Lasso}} \\ \hat\gamma_S^{\text{Lasso}}\end{pmatrix}_j\right) \text{ if } \begin{pmatrix}\hat\beta_S^{\text{Lasso}} \\ \hat\gamma_S^{\text{Lasso}}\end{pmatrix}_j \neq 0, \tag{10}$$

where $\|\hat\kappa_S\|_\infty \leq 1$.

The matrix $J_S = \frac{1}{n\sigma^2} X_S^{*\top} X_S^*$ is by construction not invertible because $p+|S| > n$. We construct a relaxed inverse M_S of J_S by using the Lasso nodewise regression technique, as presented in van de Geer et al. (2014) and in section 4.2, and we define the following desparsified estimator:

$$\begin{aligned} \begin{pmatrix}\hat\beta_S^{\text{desp}} \\ \hat\gamma_S^{\text{desp}}\end{pmatrix} &= \begin{pmatrix}\hat\beta_S^{\text{Lasso}} \\ \hat\gamma_S^{\text{Lasso}}\end{pmatrix} + M_S \tfrac{1}{n\sigma^2} X_S^{*\top}\left(Y - X_S^* \begin{pmatrix}\hat\beta_S^{\text{Lasso}} \\ \hat\gamma_S^{\text{Lasso}}\end{pmatrix}\right) \\ &= M_S \tfrac{1}{n\sigma^2} X_S^{*\top} Y + \left(I_{p+|S|} - M_S J_S\right) \begin{pmatrix}\hat\beta_S^{\text{Lasso}} \\ \hat\gamma_S^{\text{Lasso}}\end{pmatrix}. \end{aligned} \tag{11}$$

We now give some intuition of the desparsfying estimator defined in (11). It can be seen as a bias-corrected version of the Lasso (first line) or as what we could call a pseudo-least-squares estimator in a high-dimensional framework (second line). We focus on the second interpretation. Since J_S is not invertible and M_S is used as a relaxed inverse, we could use the estimator $M_S \frac{1}{n\sigma^2} X_S^{*\top} Y$. This estimator

has mean $M_S J_S(\beta_0^\top, \gamma_{0,S}^\top + \delta_S^\top/\sqrt{n})^\top$ and variance $\frac{1}{n} M_S J_S M_S^\top$. We aim to correct this bias by adding $(I_{p+|S|} - M_S J_S)(\hat{\beta}^\top, \hat{\gamma}_S^\top)^\top$ using a reasonable estimator of the parameter vector. Here and in several referenced papers, the lasso estimator is taken.

By plugging $Y = X_\beta \beta_0 + X_\gamma \delta/\sqrt{n}$ into (11), we obtain the following equalities.

$$\begin{aligned}
\begin{pmatrix} \sqrt{n}(\hat{\beta}_S^{\text{desp}} - \beta_0) \\ \sqrt{n}(\hat{\gamma}_S^{\text{desp}} - \gamma_0) \end{pmatrix} &= M_S \begin{pmatrix} J_{01}\delta \\ \pi_S J_{11}\delta \end{pmatrix} + \left(I_{p+|S|} - M_S J_S\right) \begin{pmatrix} 0_p \\ \delta_S \end{pmatrix} \\
&\quad + M_S \tfrac{1}{\sqrt{n}\sigma^2} X_S^{*\top} \epsilon - \sqrt{n}\left(I_{p+|S|} - M_S J_S\right) \begin{pmatrix} \hat{\beta}_S^{\text{Lasso}} - \beta_0 \\ \hat{\gamma}_S^{\text{Lasso}} - \frac{\delta_S}{\sqrt{n}} \end{pmatrix} \\
&= \begin{pmatrix} 0_p \\ \delta_S \end{pmatrix} + M_S \tfrac{1}{n\sigma^2} X_S^{*\top} X_{\gamma, S^c} \delta_{S^c} \\
&\quad + M_S \tfrac{1}{\sqrt{n}\sigma^2} X_S^{*\top} \epsilon - \sqrt{n}\left(I_{p+|S|} - M_S J_S\right) \begin{pmatrix} \hat{\beta}_S^{\text{Lasso}} - \beta_0 \\ \hat{\gamma}_S^{\text{Lasso}} - \frac{\delta_S}{\sqrt{n}} \end{pmatrix}.
\end{aligned} \tag{12}$$

The right hand side of the second line of equation (12) has a very clear interpretation. It consists of a sum of four elements. The first two are related to the local misspecification, the third one is a variance term and the fourth one is a bias term that is shown in Theorem 2 to be $o_p(1)$ if $S_{0,n} \subseteq S$. Before stating our theoretical results and defining the FIC, we describe how to construct the relaxed inverse M_S.

4.2 Nodewise regression

Before stating our theoretical result we briefly describe how we construct the matrix M_S which acts as a relaxed inverse of J_S. We follow the methodology of van de Geer et al. (2014). For each $j \in \{1, \ldots, p + |S|\}$ we compute

$$\hat{\eta}_j = \underset{\eta \in \mathbb{R}^{p+|S|-1}}{\arg\min} \frac{1}{2n} \left\| X_{S,j}^* - X_{S,-j}^* \eta \right\|_2^2 + \lambda_j \|\eta\|_1,$$

where $X_{S,j}^*$ is the j-th column of X_S^* and $X_{S,-j}^* \in \mathbb{R}^{n \times (p+|S|-1)}$ is X_S^* without its j-th column, and we form

$$\hat{A}_S = \begin{bmatrix} 1 & -\hat{\eta}_{1,2} & \cdots & \hat{\eta}_{1,p+|S|} \\ -\hat{\eta}_{2,1} & 1 & \cdots & \hat{\eta}_{2,p+|S|} \\ \vdots & \vdots & \ddots & \vdots \\ -\hat{\eta}_{p+|S|,1} & -\hat{\eta}_{p+|S|,2} & \cdots & \hat{\eta}_{p+|S|,p+|S|} \end{bmatrix}$$

with components of $\hat{\eta}_j$ indexed by $k \in \{1, \ldots, j-1, j+1, \ldots, p+|S|\}$. We define

$$M_S = \hat{T}_S^{-2} \hat{A}_S$$

with $\hat{T}_S^2 = \text{diag}(\hat{\tau}_1^2, \ldots, \hat{\tau}_{p+|S|}^2)$ and $\hat{\tau}_j^2 = \frac{1}{n} \left\| X_{S,j}^* - X_{S,-j}^* \hat{\eta}_j \right\|_2^2 + \lambda_j \|\hat{\eta}_j\|_1$.

4.3 Submodel containing the true active set: theoretical results

In this section, we assume that the submodel S contains the true active $S_{0,n}$ of γ_n. We state the following conditions.

(A1) For the true active set $\{1,\dots,p\} \cup \{p+j : j \in S_{0,n}\}$, the compatibility condition holds for $\hat{\Sigma}_S = \frac{1}{n} X_S^{*\top} X_S^*$ with compatibility constant $\phi_0^2 > 0$, this is for all β and γ satisfying $\|\gamma_{S_{0,n}^c}\|_1 \leq 3(\|\beta\|_1 + \|\gamma_{S_{0,n}}\|_1)$, it holds that $(\|\beta\|_1 + \|\gamma_{S_{0,n}}\|_1)^2 \leq \begin{pmatrix} \beta \\ \gamma_S \end{pmatrix}^\top \hat{\Sigma}_S \begin{pmatrix} \beta \\ \gamma_S \end{pmatrix} (p+s_n)/\phi_0^2$. Furthermore, $\max_j \hat{\Sigma}_{S,j,j} \leq N^2$ for some $0 < N < \infty$.

(A2) For each j we take $\lambda_j = O(\sqrt{\log(p+|S|)/n}$ in the nodewise regression procedure and we have $\hat{\tau}_j \geq C > 0$.

(A3) $s_n = o(n^{1/4})$.

Assumption (A1) is common in the high-dimensional literature, see for example Bühlmann and van de Geer (2011), and assumption (A2) corresponds to (B1) in Bühlmann and van de Geer (2015). (A3) is a sparsity condition, which is the same as assumption (C5).

The following lemma follows from Theorem 2.1 of van de Geer et al. (2014) applied to the model $Y = X_\beta \beta_0 + X_{\gamma,S}\gamma_{n,S} + \epsilon$ (which holds if $S_{0,n} \subseteq S$) under the local misspecification framework $\gamma_n = \gamma_0 + \delta/\sqrt{n}$.

Lemma 3. *Let us consider the linear model (8). Let S be a subset of $\{1,\dots,q_n\}$ such that $S_{0,n} \subseteq S$ and let $t > 0$ be arbitrary. Under conditions (A1), if $\lambda \geq 2N\sigma\sqrt{2(t^2+\log(p+|S|))/n}$ we have:*

$$\begin{pmatrix} \sqrt{n}(\hat{\beta}_S^{\text{desp}} - \beta_0) \\ \sqrt{n}\hat{\gamma}_S^{\text{desp}} \end{pmatrix} \doteq_d \begin{pmatrix} C_S \\ D_S \end{pmatrix} + \Delta_1, \tag{13}$$

with

$$\begin{pmatrix} C_S \\ D_S \end{pmatrix} \sim \mathcal{N}_{p+|S|}\left(\begin{pmatrix} 0_p \\ \delta_S \end{pmatrix}, M_S J_S M_S^\top \right)$$

and

$$\mathbb{P}\left[\|\Delta_1\|_\infty \geq 8\sqrt{n}\left(\max_j \frac{\lambda_j}{\hat{\tau}_j^2}\right) \frac{\lambda(p+s_n)}{\phi_0^2} \right] \leq 2\exp(-t^2),$$

with λ_j and $\hat{\tau}_j^2$ being the tuning parameter and the residual sum of squares of the regression of $X_{S,j}^*$ on $X_{S,-j}^*$ in the nodewise regression procedure.

Using Lemma 3, we can obtain the distribution of the focus estimator.

Theorem 2. *Let consider the linear model (8). Let S be a subset of $\{1,\dots,q\}$ such that $S_{0,n} \subseteq S$ and let $t > 0$ be arbitrary. Under conditions (A1), (A2) and (A3), if $\lambda \geq 2N\sigma\sqrt{2(t^2+\log(p+|S|))/n}$ we have for $\hat{\mu}_S = \mu(\hat{\beta}_S^{\text{desp}}, \hat{\gamma}_S^{\text{desp}}, 0_{|S^c|}^\top)$ and $\mu_{\text{true}} = \mu(\beta_0, \gamma_n)$*

$$\sqrt{n}(\hat{\mu}_S - \mu_{\text{true}}) \doteq_d \Lambda_S + \Delta_2,$$

where

$$\Lambda_S = \left(\frac{\partial \mu}{\partial \theta}\right)^\top C_S + \left(\frac{\partial \mu}{\partial \gamma_S}\right)^\top D_S - \left(\frac{\partial \mu}{\partial \gamma}\right)^\top \delta = \begin{pmatrix} \frac{\partial \mu}{\partial \theta} \\ \frac{\partial \mu}{\partial \gamma} \end{pmatrix}^\top \pi_S^{*\top} M_S \begin{pmatrix} U \\ V_S \end{pmatrix}$$

with $(U^\top, V_S^\top) \sim \mathcal{N}_{p+|S|}(0, J_S)$

and

$$\mathbb{P}\left[\Delta_2 \geq 8K(p+|S|)\sqrt{n}\left(\max_j \frac{\lambda_j}{\hat{\tau}_j^2}\right)\frac{\lambda(p+s_n)}{\phi_0^2}\right] \leq 2\exp(-t^2).$$

Using a regularization parameter λ of order $\sqrt{\log(p+|S|)/n}$ and under assumption (A2), Δ_2 can be neglected if $s_n = o(\sqrt{n}/\{(p+|S|)\log(p+|S|)\})$ which holds thanks to (A3) and the fact that p and S are fixed.

For a model S, under conditions of Theorem 2 and with adequate tuning parameters and sparsity assumption, the limiting distribution Λ_S of $\sqrt{n}(\hat{\mu}_S - \mu_{\text{true}})$ is normal with mean $E(\Lambda_S) = 0$ and variance

$$\text{Var}(\Lambda_S) \quad = \quad \begin{pmatrix} \frac{\partial \mu}{\partial \theta} \\ \frac{\partial \mu}{\partial \gamma} \end{pmatrix}^\top \pi_S^{*\top} M_S J_S M_S^\top \pi_S^* \begin{pmatrix} \frac{\partial \mu}{\partial \theta} \\ \frac{\partial \mu}{\partial \gamma} \end{pmatrix}.$$

It is logical to observe a null bias because we make the assumption that the true active set is included in the considered submodel. The limiting mean squared error is thus

$$\text{MSE}(S,\delta) = \begin{pmatrix} \frac{\partial \mu}{\partial \theta} \\ \frac{\partial \mu}{\partial \gamma} \end{pmatrix}^\top \pi_S^{*\top} M_S J_S M_S^\top \pi_S^* \begin{pmatrix} \frac{\partial \mu}{\partial \theta} \\ \frac{\partial \mu}{\partial \gamma} \end{pmatrix}$$

and the FIC is defined as $\text{FIC}(S,\delta) = \widehat{\text{MSE}(S,\delta)}$.

4.4 Arbitrary submodel

For an arbitrary submodel indexed by S, Lemma 3 does not necessarily hold because we cannot guarantee that all active variables are in the chosen submodel. For the purpose of model selection, an estimator of the mean squared error of the focus is needed. We propose to use the approximations

$$\text{E}\begin{bmatrix} \sqrt{n}(\hat{\beta}_S^{\text{desp}} - \beta_0) \\ \sqrt{n}\hat{\gamma}_S^{\text{desp}} \end{bmatrix} \approx \begin{pmatrix} 0_p \\ \delta_S \end{pmatrix} + M_S \frac{1}{n\sigma^2} X_S^{*\top} X_{\gamma,S^c} \delta_{S^c}; \quad \text{Var}\begin{bmatrix} \sqrt{n}(\hat{\beta}_S^{\text{desp}} - \beta_0) \\ \sqrt{n}\hat{\gamma}_S^{\text{desp}} \end{bmatrix} \approx M_S J_S M_S^\top,$$

based on (12). This leads to the following definition of a high-dimensional FIC for a general submodel S:

$$\text{FIC}(S) = \widehat{\text{MSE}(S)}$$

with

$$\text{MSE}(S) = \begin{pmatrix} \frac{\partial \mu}{\partial \theta} \\ \frac{\partial \mu}{\partial \gamma} \end{pmatrix}^\top \left(B_S' \delta\delta^\top B_S'^t + \pi_S^{*\top} M_S J_S M_S^\top \pi_S^*\right) \begin{pmatrix} \frac{\partial \mu}{\partial \theta} \\ \frac{\partial \mu}{\partial \gamma} \end{pmatrix}$$

and

$$B_S' = \left(\pi_S^{*\top} M_S \begin{pmatrix} J_{01} \\ \pi_S J_{11} \end{pmatrix} - \begin{pmatrix} 0_{p\times q_n} \\ I_{q_n} \end{pmatrix}\right)\left(I_q - \pi_S^\top \pi_S\right).$$

This formula corresponds to (7) if $M_S = J_S^{-1}$.

Note that this corresponds to approximate the distribution of $\begin{pmatrix} \hat{\beta}_S^{\text{desp}} \\ \hat{\gamma}_S^{\text{desp}} \end{pmatrix}$ by the one of

$$\begin{pmatrix} \hat{\beta}_S^{\text{desp}} \\ \hat{\gamma}_S^{\text{desp}} \end{pmatrix} - \left(I_{p+|S|} - M_S J_S\right)\begin{pmatrix} \hat{\beta}_S^{\text{Lasso}} \\ \hat{\gamma}_S^{\text{Lasso}} \end{pmatrix} = M_S \frac{1}{n\sigma^2} X_S^{*\top} Y + \left(I_{p+|S|} - M_S J_S\right)\begin{pmatrix} \beta_0 \\ \delta/\sqrt{n} \end{pmatrix},$$

which can be seen as an *oracle desparsified estimator*.

5 Practical considerations

The results of last two sections for linear models are summarized in Table 1. We observe that the expressions giving the limiting variance of $\hat{\mu}_S$ are very similar. In the high-dimensional submodel case, J_S^{-1} is not available and is thus replaced by $M_S J_S M_S$. Regarding the bias, it is possible to show in both cases that if $S_{0,n} \subset S$ then the bias expression reduces to 0.

In practice, when computing the squared bias, we need to estimate $\delta\delta^\top$. There are several possibilities to do it but none of them produces a consistent estimator. We list here four possibilities. A first natural choice is to use the Lasso estimator $\hat{\delta}^{\text{Lasso}} = \sqrt{n}\hat{\gamma}^{\text{Lasso}}$ where

$$(\hat{\beta}^{\text{Lasso}}, \hat{\gamma}^{\text{Lasso}}) = \arg\min_{\beta,\gamma} \frac{1}{2n} \left\| Y - X^* \begin{pmatrix} \beta \\ \gamma \end{pmatrix} \right\|_2^2 + \lambda \|\beta\|_1 + \lambda \|\gamma\|_1 . \tag{14}$$

A second possibility is to use the more sophisticated adaptive Lasso $\hat{\delta}^{\text{adap}} = \sqrt{n}\hat{\gamma}^{\text{adap}}$ (see Zou, 2006) where

$$(\hat{\beta}^{\text{adap}}, \hat{\gamma}^{\text{adap}}) = \arg\min_{\beta,\gamma} \frac{1}{2n} \left\| Y - X^* \begin{pmatrix} \beta \\ \gamma \end{pmatrix} \right\|_2^2 + \lambda \sum_{j=1}^{p} w_j \beta_j + \lambda \sum_{j=p+1}^{p+q} w_j \gamma_j. \tag{15}$$

with $w_j = \begin{cases} \frac{1}{n^{-1/2}+|\hat{\beta}_j^{\text{Lasso}}|} & \text{for } 1 \le j \le p \\ \frac{1}{n^{-1/2}+|\hat{\gamma}_j^{\text{Lasso}}|} & \text{for } p+1 \le j \le p+q_n. \end{cases}$ This can provide a better estimator of δ in view of asymptotic results of Zou (2006). A third possibility is to use the desparsified estimator of the full model $\hat{\delta}^{\text{desp}} = \sqrt{n}\hat{\gamma}^{\text{desp}}$ defined as

$$\begin{pmatrix} \hat{\beta}^{\text{desp}} \\ \hat{\gamma}^{\text{desp}} \end{pmatrix} = \begin{pmatrix} \hat{\beta}^{\text{Lasso}} \\ \hat{\gamma}^{\text{Lasso}} \end{pmatrix} + M \frac{1}{n\sigma^2} X^{*\top} \left(Y - X^t \begin{pmatrix} \hat{\beta}^{\text{Lasso}} \\ \hat{\gamma}^{\text{Lasso}} \end{pmatrix} \right),$$

where M is a relaxed inverse of the Fisher information matrix $J = \frac{1}{n\sigma^2} X^{*\top} X^*$ obtained by the nodewise regression technique. The fourth possibility follows from Lemma 3 applied to $S = (1, \ldots, q_n)$. Under suitable conditions we have $\hat{\delta}^{\text{desp}} \doteq_d \mathcal{N}_q(\delta, \hat{\Omega}) + o_P(1)$ where $\hat{\Omega} = (MJM)_{-p,-p}$ is obtained by deleting the first p rows and the first p columns of MJM. Thus $\delta^{\text{desp}}\delta^{\text{desp},\top}$ has mean $\delta\delta^\top + \hat{\Omega}$. This leads to a fourth possibility for estimating $\delta\delta^\top$: to use $\hat{\delta}^{\text{desp}}\hat{\delta}^{\text{desp},t} - \hat{\Omega}$. In case this quantity would be negative, it can be truncated to zero. To summarize, we propose the four following ways to estimate $\delta\delta^\top$ in the FIC formula: (1) $\hat{\delta}^{\text{Lasso}}(\hat{\delta}^{\text{Lasso}})^\top$, (2) $\hat{\delta}^{\text{adap}}(\hat{\delta}^{\text{adap}})^\top$, (3) $\hat{\delta}^{\text{desp}}(\hat{\delta}^{\text{desp}})^\top$, (4) $\hat{\delta}^{\text{desp}}(\hat{\delta}^{\text{desp}})^\top - \hat{\Omega}$.

6 Simulation study

We perform a simulation study to illustrate the benefits of the high-dimensional FIC. We consider the linear model $Y_i = X_i\gamma_n + \sigma_\epsilon \epsilon_i$ with $\epsilon_i \sim \mathcal{N}(0,1)$ for $i = 1, \ldots, n$. We consider sample sizes $n = 100$ and

	Low-dimensional submodel	High-dimensional submodel
Estimator of $(\beta_0, \gamma_{n,S})$	least-squares estimator: $\begin{pmatrix}\hat\beta_S^{LS}\\ \hat\gamma_S^{LS}\end{pmatrix} = (X_S^{*\top}X_S^*)^{-1}X_S^{*\top}Y$	desparsified estimator: $\begin{pmatrix}\hat\beta_S^{\text{desp}}\\ \hat\gamma_S^{\text{desp}}\end{pmatrix} = \begin{pmatrix}\hat\beta_S^{\text{Lasso}}\\ \hat\gamma_S^{\text{Lasso}}\end{pmatrix} + M_S\frac{1}{n\sigma^2}X_S^{*\top}\left(Y - X_S^*\begin{pmatrix}\hat\beta_S^{\text{Lasso}}\\ \hat\gamma_S^{\text{Lasso}}\end{pmatrix}\right)$
Estimator $\hat\mu_S$ of μ_{true}	$\mu(\hat\beta_S^{LS}, \hat\gamma_S^{LS}, 0_{\mid S^c})$	$\mu(\hat\beta_S^{\text{desp}}, \hat\gamma_S^{\text{desp}}, 0_{\mid S^c})$
Bias of $\sqrt{n}\hat\mu_S$	$\left(\frac{\partial\mu}{\partial(\beta,\gamma)}\right)^\top\left(\pi_S^{*\top}J_S^{-1}\begin{pmatrix}J_{01}\\ \pi_S J_{11}\end{pmatrix} - \begin{bmatrix}0_{p\times q}\\ I_q\end{bmatrix}\right)\delta$	$\left(\frac{\partial\mu}{\partial(\beta,\gamma)}\right)^\top\left(\pi_S^{*\top}M_S\begin{pmatrix}J_{01}\\ \pi_S J_{11}\end{pmatrix} - \begin{bmatrix}0_{p\times q}\\ I_q\end{bmatrix}\right)(I_q - \pi_S^\top\pi_S)\,\delta$
Variance of $\sqrt{n}\hat\mu_S$	$\left(\frac{\partial\mu}{\partial(\beta,\gamma)}\right)^\top\pi_S^{*\top}J_S^{-1}\pi_S^*\left(\frac{\partial\mu}{\partial(\beta,\gamma)}\right)$	$\left(\frac{\partial\mu}{\partial(\beta,\gamma)}\right)^\top\pi_S^{*\top}M_S J_S M_S^\top\pi_S^*\left(\frac{\partial\mu}{\partial(\beta,\gamma)}\right)$

Table 1: Estimator $\hat\mu_S$ of the focus and its bias and variance for a low-dimensional and a high-dimensional submodel in the context of a linear model.

$n = 200$ and two different possibilities for the dimension q of the paramater γ_n: $q = 80$ and $q = 200$. The case $q = 200$ corresponds to high-dimensional data for which the classical FIC can not be used. We generate the true model according to four scenarios:

- Case 1: $\gamma_n = 10c(1, -1, 1, -1, 1, 0, \ldots, 0)/\sqrt{n}$ and X_i from $\mathcal{N}_q(0, I_q)$ for $i = 1, \ldots, n$.
- Case 2: γ_n as in case 1 and X_i from $\mathcal{N}_q(0, \Sigma)$ for $i = 1, \ldots, n$ with $\Sigma_{jj} = 1$ and $\Sigma_{jk} = 0.5$ for $j \neq k$.
- Case 3: $\gamma_n = 10c(1, -\frac{1}{2}, \frac{1}{3}, -\frac{1}{4}, \ldots, \pm\frac{1}{q})/\sqrt{n}$ and X_i as in case 1
- Case 4: γ_n as in case 3 and X_i as in case 2.

Cases 1 and 2 correspond to sparse models and cases 2 and 4 correspond to models with correlation between variables. The parameter c controls the amplitude of the components of γ_n. We consider three different focuses. The first focus is the prediction $\mu_1(\gamma_n) = X_0\gamma_n$ for a new value X_0 of the covariate vector with the components of X_0 randomly generated from $\mathbb{U}[-1, 1]$. The second focus is the first coefficient of γ_n, that is $\mu_2(\gamma_n) = \gamma_{n,1}$ and the third focus is the last coefficient of γ_n, that is $\mu_3(\gamma_n) = \gamma_{n,q}$. Note that the true value of the last focus is 0 for the sparse settings (cases 1 and 2).

We compare predictions of the focus μ_j for two types of methods: (i) we compute a penalized estimator of γ_n in the full model and make prediction based on this parameter estimate and (ii) we use the high-dimensional FIC as described in Sections 3 and 4. We consider two penalized estimators, the Lasso and the adaptive Lasso, with the tuning paramaters chosen by 10-fold cross-validation. Other tuning parameter choices are possible too. These two penalized estimators are also used to estimate δ in the FIC procedure. We thus obtain four different predictions of μ_j. For the estimation of σ_ϵ^2, we follow the recommendation of Reid et al. (2016) and use $\hat\sigma_\epsilon^2 = \text{RSS}/(n - \widehat{\text{df}})$ with $\widehat{\text{df}}$ the number of non-zero coefficients of the penalized estimator of γ_n.

Because the number of covariates is large, it is computationally impossible to obtain the FIC of every possible submodel. Instead, we propose to use a backward-forward stepwise procedure with two possible starting sets: the empty set and the set $\{j : \hat\delta_j \neq 0\}$ of active components of the estimator of δ. The

two procedures usually converge to two different subsets S_1 and S_2 and we keep the one that gives the smallest FIC value. More refined procedures can be used to improve the selection search. It is for example possible to do some pseudo-exhaustive search by computing the FIC of all submodels upto a certain size d.

In Tables 2 to 4, we report the averaged squared errors of the estimators for the three different focuses over 1000 simulated datasets for different settings. In Table 2 we consider settings with 80 covariates and in Table 3 we increase the number of covariates to 200, obtaining high-dimensional data for which the traditional FIC could not be used. Results for these two tables are similar. We observe that all methods perform well for the third focus $\mu_3 = \gamma_q$. Regarding focuses 1 and 2 we observe that the FIC procedures outperform the penalized estimators for the sparse settings (cases 1 and 2), the ones that are supported by the theory. For non-sparse settings (cases 3 and 4), the different methods are equally competitive. The presence of correlation makes things slightly more complicated.

In Table 4, we compare the sensitivity of the different methods to the standard noise level σ_ϵ. We observe that in the sparse cases, the FIC takes much more advantage of the decrease of the noise level. For $\sigma_\epsilon = 0.25$ the FIC largely outperforms the penalized methods while for $\sigma_\epsilon = 1$ the methods are equally competitive.

We conclude this simulation study by a remark on the size of the models selected by the FIC. We observed in our simulations that the models selected by the FIC procedure are very often of size smaller than 5. It turns out that it is often possible to find a small submodel S whose FIC is smaller than the FIC of S_{true}, the active set of the model having generated the data. On Figure 1, we illustrate this by giving the scatter plot of FIC(S) versus $\hat{\mu}_S$ for every possible submodel of size smaller or equal to 3. The setting is chosen to have many true non-zero coefficients (20) so that we expect the bias to be large for models of size only 3. We also choose a small value of the standard noise ($\sigma_\epsilon = 0.1$) to increase the weight of the squared bias in the FIC expression. We see on the left figure that many of the submodels exhibit large values of FIC but more importantly we also notice on the right figure that for some of the small models (about 3% of them), the FIC value is smaller than the FIC of the true model. For such submodels, the estimator $\hat{\mu}_S$ is very close to the true value (the grey horizontal line). This should be considered one of the strong features of the FIC.

7 Real data example: the riboflavin data

We apply the high-dimensional FIC procedure on the riboflavin data that can be found in the R package *hdi* (Meier et al., 2014). The data contains 71 observations, 4088 predictors (gene expressions) and a response variable measuring the riboflavin production of the Bacillus subtilis bacteria. This dataset has been used by many authors in the high-dimensional literature including van de Geer et al. (2014) and Javanmard and Montanari (2014). We center the response variable and randomly split the data into a training set $(X_{\text{train}}, Y_{\text{train}})$ of size 50 and a testing set $(X_{\text{test}}, Y_{\text{test}})$ of size 21. We then consider the linear model $Y_{\text{train}} = X_{\text{train}}\beta + \epsilon$ with $\epsilon \sim N(0, \sigma_\epsilon^2)$ and the 21 focuses $\mu_j = X_{\text{test}}^j\beta$ for $j = 1, \ldots, 21$.

	$n = 100$, $q = 80$						$n = 200$, $q = 80$					
	$c = 1$			$c = 2$			$c = 1$			$c = 2$		
Focus:	μ_1	μ_2	μ_3	μ_1	μ_2	μ_3	μ_1	μ_2	μ_3	μ_1	μ_2	μ_3
Case 1: sparsity and no correlation												
Lasso	0.023	0.016	0.000	0.125	0.087	0.000	0.003	0.002	0.000	0.021	0.015	0.000
Adap. Lasso	0.022	0.015	0.000	0.132	0.088	0.000	0.002	0.001	0.000	0.020	0.015	0.000
FIC Lasso	0.003	0.001	0.000	0.006	0.001	0.000	0.001	0.000	0.000	0.001	0.000	0.000
FIC Adap. Lasso	0.003	0.001	0.000	0.009	0.001	0.000	0.001	0.000	0.000	0.001	0.000	0.000
Case 2: sparsity and correlation												
Lasso	0.026	0.013	0.000	0.150	0.074	0.000	0.005	0.002	0.000	0.019	0.010	0.000
Adap. Lasso	0.021	0.010	0.000	0.150	0.072	0.000	0.002	0.001	0.000	0.017	0.010	0.000
FIC Lasso	0.008	0.003	0.000	0.028	0.013	0.000	0.003	0.001	0.000	0.004	0.002	0.000
FIC Adap. Lasso	0.007	0.003	0.000	0.030	0.017	0.000	0.002	0.001	0.000	0.003	0.002	0.000
Case 3: no sparsity and no correlation												
Lasso	0.028	0.003	0.000	0.091	0.010	0.001	0.009	0.001	0.000	0.017	0.001	0.000
Adap. Lasso	0.028	0.002	0.000	0.089	0.003	0.001	0.009	0.000	0.000	0.017	0.001	0.000
FIC Lasso	0.026	0.002	0.001	0.080	0.004	0.001	0.009	0.000	0.000	0.016	0.001	0.000
FIC Adap. Lasso	0.028	0.002	0.000	0.088	0.003	0.001	0.010	0.000	0.000	0.017	0.001	0.000
Case 4: no sparsity and correlation												
Lasso	0.038	0.005	0.001	0.114	0.013	0.001	0.015	0.001	0.000	0.024	0.001	0.001
Adap. Lasso	0.039	0.003	0.001	0.112	0.005	0.001	0.015	0.001	0.000	0.025	0.001	0.000
FIC Lasso	0.037	0.003	0.001	0.105	0.005	0.001	0.015	0.001	0.000	0.024	0.001	0.001
FIC Adap. Lasso	0.039	0.002	0.001	0.111	0.005	0.001	0.015	0.001	0.000	0.025	0.001	0.001

Table 2: Averaged squared errors of the estimators for the three different focuses over 1000 simulated datasets and for different settings. μ_1 is a random new observation, $\mu_2 = \gamma_1$ and $\mu_3 = \gamma_q$. The number of covariates is $q = 80$, the standard noise is $\sigma_\epsilon = 0.25$ and c is a parameter controlling the amplitude of the components of γ_n.

	$n = 100$, $q = 200$						$n = 200$, $q = 200$					
	$c = 1$			$c = 2$			$c = 1$			$c = 2$		
Focus:	μ_1	μ_2	μ_3	μ_1	μ_2	μ_3	μ_1	μ_2	μ_3	μ_1	μ_2	μ_3
Case 1: sparsity and no correlation												
Lasso	0.026	0.016	0.000	0.141	0.088	0.000	0.004	0.002	0.000	0.022	0.015	0.000
Adap. Lasso	0.024	0.016	0.000	0.141	0.091	0.000	0.002	0.001	0.000	0.022	0.015	0.000
FIC Lasso	0.003	0.001	0.000	0.006	0.001	0.000	0.001	0.000	0.000	0.001	0.000	0.000
FIC Adap. Lasso	0.003	0.001	0.000	0.010	0.001	0.000	0.001	0.000	0.000	0.001	0.000	0.000
Case 2: sparsity and correlation												
Lasso	0.031	0.014	0.000	0.161	0.074	0.000	0.006	0.003	0.000	0.021	0.011	0.000
Adap. Lasso	0.023	0.011	0.000	0.159	0.074	0.000	0.002	0.001	0.000	0.020	0.010	0.000
FIC Lasso	0.009	0.004	0.000	0.027	0.014	0.000	0.003	0.001	0.000	0.004	0.002	0.000
FIC Adap. Lasso	0.006	0.003	0.000	0.033	0.016	0.000	0.001	0.001	0.000	0.003	0.002	0.000
Case 3: no sparsity and no correlation												
Lasso	0.050	0.006	0.000	0.150	0.018	0.000	0.017	0.001	0.000	0.035	0.002	0.000
Adap. Lasso	0.048	0.003	0.000	0.193	0.008	0.000	0.020	0.001	0.000	0.086	0.002	0.000
FIC Lasso	0.047	0.003	0.000	0.137	0.008	0.000	0.016	0.001	0.000	0.034	0.001	0.000
FIC Adap. Lasso	0.048	0.002	0.000	0.188	0.007	0.000	0.020	0.001	0.000	0.084	0.002	0.000
Case 4: no sparsity and correlation												
Lasso	0.065	0.010	0.000	0.176	0.023	0.000	0.024	0.003	0.000	0.047	0.003	0.000
Adap. Lasso	0.063	0.004	0.000	0.252	0.010	0.000	0.028	0.001	0.000	0.115	0.003	0.000
FIC Lasso	0.062	0.005	0.000	0.162	0.011	0.000	0.023	0.001	0.000	0.046	0.002	0.000
FIC Adap. Lasso	0.062	0.004	0.000	0.242	0.010	0.000	0.028	0.001	0.000	0.111	0.003	0.000

Table 3: Averaged squared errors of the estimators for the three different focuses over 1000 simulated datasets and for different settings. μ_1 is a random new observation, $\mu_2 = \gamma_1$ and $\mu_3 = \gamma_q$. The number of covariates is $q = 200$, the standard noise is $\sigma_\epsilon = 0.25$ and c is a parameter controlling the amplitude of the components of γ_n.

Focus:	μ_1			μ_2			μ_3		
Standard noise: σ_ϵ	1	0.5	0.25	1	0.5	0.25	1	0.5	0.25
Case 1: sparsity and no correlation									
Lasso	0.086	0.024	0.023	0.036	0.013	0.016	0.001	0.000	0.000
Adap. Lasso	0.060	0.015	0.022	0.014	0.007	0.015	0.001	0.000	0.000
FIC Lasso	0.063	0.013	0.003	0.013	0.003	0.001	0.002	0.000	0.000
FIC Adap. Lasso	0.063	0.011	0.003	0.012	0.003	0.001	0.002	0.000	0.000
Case 2: sparsity and correlation									
Lasso	0.166	0.042	0.026	0.066	0.018	0.013	0.002	0.000	0.000
Adap. Lasso	0.135	0.024	0.021	0.029	0.007	0.010	0.003	0.000	0.000
FIC Lasso	0.140	0.029	0.008	0.039	0.009	0.003	0.003	0.001	0.000
FIC Adap. Lasso	0.142	0.024	0.007	0.027	0.006	0.003	0.004	0.001	0.000
Case 3: no sparsity and no correlation									
Lasso	0.125	0.056	0.028	0.040	0.009	0.003	0.001	0.001	0.000
Adap. Lasso	0.137	0.057	0.028	0.016	0.004	0.002	0.002	0.001	0.000
FIC Lasso	0.126	0.054	0.026	0.016	0.005	0.002	0.002	0.001	0.001
FIC Adap. Lasso	0.149	0.059	0.028	0.015	0.004	0.002	0.003	0.001	0.000
Case 4: no sparsity and correlation									
Lasso	0.188	0.086	0.038	0.084	0.018	0.005	0.002	0.001	0.001
Adap. Lasso	0.204	0.088	0.039	0.035	0.008	0.003	0.003	0.001	0.001
FIC Lasso	0.195	0.085	0.037	0.045	0.011	0.003	0.003	0.001	0.001
FIC Adap. Lasso	0.223	0.091	0.039	0.034	0.008	0.002	0.005	0.002	0.001

Table 4: Averaged squared errors of the estimators for the three different focuses over 1000 simulated datasets and for different settings. μ_1 is a random new observation, $\mu_2 = \gamma_1$ and $\mu_3 = \gamma_q$. Parameters are $q = 80$, $n = 100$ and $c = 1$. Results are given for different values of the standard noise: $\sigma_\epsilon = 1, 0.5, 0.25$.

In a first step we compute a Lasso estimator $\hat{\beta}^{\text{Lasso}}$ of β with the tuning paramaters chosen by 10-fold cross-validation and we obtain estimators $\hat{\mu}_j^{\text{Lasso}} = X_{\text{test}}^j \hat{\beta}^{\text{Lasso}}$ of the 21 focuses. In a second step we apply our FIC procedure. For each of the 21 focuses, we search for a submodel that provides a small FIC value. As in the simulation study, we apply a backward-forward stepwise procedure with two possible starting sets: the empty set and the set selected by the Lasso. We denote by S_1^j and S_2^j the sets obtained for the focus j with these two choices of starting sets. We then keep the best of the two by defining $S^j = \arg\min_{S \in \{S_1^j, S_2^j\}} \text{FIC}_j(S)$. We compute the corresponding estimator $\hat{\beta}_{S^j}$ and obtain an estimator $\hat{\mu}_j^{\text{FIC}} = X_{\text{test}}^j \hat{\beta}_{S^j}$ of the focus μ_j. We then compute the mean squared prediction errors $1/21 \sum_{j=1}^{21} (Y_{\text{test}}^j - \hat{\mu}_j)^2$ for $\hat{\mu}_j = \hat{\mu}_j^{\text{Lasso}}$ and $\hat{\mu}_j = \hat{\mu}_j^{\text{FIC}}$. For comparison purpose, we also compute estimators of the focuses with S_1^j and S_2^j. Note that each computation of the FIC takes about one millisecond on a regular computer. Thus, each step of the stepwise procedure takes about four seconds.

The results are reported in Table 5. We observe that the three strategies for performing the FIC optimization are very competitive and all outperform the Lasso (0.180, 0.177 and 0.182 versus 0.235 for

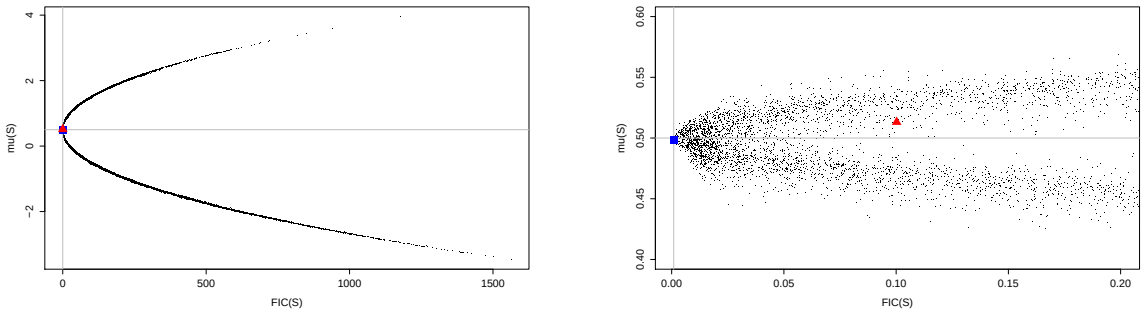
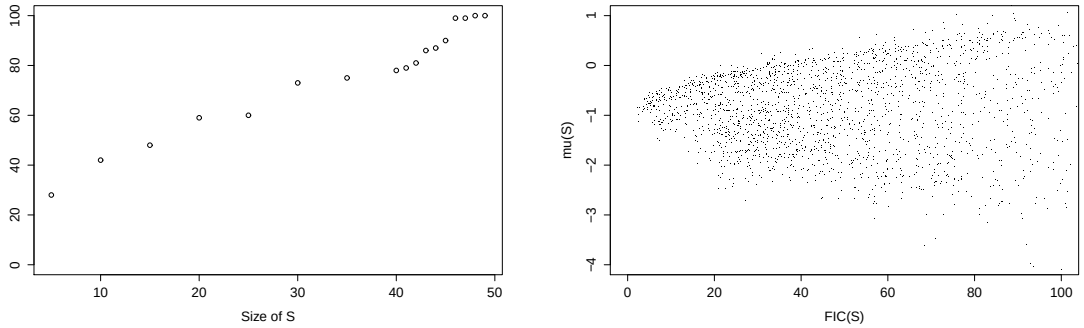


Figure 1: Parameters: $n = 100$, $q = 80$, $p = 0$, $s_0 = 20$, $\sigma_\epsilon = 0.1$, $c = 1$, μ_1. Scatterplots of $\hat{\mu}(S)$ versus FIC(S) for all the 85401 possible models of size smaller or equal to 3. The true δ and the true σ_ϵ are used in the FIC computations. The red triangle corresponds to the true model of size 20 and the blue square corresponds to the model minimizing the FIC amongst the models of size smaller than 3. The right figure is a zoom of the left figure.

the Lasso). We also observe that for two third of the focuses (14 out of 21), the set S_1 was chosen, corresponding to work with the empty set as starting set. In Table 6, we report information about the variables selected by the different procedures. As expected, the set S_1^j is generally smaller (4.7) than the set S_2^j (10.7). Furthermore, we note that only two variables are selected at least three times by FIC 1 and none of them is also selected by the Lasso. Conversely, all the 10 variables selected at least three times by FIC 2 are also selected by the Lasso. To conclude, we observe that the FIC uses for each prediction much fewer variables than the Lasso (6.7 versus 27) but in total the number of different variables used by the FIC for the 21 predictions is much larger than for the Lasso (120 versus 27). This is a key feature of the FIC.



(a) (b)

Figure 2: (a) For different subset sizes, the number of times that the desparsified FIC of Section 4 is smaller than the OLS FIC of Section 3 for 100 random subsets and for the first focus. (b) Scatterplot of $\hat{\mu}(S)$ versus FIC(S) for the desparsified FIC of Section 4 for models going from size 5 to 100.

We end this section by numerically comparing the behaviour of the low-dimensional FIC (OLS) introduced in Section 3 to the high-dimensional FIC (desparsified) presented in Section 4. Summary of the formulas can be found back in Table 1. We refer to these two FIC as OLS FIC and desparsified FIC. We consider the first focus $\mu_1 = X^1_{\text{test}}\beta$ of the real data example and compute the two FIC values for different subsets. Note that these two FIC aim at estimating the mean squared error of two different estimators (see Table 1). More concretely, we consider subsets of the 4088 covariates of size going from 5 to 49. Recall that the training sample size is 50. For each of the considered sizes we consider 100 different subsets.

In Table 7 we give the two FIC values and the two predictions for the first considered subset of each size. Note that the true value of the focus is unknown but we can expect it to be close to $y^1_{\text{test}} = -1.13$. We see that for small subsets the results are very close to each other. When $|S|$ increases, M_S gets further away from J_S^{-1} so that the estimators become more and more different. When $|S|$ is close to the sample size 50 the quality of the OLS estimator deteriorates and it is better to use the desparsified estimator. In Figure 2(a) we count how many times out of 100 random choices of the subsets the desparsified FIC is smaller than the OLS FIC. We observe that from size 20 the desparsified FIC tends to outperfom the OLS FIC. This gets more and more pronounced when $|S|$ gets closer to the sample size. In any case for each S we can always compute both FIC and keep the smaller one. Recall that they refer to different estimators. In Figure 2(b), we give the scatterplot of FIC(S) versus $\hat{\mu}_S$ for 2300 submodels of size $5, 10, \ldots, 40, 41, \ldots, 49, 50, 60, \ldots, 100$ (100 submodels of each size). We observe that the FIC obtained with the desparsified procedure behaves as it is supposed to: the FIC aims at estimating the expected value of $50(\hat{\mu}_S - \mu_{\text{true}})^2$ and we do observe a quadratic shape, which is slightly altered due to the difficulty to estimate δ in this high-dimensional example.

8 Extensions and discussion

8.1 Focused selection for high-dimensional generalized linear models

The results of Section 4 can be extended to high-dimensional generalized linear models (GLM). Let us consider observations $Y_1, \ldots, Y_n$ where Y_i has density $f(y, X_i, \theta_0, \gamma_0 + \delta/\sqrt{n})$ for $i = 1, \ldots, n$ with f from the exponential family of distributions. Consider a high-dimensional submodel S containing the true active set $S_{0,n}$ Let us write $\beta_S = \begin{pmatrix} \theta \\ \gamma_S \end{pmatrix}$ and denote the loss function for an observation (y, x) by $\rho_{\beta_S}(y, x) = -\log f(y, x, \theta, \gamma_S, \gamma_{0,S^c})$. We define the first an second partial derivatives of the loss function as $\dot{\rho}_{\beta_S} = \frac{\partial}{\partial \beta_S}\rho_{\beta_S}$ and $\ddot{\rho}_{\beta_S} = \frac{\partial^2}{\partial \beta_S \partial \beta_S^\top}\rho_{\beta_S}$ and use the following notation: for a function g we write $P_n g = \frac{1}{n}\sum_{i=1}^n g(Y_i, X_i)$. We use the penalized estimator

$$\begin{pmatrix} \hat{\theta}^L_S \\ \hat{\gamma}^L_S \end{pmatrix} = \hat{\beta}^L_S = \underset{\beta_S = (\theta, \gamma_S)}{\arg\min}\ P_n \rho_{\beta_S} + \lambda \left\|\beta_S\right\|_1 .$$

Similarly to van de Geer et al. (2014), we define $\hat{\Sigma}_S = P_n \ddot{\rho}_{\hat{\beta}^L_S}$ and $\hat{M}_S$ as a relaxed inverse of $\hat{\Sigma}_S$ obtained by the Lasso nodewise regression. Note that $\hat{\Sigma}_S$ corresponds to the empirical Fisher matrix estimated in

$(\hat{\theta}_S^L, \hat{\gamma}_S^L, \gamma_{0,S^c})$. We define the following desparsified estimator

$$\begin{pmatrix} \hat{\theta}_S^{\text{desp}} \\ \hat{\gamma}_S^{\text{desp}} \end{pmatrix} = \begin{pmatrix} \hat{\theta}_S^L \\ \hat{\gamma}_S^L \end{pmatrix} - \hat{M}_S P_n \dot{\rho}_{\hat{\theta}_S^L, \hat{\gamma}_S^L}. \tag{16}$$

Using results from van de Geer et al. (2014), we can show that

$$\begin{pmatrix} \hat{\theta}_S^{\text{desp}} - \theta_0 \\ \hat{\gamma}_S^{\text{desp}} - \gamma_{0,S} \end{pmatrix} = \begin{pmatrix} 0 \\ \delta_S \end{pmatrix} - \hat{M}_S P_n \dot{\rho}_{\hat{\beta}_S^{\text{true}}} + o_P(n^{-1/2}) \tag{17}$$

and that $\hat{M}_S P_n \dot{\rho}_{\hat{\beta}_S^L} \dot{\rho}_{\hat{\beta}_S^L}^\top \hat{M}_S^\top$ is a consistent estimator of the variance of $\sqrt{n}(\hat{\theta}_S^{\text{desp},t} - \theta_0^\top, \hat{\gamma}_S^{\text{desp},t} - \gamma_{0,S}^\top)^\top$. This leads to a result similar to Theorem 2 where $M_S J_S M_S$ is replaced by $\hat{M}_S P_n \dot{\rho}_{\hat{\beta}_S^L} \dot{\rho}_{\hat{\beta}_S^L}^\top \hat{M}_S^\top$.

8.2 Model averaging in high-dimensional models

Averaging estimators across several good models is another interesting route, also in the high-dimensional setting. Since estimators in a selected model can be written as model averaged estimators assigning weight one to the estimator in the selected model, and weight zero to all other models, the tool of model averaging is important to study proper post-selection inference.

Let the weighted estimator be obtained in the following way

$$\hat{\mu}_{\text{avg}} = \sum_{S \in \mathcal{A}} w_S(\hat{\delta}) \hat{\mu}_S,$$

where $\hat{\mu}_S = \mu(\hat{\theta}_S, \hat{\gamma}_S, \gamma_{0,S^c})$ and $\mathcal{A}$ is the set of models under consideration for averaging, this does not need to be the set of all possible submodels of the largest available model. The weight for each S, may be deterministic, e.g., assigning equal weight to each of the models, or a predetermined weight that is not data-dependent, or the weights may be data-driven, e.g., $w_S(\hat{\delta}) = I\{S = \arg\min_{S' \in \mathcal{A}} \text{FIC}(S', \hat{\delta})\}$ in the FIC selection case.

Using Lemma 3, or equation (17) in the GLM case, we obtain that the desparsified estimator $\hat{\delta}^{\text{desp}} \doteq_d \tilde{\delta} \sim \mathcal{N}_q(\delta, \Omega)$. Using the joint convergence of the random weights $w_S(\hat{\delta})$ and $\sqrt{n}(\hat{\mu}_S - \mu_{\text{true}})$ for $S \in \mathcal{A}$ to their respective limits, we obtain the following Corollary to Theorem 2.

Corollary 1. *Assume the local misspecification setting. Under the assumptions of Theorem 2, for a set of weights such that* $\sum_{S \in \mathcal{A}} w_S(d) = 1$ *for all* d *and with at most a countable number of discontinuities,*

$$\sqrt{n}\left\{\sum_{S \in \mathcal{A}} w_S(\hat{\delta}) \hat{\mu}_S - \mu_{\text{true}}\right\} \to_d \Lambda_{\text{avg}} = \sum_{S \in \mathcal{A}} w_S(\tilde{\delta}) \Lambda_S.$$

The limiting variable is in case of deterministic weights again normal. For random weights the limit distribution is a sum of products of the random weights and the normal limits Λ_S, which is in general not longer normally distributed. The mean of the limit random variable depends on the random weights $w_S(\tilde{\delta})$ and its correlation with the random variables C_S, D_S,

$$E(\Lambda_{\text{avg}}) = -\left(\frac{\partial \mu}{\partial \gamma}\right)^\top \delta + \sum_{S \in \mathcal{A}} E\left[w_S(\tilde{\delta})\left\{\left(\frac{\partial \mu}{\partial \theta}\right)^\top C_S + \left(\frac{\partial \mu}{\partial \gamma_S}\right)^\top D_S\right\}\right].$$

See also Claeskens et al. (2016) for a calculation of the first two moments of weighted forecasts under weak assumptions avoiding normality.

The use of the proper limiting distribution for model averaged estimators and for post-selection estimators in general, to obtain inference on the averaged estimator is a worthwhile topic of further research, as is the determination of the choice of the weights in the high-dimensional setting.

8.3 Other extensions

Since in high-dimensional models the desparsification is crucial to work with the asymptotic distribution of the focus estimator, which may depend on parameters that are finally set to zero as well as on parameters that are estimated non-zero, extensions to other types of models require first a study of the asymptotic distribution of a desparsified estimator. Gueuning and Claeskens (2016) obtained such a result for partially linear single-index models of the form $Y = \eta(z^\top \alpha) + x^\top \beta + \varepsilon$. In such model both model parts may contain high-dimensional variables, though for theoretical reasons only the dimension of the vector x is allowed to grow with the sample size. An extension of the FIC to such models with focuses that depend on the model parameters α, β, will go along the same lines. In the setting of weighted composite quantile estimation, this would be a generalization of the work on focused selection and model averaging by Xu et al. (2014) to the high-dimensional case.

The present paper also paves the way for extensions of some variations of the FIC to the high-dimensional framework. For example, the *weighted FIC* introduced by Claeskens and Hjort (2008a) aims at selecting a model that performs well for handling a range of similar tasks. Rather than minimizing a mean squared error we could consider selecting a model that minimizes another expected loss function, such as the expected value of a weighted version of the squared error loss,

$$\int n\{\hat{\mu}_S(x) - \mu_{\text{true}}(x)\}^2 d\nu(x),$$

where the dependence of the focus on, say, a covariate vector x is explicitly introduced in the notation. The choice of the weight function ν allows to specify a domain in the covariate space for which a good estimator of μ_{true} is sought. Another use could be to downweight certain regions in order to obtain a more outlier resistant estimator. An estimator of the limit version of this expected weighted loss version leads to the average-focused information criterion for the high-dimensional setting.

To conclude, this paper is the first one to obtain the focused information criterion for high-dimensional data where the parameter length is allowed to grow and even exceed the sample size. Due to the use of a desparsified estimator, the criterion is able to deal with high-dimensional submodels. In addition, we have obtained an alternative formula for FIC in the low-dimensional case that not only deals with the high-dimensionality of the model, but that also is of interest in low-dimensional models by its avoidance to invert the information matrix in a largest model. This paper may pave the way for other applications, estimation methods and models where there is a high-dimensional parameter and where focused selection could bring its benefits of better, targeted, estimators.

9 Proofs

9.1 Proofs for Theorem 1

The proof requires adjustments of the proofs of Lemmas 1, 2 and 3 of Hjort and Claeskens (2003).

Proof of Lemma 1. We use the univariate Lindeberg-Feller theorem (see Serfling (1980) section 1.9) for independant but not i.i.d. random variables and the Cramér-Wold device to obtain the appropriate multivariate normality. To ease the notations, we define the following $p+|S|-$dimensional vectors:

$$\begin{aligned}
W_i &= \begin{pmatrix} U(y_i|x_i) \\ V_S(y_i|x_i) \end{pmatrix} \text{ for } i = 1, \ldots, n, \\
\bar{W}_n &= \frac{1}{n}\sum_{i=1}^{n} W_i, \\
T_i &= W_i - \begin{pmatrix} J_{01}(x_i)\delta/\sqrt{n} \\ \pi_S J_{11}(x_i)\delta/\sqrt{n} \end{pmatrix} \text{ for } i = 1, \ldots, n, \\
\bar{T}_n &= \frac{1}{n}\sum_{i=1}^{n} T_i = \bar{W}_n - \begin{pmatrix} J_{n,01}\delta/\sqrt{n} \\ \pi_S J_{n,11}\delta/\sqrt{n} \end{pmatrix}.
\end{aligned}$$

We want to prove that

$$\sqrt{n}\bar{T}_n \xrightarrow{d} T = \mathcal{N}_{p+|S|}(0, J_S). \tag{18}$$

Let $j \in \{1, \ldots, p+|S|\}$ and let us show first that it holds that

$$\sqrt{n}\bar{T}_n^j \xrightarrow{d} T^j = N(0, (J_S)_{j,j}) \tag{19}$$

with $\bar{T}_n^j$ the j-th component of $\bar{T}_n$.

We have

$$\begin{aligned}
\mathrm{E}[W_i^j] &= \int W_i^j f_0(y|x_i) \left\{1 + V(y|x_i)^\top \delta/\sqrt{n} + R(y|x_i, \delta/\sqrt{n})\right\} dy \\
&= 0 + e_j^\top \begin{pmatrix} J_{01}(x_i)\delta/\sqrt{n} \\ \pi_S J_{11}(x_i)\delta/\sqrt{n} \end{pmatrix} + \int W_i^j f_0(y|x_i) R(y|x_i, \delta/\sqrt{n}) dy.
\end{aligned}$$

By conditions (C1) and (C5), the last term is $o(n^{-1/2})$ so that $\mathrm{E}[T_i^j] = o(n^{-1/2})$.

Furthermore

$$\begin{aligned}
\mathrm{E}[(W_i^j)^2] &= \int (W_i^j)^2 f_0(y|x_i) \left\{1 + V(y|x_i)^\top \delta/\sqrt{n} + R(y|x_i, \delta/\sqrt{n})\right\} dy \\
&= J_S(x_i)_{j,j} + \int (W_i^j)^2 f_0(y|x_i) V(y|x_i)^\top \delta_n/\sqrt{n}\, dy + \int (W_i^j)^2 f_0(y|x_i) R(y|x_i, \delta/\sqrt{n})\, dy \\
&= J_S(x_i)_{j,j} + o(1)
\end{aligned}$$

where the last equality comes from (C2), (C3) and (C5). This implies that $\mathrm{Var}(T_i^j) = J_S(x_i)_{j,j} + o(1)$. Now, applying the Lindeberg-Feller theorem to $\{T_i^j\}_i$ we have

$$\frac{\bar{T}_n^j - \frac{1}{n}\sum_{i=1}^n \mathrm{E}[T_i^j]}{\sqrt{\frac{1}{n^2}\sum_{i=1}^n \mathrm{Var}(T_i^j)}} \xrightarrow{d} \mathcal{N}(0,1)$$

which is equivalent to

$$\frac{\sqrt{n}\bar{T}_n^j}{\sqrt{(J_{n,S})_{j,j}}} \xrightarrow{d} \mathcal{N}(0,1)$$

and because $(J_{n,S})_{j,j}$ converges to $(J_S)_{j,j}$ this is also equivalent to (19). The Lindeberg condition for applying the Lindeberg-Feller theorem requires

$$\frac{1}{n}\sum_{i=1}^n \frac{1}{B_n^2/n} \int_{|t-E(T_i^j)|>\epsilon B_n} \left\{t - E(T_i^j)\right\}^2 dF_i(t) \to 0 \text{ for each } \epsilon > 0$$

with $B_n^2 = \sum_{i=1}^n \mathrm{Var}(T_i^j)$ and $F_i(t)$ the distribution function of T_i^j. This is satisfied if

$$\int_{|T_i^j|>\epsilon\sqrt{n}} (T_i^j)^2 f_{\text{true}}(y|x_i)dy \to 0 \text{ for each } \epsilon > 0$$

which holds thanks to conditions (C2) and (C3). Thus, (19) is proven.

We now apply the Cramér-Wold device that states that

$$\sqrt{n}\,\bar{T}_n \xrightarrow{d} T \quad \text{if and only if} \quad \sqrt{n}\,a^\top \bar{T}_n \xrightarrow{d} a^\top T \;\; \forall a \in \mathbb{R}^{p+|S|}.$$

Let consider an arbitrary $a \in \mathbb{R}^{p+|S|}$. Using (19), it is clear that $\sqrt{n}a^\top\bar{T}_n = \sum_{j=1}^{p+|S|}\sqrt{n}a_j\bar{T}_n^j$ tends to a normal distribution with mean 0 and variance given by $\sum_{j,k=1}^{p+|S|} a_j a_k (J_S)_{j,k} = a^\top J_S a$. Indeed,

$$\begin{aligned}\mathrm{Var}(\sqrt{n}a^\top\bar{T}_n) &= \sum_{j,k=1}^{p+|S|} n\, a_j a_k \mathrm{Cov}(\bar{T}_n^j, \bar{T}_n^k) \\ &= \sum_{j,k=1}^{p+|S|} \frac{1}{n}\sum_{i=1}^n a_j a_k \mathrm{Cov}(T_i^j, T_i^k) = \sum_{j,k=1}^{p+|S|} a_j a_k (J_{n,S})_{j,k} = a^\top (J_{n,S}) a\end{aligned}$$

which tends to $a^\top J_S a$. This implies that $\sqrt{n}\,a^\top\bar{T}_n \xrightarrow{d} a^\top T$ so that (18) holds. □

Proof of Lemma 2. By Lemma 1, it suffices to show that

$$\begin{pmatrix} \sqrt{n}(\hat{\theta}_S - \theta_0) \\ \sqrt{n}(\hat{\gamma}_S - \gamma_{0,S}) \end{pmatrix} \doteq_d J_S^{-1} \begin{pmatrix} \sqrt{n}\bar{U}_n \\ \sqrt{n}\bar{V}_{n,S} \end{pmatrix},$$

which can be done by using traditional arguments for maximum likelihood estimators (see for example Serfling (1980) section 4.2.2). We give here the explicit derivations.

Writing $\hat{\beta} = (\hat{\theta}_S, \hat{\gamma}_S)$ and $\beta_0 = (\theta_0, \gamma_{0,S})$, a Taylor expansion of $\frac{1}{n}\sum_{i=1}^n \frac{\partial}{\partial\beta}\log f(y_i|x_i,\hat{\beta})$ around β_0 gives

$$0 = \begin{pmatrix} \bar{U}_n \\ \bar{V}_{n,S} \end{pmatrix} + \frac{1}{n}\sum_{i=1}^n \frac{\partial^2}{\partial\beta\partial\beta^\top}\log f(y_i|x_i,\beta_0)(\hat{\beta}-\beta_0) + \frac{1}{2}\sum_{j=1}^{p+|S|}(\hat{\beta}-\beta_0)^\top \frac{1}{n}\sum_{i=1}^n \frac{\partial^3}{\partial\beta_j\partial\beta\partial\beta^\top}\log f(y_i|x_i,\tilde{\beta})(\hat{\beta}-\beta_0) \tag{20}$$

with $\tilde{\beta}$ between β_0 and $\hat{\beta}$. By condition (C4), there exists a function $H(x)$ with finite mean such that $\left|\frac{\partial^3}{\partial\beta_j\partial\beta_k\partial\beta_l}\log f(y|x,\beta)\right| \leq H(x)$ for each $1 \leq j,k,l \leq p+|S|$ in a neighbourhood of β_0. Furthermore, because $\tilde{\beta}$ lies between β_0 and $\hat{\beta}$, we can write $\frac{\partial^3}{\partial\beta_j\partial\beta_k\partial\beta_l}\log f(y|x,\beta) = H(x_i)\xi^i_{jkl}$ with $|\xi^i_{jkl}|\leq 1$. Denoting by ξ^i_j the $(p+|S|)\times(p+|S|)$ matrix whose element (k,l) is ξ^i_{jkl}, the last term of (20) can be expressed as

$$\frac{1}{2}(\hat{\beta}-\beta_0)^\top \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{p+|S|} H(x_i)\xi^i_j(\hat{\beta}-\beta_0).$$

We now define $C_n = \frac{1}{n}\sum_{i=1}^n H(x_i)$ and $\xi^* = \frac{1}{C_n}\frac{1}{n}\sum_{i=1}^n\sum_{j=1}^{p+|S|} H(x_i)\xi^i_j$ and see that this last term is also equal to $\frac{1}{2}(\hat{\beta}-\beta_0)^\top C_n\xi^*(\hat{\beta}-\beta_0)$. Note that each component of the matrix ξ^* is smaller than $p+|S|$ in absolute value. Defining $B_n = \frac{1}{n}\sum_{i=1}^n \frac{\partial^2}{\partial\beta\partial\beta^\top}\log f(y_i|x_i,\beta_0)$, the equation (20) can now be rewritten as

$$0 = \begin{pmatrix} \bar{U}_n \\ \bar{V}_{n,S} \end{pmatrix} + B_n(\hat{\beta}-\beta_0) + \frac{1}{2}(\hat{\beta}-\beta_0)^\top C_n\xi^*(\hat{\beta}-\beta_0)$$

or equivalently as

$$\begin{pmatrix} \bar{U}_n \\ \bar{V}_{n,S} \end{pmatrix} = -(B_n + \frac{1}{2}(\hat{\beta}-\beta_0)^\top C_n\xi^*)\,(\hat{\beta}-\beta_0).$$

We observe that B_n converges to J_S, that C_n converges to $\mathrm{E}[H(X)]$ which is finite by (C4), that all the elements of ξ^* are bounded by $p+|S|$ (which is finite) and that $\hat{\beta}$ tends to β_0. All of this implies that $-(B_n + \frac{1}{2}(\hat{\beta}-\beta_0)^\top C_n\xi^*)$ tends to J_S, which ends the proof of Lemma 2.

□

Proof of Theorem 1. Taylor expansions of $\hat{\mu}_S$ and μ_{true} around $\mu_0 = \mu(\theta_0,\gamma_0)$ give

$$\sqrt{n}(\hat{\mu}_S - \mu_{\text{true}}) = \left(\frac{\partial\mu}{\partial(\theta,\gamma_S)}\right)^\top \begin{pmatrix} \sqrt{n}(\hat{\theta}_S-\theta_0) \\ \sqrt{n}(\hat{\gamma}_S-\gamma_{0,S}) \end{pmatrix} - \left(\frac{\partial\mu}{\partial\gamma}\right)^\top \delta + R_1 - R_2$$

with

$$R_1 = \frac{1}{2}n^{-1/2}\delta^\top \frac{\partial^2\mu}{\partial\gamma\partial\gamma^\top}_{|(\theta_0,\tilde{\gamma}_1)}\delta$$

and

$$R_2 = \frac{1}{2}n^{-1/2}\begin{pmatrix} \sqrt{n}(\hat{\theta}_S-\theta_0) \\ \sqrt{n}(\hat{\gamma}_S-\gamma_{0,S}) \end{pmatrix}^\top \frac{\partial^2\mu}{\partial(\theta,\gamma_S)\partial(\theta,\gamma_S)^\top}_{|(\tilde{\theta},\tilde{\gamma}_S)} \begin{pmatrix} \sqrt{n}(\hat{\theta}_S-\theta_0) \\ \sqrt{n}(\hat{\gamma}_S-\gamma_{0,S}) \end{pmatrix}$$

with $\tilde{\gamma}_1$ between γ_0 and $\gamma_0+\delta/\sqrt{n}$ and $(\tilde{\theta},\tilde{\gamma}_S)$ between (θ_0,γ_0) and $(\hat{\theta}_S,\hat{\gamma}_S,\gamma_{0,S^c})$. By (C5) and Lemma 1 2, $R_1 = o(1)$ and $R_2 = o_P(1)$ which implies that

$$\sqrt{n}(\hat{\mu}_S - \mu_{\text{true}}) \;\doteq_d\; \left(\frac{\partial\mu}{\partial(\theta,\gamma_S)}\right)^\top \begin{pmatrix} \sqrt{n}(\hat{\theta}_S-\theta_0) \\ \sqrt{n}(\hat{\gamma}_S-\gamma_{0,S}) \end{pmatrix} - \left(\frac{\partial\mu}{\partial\gamma}\right)^\top \delta.$$

Lemma 2 and algebraic manipulations end the proof. □

9.2 Proofs for Theorem 2

Proof of Lemma 3. Because $S_{0,n} \subseteq S$, it holds that $Y = X_\beta \beta_0 + X_{\gamma,S}\gamma_{n,S} + \epsilon$ with $\gamma_{n,S} = \delta_S/\sqrt{n}$. As conditions of Theorem 2.1 of van de Geer et al. (2014) hold for this linear model we have

$$\begin{pmatrix} \sqrt{n}(\hat{\beta}_S^{\text{desp}} - \beta_0) \\ \sqrt{n}(\hat{\gamma}_S^{\text{desp}} - \delta_S/\sqrt{n}) \end{pmatrix} \doteq_d \; W + \Delta_1,$$

with

$$W \sim \mathcal{N}_{p+|S|}\left(\begin{pmatrix} 0_p \\ 0_{|S|} \end{pmatrix} \; , \; M_S J_S M_S^\top \right)$$

and

$$\mathbb{P}\left[\|\Delta_1\|_\infty \geq 8\sqrt{n}\left(\max_j \frac{\lambda_j}{\hat{\tau}_j^2} \right) \frac{\lambda(p+s_n)}{\phi_0^2} \right] \leq 2\exp(-t^2)$$

which ends the proof of Lemma 3.

Proof of Theorem 2. The proof is straightforward using Lemma 3 and the same reasoning as for Theorem 1.

Acknowledgements

We acknowledge the support of the Fund for Scientific Research Flanders, KU Leuven grant GOA/12/14 and of the IAP Research Network P7/06 of the Belgian Science Policy. The computational resources and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by the Hercules Foundation and the Flemish Government - department EWI.

References

Behl, P., Claeskens, G., and Dette, H. (2014). Focussed model selection in quantile regression. *Statistica Sinica*, 24(2):601–624.

Bühlmann, P. and van de Geer, S. (2011). *Statistics for high-dimensional data*. Springer Series in Statistics. Springer, Heidelberg. Methods, theory and applications.

Bühlmann, P. and van de Geer, S. (2015). High-dimensional inference in misspecified linear models. *Electronic Journal of Statistics*, 9(1):1449–1473.

Claeskens, G. (2012). Focused estimation and model averaging with penalization methods: an overview. *Statistica Neerlandica*, 66(3):272–287.

Claeskens, G., Croux, C., and Van Kerckhoven, J. (2007). Prediction-focused model selection for autoregressive models. *Australian & New Zealand Journal of Statistics*, 49(4):359–379.

Claeskens, G. and Hjort, N. L. (2003). The focused information criterion. *Journal of the American Statistical Association*, 98:900–916.

Claeskens, G. and Hjort, N. L. (2008a). Minimizing average risk in regression models. *Econometric Theory*, 24(02):493–527.

Claeskens, G. and Hjort, N. L. (2008b). *Model selection and model averaging*. Cambridge series in statistical and probabilistic mathematics. Cambridge University Press, Cambridge, New York.

Claeskens, G., Magnus, J., Vasnev, A., and Wang, W. (2016). "The forecast combination puzzle: A simple theoretical explanation". *International Journal of Forecasting*, 32:754 – 762.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360.

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

Gueuning, T. and Claeskens, G. (2016). Confidence intervals for high-dimensional partially linear single-index models. *Journal of Multivariate Analysis*, 149:13–29.

Hjort, N. L. and Claeskens, G. (2003). Frequentist model average estimators. *Journal of the American Statistical Association*, 98(464):879–899.

Hjort, N. L. and Claeskens, G. (2006). Focused information criteria and model averaging for the Cox hazard regression model. *Journal of the American Statistical Association*, 101(476):1449–1464.

Javanmard, A. and Montanari, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research*, 15:2869–2909.

Kim, Y., Kwon, S., and Choi, H. (2012). Consistent model selection criteria on high dimensions. *The Journal of Machine Learning Research*, 13(1):1037–1057.

Luo, S. and Chen, Z. (2013). Extended BIC for linear regression models with diverging number of relevant features and high or ultra-high feature spaces. *Journal of Statistical Planning and Inference*, 143(3):494–504.

Meier, L., Meinshausen, N., and Dezeure, R. (2014). *hdi: High-Dimensional Inference*. R package version 0.1-2.

Pircalabelu, E., Claeskens, G., Jahfari, S., and Waldorp, L. (2016). A focused information criterion for graphical models in fMRI connectivity with high-dimensional data. *Annals of Applied Statistics*, 9(4):2179–2214.

Pircalabelu, E., Claeskens, G., and Waldorp, L. (2015). A focused information criterion for graphical models. *Statistics and Computing*, 25(6):1071–1092.

Reid, S., Tibshirani, R., and Friedman, J. (2016). A study of error variance estimation in lasso regression. *Statistica Sinica*, 26(1):35–67.

Rohan, N. and Ramanathan, T. V. (2011). Order selection in arma models using the focused information criterion. *Australian & New Zealand Journal of Statistics*, 53(2):217–231.

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. John Wiley Sons, Inc.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288.

van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202.

Vansteelandt, S., Bekaert, M., and Claeskens, G. (2012). On model selection and model misspecification in causal inference. *Statistical Methods in Medical Research*, 21(1, SI):7–30.

Wang, H., Li, B., and Leng, C. (2009). Shrinkage tuning parameter selection with a diverging number of

parameters. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(3):671–683.

Wang, H., Li, Y., and Sun, J. (2015). Focused and model average estimation for regression analysis of panel count data. *Scandinavian Journal of Statistics*, 42(3):732–745.

Xu, G., Wang, S., and Huang, J. Z. (2014). Focused information criterion and model averaging based on weighted composite quantile regression. *Scandinavian Journal of Statistics*, 41(2):365–381.

Yang, H., Liu, Y., and Liang, H. (2015). Focused information criterion on predictive models in personalized medicine. *Biometrical Journal*, 57(3):422–440.

Zhang, C.-H. and Zhang, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):217–242.

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

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429.

	Squared prediction errors				FIC values	
Focus number	Lasso	Best FIC	FIC 1	FIC 2	Value FIC 1	Value FIC 2
1	0.189	0.057	0.057	0.076	**1.138**	1.248
2	0.102	0.304	0.324	0.304	1.312	**1.271**
3	0.380	0.241	0.241	0.326	**0.479**	1.983
4	0.036	0.047	0.047	0.035	**0.368**	0.907
5	0.017	0.039	0.039	0.031	**0.475**	1.308
6	0.288	0.118	0.118	0.107	**0.916**	1.008
7	0.003	0.182	0.101	0.182	3.064	**2.899**
8	0.816	0.733	0.733	0.639	**0.305**	0.594
9	0.061	0.047	0.047	0.050	**0.271**	0.525
10	1.706	0.951	0.915	0.951	2.643	**1.989**
11	0.011	0.003	0.003	0.000	**0.344**	0.454
12	0.044	0.001	0.001	0.001	1.191	**1.117**
13	0.635	0.370	0.399	0.370	1.230	**0.806**
14	0.081	0.009	0.009	0.009	**1.224**	1.358
15	0.188	0.130	0.130	0.114	**0.357**	0.500
16	0.049	0.001	0.000	0.001	1.558	**0.863**
17	0.045	0.034	0.034	0.045	**0.220**	0.715
18	0.003	0.002	0.005	0.002	0.670	**0.666**
19	0.021	0.016	0.016	0.028	**0.011**	0.358
20	0.002	0.000	0.000	0.001	**0.354**	0.394
21	0.254	0.502	0.502	0.549	**1.229**	1.470
Average	0.235	0.180	0.177	0.182		

Table 5: Squared prediction errors and FIC values for the 21 focuses of the riboflavin data. The FIC search is done trough a stepwise procedure with as starting set the empty set for FIC 1 and the active set of the Lasso for FIC 2. Best FIC consists is obtain by keeping the submodel that gives the smallest of the two FIC values.

	Lasso	Best FIC	FIC 1	FIC 2
Average number of selected variables	27	6.7	4.6	10.7
Number of variables selected at least once	27	120	77	177
Number of variables selected at least 3 times	27	5	2	10

Table 6: Information on the variables selected for the 21 focuses of the riboflavin data. The FIC search is done trough a stepwise procedure with as starting set the empty set for FIC 1 and the active set of the Lasso for FIC 2. Best FIC consists is obtained by keeping the submodel that gives the smallest of the two FIC values.

$\|S\|$	FIC OLS	FIC desp.	Pred. OLS	Pred. desp.
5	**12.46**	12.56	-0.42	-0.42
10	45.39	**45.01**	0.10	0.10
15	26.23	**26.21**	-0.22	-0.21
20	39.33	**39.18**	-0.18	-0.18
25	**21.43**	21.63	-0.44	-0.43
30	17.14	**15.47**	-0.73	-0.75
35	**21.81**	22.67	-0.67	-0.65
40	**31.90**	38.50	-0.32	-0.33
41	**32.10**	39.29	-0.32	-0.45
42	**32.85**	37.50	-0.29	-0.39
43	67.16	**58.84**	0.42	0.09
44	55.61	**46.93**	0.22	-0.43
45	**92.91**	130.30	0.57	0.70
46	114.14	**85.71**	0.83	0.46
47	121.62	**98.36**	0.47	0.52
48	416.60	**119.20**	1.44	0.77
49	924.01	**114.23**	4.51	0.73
50	n/a	89.99	n/a	0.55
60	n/a	79.35	n/a	0.15
70	n/a	66.49	n/a	-0.26
80	n/a	133.70	n/a	0.73
90	n/a	66.53	n/a	0.02
100	n/a	98.10	n/a	-0.43

Table 7: OLS FIC of Section 3 and desparsified FIC of Section 4 and their corresponding predictions for one random subset for each considered size. The focus is $\mu_1 = X^1_{\text{test}}\beta$ whose true value is unknown but should be close to $y^1_{\text{test}} = -1.13$.

A High-dimensional Focused Information Criterion

Thomas Gueuning and Gerda Claeskens

ORSTAT and Leuven Statistics Research Center
KU Leuven, Faculty of Economics and Business
Naamsestraat 69, 3000 Leuven, Belgium
thomas.gueuning@kuleuven.be, gerda.claeskens@kuleuven.be

March 21, 2017

Abstract

The focused information criterion for model selection is constructed to select the model that best estimates a particular quantity of interest, the focus, in terms of mean squared error. We extend this focused selection process to the high-dimensional regression setting with potentially a larger number of parameters than the size of the sample. We distinguish two cases: (i) the case where the considered submodel is of low-dimension and (ii) the case where it is of high-dimension. In the former case, we obtain an alternative expression of the low-dimensional focused information criterion that can directly be applied. In the latter case we use a desparsified estimator that allows us to derive the mean squared error of the focus estimator. We illustrate the performance of the high-dimensional focused information criterion with a numerical study and a real dataset.

Keywords: Desparsified estimator; Focused information criterion; High-dimensional data; Variable selection.

Running headline: A high-dimensional FIC

1 Introduction

We extend the theory of the focused information criterion (FIC) for variable selection in parametric models to allow a diverging dimension of the parameter, permitting us to apply the method on high-dimensional data where the number of parameters may exceed the sample size. To do so, we extend the desparsified estimator of van de Geer et al. (2014) to the local misspecification framework. The FIC philosophy puts less emphasis on which variables are in the model but rather on the accuracy of the estimator of a focus, which is a differentiable function of the model parameters. The accuracy of the estimation is assessed via the mean squared error (MSE).

For example in the context of prediction with linear models, the FIC permits to use different variables to make predictions for different new observations of the covariate vector. We illustrate this on a real data set containing 4088 variables and 71 observations that we split in a training set of size 50 and a testing set of size 21. Whereas the usual approach consists in using the same penalized estimator and thus the same covariates to obtain the 21 predictions, the FIC allows us to use different covariates for each of the 21 different predictions. In our example, the mean squared prediction error is improved from 0.235 with a penalized estimator approach to 0.180 with the FIC approach.

The FIC has been introduced by Claeskens and Hjort (2003) for low-dimensional likelihood models, see

also Claeskens and Hjort (2008b, Ch. 6). This approach of focused selection has further been extended to several application areas including panel count data (Wang et al., 2015), graphical models (Pircalabelu et al., 2015) and personalized medicine (Yang et al., 2015). Focused selection for quantile regression has been studied by (Behl et al., 2014), and for weighted composite quantile regression by Xu et al. (2014). Focused selection for causal inference has been obtained by (Vansteelandt et al., 2012). Other model classes where focused selection has been studied include time series models (Rohan and Ramanathan, 2011; Claeskens et al., 2007), partially linear models (Zhang and Liang, 2011) and survival data (Hjort and Claeskens, 2006), without being complete in this overview.

Variable selection and estimation for high-dimensional data is most often performed simultaneously by using penalization methods; for an overview, see Fan and Lv (2010). The use of lasso-type estimators (Tibshirani, 1996) and its variations is currently well known. For theoretical results, see Bühlmann and van de Geer (2011). However, one should realize that also such methods, as do most other variable selection procedures, aim at selecting one 'best' model that one hence is supposed to use to estimate all quantities of interest related to that dataset. In contrast, the focused information criterion (FIC) may select different models for different quantities interest, which we call the focuses.

The introduction of the FIC for a diverging number of parameters is important and has a large application area. Claeskens (2012) gave a FIC formula for penalized estimators but required the dimension of the parameters to be fixed. Thus the small n (sample size) – large p (number of parameters) case is asymptotically not covered by that work. That form of FIC for penalized estimators with a fixed dimension is used by Pircalabelu et al. (2016) for high-dimensional graphical models.

Besides penalization procedures, several other variable selection procedures have been developed for high-dimensional data. In particular, Luo and Chen (2013) establish the consistency of the extended Bayesian information criterion (EBIC) with a diverging number of relevant features but need to restrict to low-dimensional submodels. Kim et al. (2012) obtain the consistency of the generalized information criterion (GIC) and Wang et al. (2009) propose a modified BIC (mBIC) whose consistency is shown for a number of parameters that diverges slower than the sample size.

The paper is organized as follows. In Section 2, we define the general framework and recall the classical FIC formula for fixed dimensions. In Section 3, we introduce the FIC for high-dimensional data when the considered submodel is of low dimension. This also provides an alternative formula in the classical FIC setting. In Section 4 we consider the high-dimensional submodel case in which $p + |S| > n$ and restrict to linear models. In that case the maximum likelihood estimator is not available because the Fisher information (sub)matrix is not invertible. To tackle this problem we use a desparsifying estimator, following the idea of van de Geer et al. (2014), Javanmard and Montanari (2014) and Zhang and Zhang (2014). In Section 5, we give some practical considerations for the computation of the FIC, including information over the estimation of $\delta\delta^\top$ and in Section 6 we give numerical results. In Section 7, we illustrate the FIC procedure on the real data set riboflavin from package *hdi* and compare it to a regular penalization approach. Section 8 provides some insights over the extension of results of Section 4 to the generalized linear models. All proofs are given in Section 9.

2 Model, notations and limitations of the current FIC literature

2.1 Notation and framework

Let $Y_1, \ldots, Y_n$ be independent response values with concomitant covariates $x_1, \ldots, x_n$, such that Y_i has a density $f(y|x_i, \theta_0, \gamma_n)$. The vector γ_n contains all the parameters on which we want to perform variable selection and has length q_n that is allowed to grow with n. The parameter vector θ_0 is of fixed length p and contains all the parameters that we want to include in every considered model. These parameters are protected. For example, for a linear model $Y_i = \beta_0 + x_i^\top \beta + \sigma\epsilon_i$ with $\epsilon_i \sim \mathcal{N}(0,1)$, it is quite common to include the parameters σ and β_0 in every considered submodel. Thus a natural choice would be $\theta_0 = (\sigma, \beta_0)$. It might also be relevant in some cases to include some of the components of β in the protected set (e.g. some variables that are known beforehand to be important). The covariate x_i is of diverging length r_n. Very often r_n and q_n are of the same order but it is not necessary the case. We present two simple examples that illustrate the link between p, q_n and r_n.

First assume that we want to fit a linear model for $(Y_i, x_i), i = 1, \ldots, n$ and that we want to include the first three components of x_i in every considered model. These three components might for instance be the age, the weight and the height of an individual while all other components might consist of blood information or gene expressions. The full model, the largest model under consideration, is then $Y_i = \beta_0 + \sum_{j=1}^{3} x_{i,j}\beta_j + \sum_{j=4}^{r_n} x_{i,j}\beta_j + \sigma\epsilon_i$ and we have $\theta_0 = (\sigma, \beta_0, \beta_1, \beta_2, \beta_3)$ and $\gamma_n = (\beta_4, \ldots, \beta_{r_n})$. Thus $p = 5$ and $q_n = r_n - 3$.

In a second example, assume that we still want to fit a linear model for (Y_i, x_i) but that none of the components of x_i should be protected. Furthermore assume that we want to consider interaction terms as well as first and second order terms in the possible models. Then $p = 2$ (for the error standard deviation level and the intercept) and $q_n = r_n + r_n(r_n + 1)/2$.

As in the earlier studies about FIC (e.g. Claeskens and Hjort, 2003; Claeskens, 2012) we consider the local misspecification framework where $\gamma_n = \gamma_0 + \delta/\sqrt{n}$, with the major difference that the length q_n of δ is diverging. Each component of δ is $O(1)$. This framework allows us to study the mean squared error (MSE) of the estimator of the further-defined focus, with a balance between the squared bias and the variance, without having the bias or the variance dominating the mean squared error expression. We refer to Claeskens and Hjort (2008b, Sec. 5.5) for more details regarding the local misspecification setting.

Taking $\gamma_n = \gamma_0$, a known value, corresponds to working with the simplest model, often called *the narrow model*. In the two examples given hereabove, it is natural to choose $\gamma_{0,j} = 0$ for each j: the simplest model consists in not including the unprotected variables. In other cases, $\gamma_{0,j}$ might be nonzero, see for instance example 5.4 in Claeskens and Hjort (2008b) in which the skewing logistic regression model $p_i = H(x_i^\top \beta + z_i^\top \alpha)^\kappa$ is considered. In that example, κ is an unprotected parameter that takes value 1 in the narrow model.

We denote by $S_{0,n} = \{j : \delta_j \neq 0\}$ the active set of coefficients where we emphasize in the notation the fact that the length of δ is growing with n and we write $s_n = |S_{0,n}|$, the number of elements of $S_{0,n}$. We

consider subsets S of $\{1, \ldots, q_n\}$ and denote by (sub)model S the model containing θ and those parameters γ_j with j belonging to S. This model corresponds to working with a density $f(y|x, \theta, \gamma_S, \gamma_{0,S^c})$ with S^c the complementary set of S. The slight abuse of notation groups the components of γ by whether their index is present or absent in S. When fitting a model with the p protected parameters in θ, and $|S|$ added parameters, in total $p + |S|$ parameters need to be estimated. Let us denote by $(\hat{\theta}_S, \hat{\gamma}_S)$ an estimator of $(\theta_0, \gamma_{n,S})$. See Sections 3 and 4 for more details.

2.2 The focused information criterion

Following the FIC philosophy, we are interested in estimating as accurately as possible (in terms of MSE) a particular quantity of interest $\mu_{\text{true}} = \mu(\theta_0, \gamma_n)$, called *the focus*. A model that is best in terms of MSE for one focus μ, might not be the best for another focus. This leads to a tailored model choice where one first specifies the focus and then searches for the best model for that particular goal. In this sense it should be clear that the FIC is not constructed to aim for selection consistency.

We make the assumption that μ is differentiable with respect to θ and γ such that $\|[(\frac{\partial \mu}{\partial \theta})^\top, (\frac{\partial \mu}{\partial \gamma_S})^\top]\|_\infty = K = O(1)$ in a neighborhood of θ_0, γ_0. Several examples of such quantities of interest are given in Claeskens and Hjort (2008b). The focus might for example be the prediction for a particular subgroup of the population, the estimation of the impact of one particular covariate on the response or a particular quantile for a specific value of the covariates. The goal is to find the submodel S whose corresponding estimator $\hat{\mu}_S = \mu(\hat{\theta}_S, \hat{\gamma}_S, \gamma_{0,S^c})$ of the focus is the best in terms of mean squared error. For a submodel S we are thus interested in the limiting distribution Λ_S of $\sqrt{n}(\hat{\mu}_S - \mu_{\text{true}})$. The focused information criterion estimates the corresponding limiting mean squared error. Thus,

$$\text{FIC}(S) = \widehat{\text{E}(\Lambda_S)^2} + \widehat{\text{Var}(\Lambda_S)}.$$

Different models, say indexed by S and S', might have different values for the bias and variance of the submodel-based estimator of μ, thus Λ_S might be different from $\Lambda_{S'}$. Hence, models can be ranked based on their FIC value. The model S with the smallest FIC(S) value amongst all considered models, is selected as the best one for the purpose of estimating the focus μ.

In the low-dimensional framework with γ_n and δ of fixed dimensions $q \times 1$, Claeskens and Hjort (2003) show that if $(\hat{\theta}_S, \hat{\gamma}_S)$ is the maximum likelihood estimator, the limiting MSE of Λ_S is

$$\text{MSE}(S) = \omega^\top (I_q - G_S)\delta\delta^\top (I_q - G_S)^\top \omega + \left(\frac{\partial \mu}{\partial \theta}\right)^\top J_{00}^{-1} \frac{\partial \mu}{\partial \theta} + \omega^\top G_S Q_S G_S^\top \omega, \tag{1}$$

with the $(p+q) \times (p+q)$ Fisher information matrix J and its inverse matrix denoted by

$$J = \begin{pmatrix} J_{00} & J_{01} \\ J_{10} & J_{11} \end{pmatrix}, \quad J^{-1} = \begin{pmatrix} J^{00} & J^{01} \\ J^{10} & J^{11} \end{pmatrix},$$

where $Q = J^{11}$, $G_S = \pi_S^\top Q_S \pi_S Q^{-1}$, $Q_S = J^{11,S}$, $\omega = J_{10} J_{00}^{-1} \frac{\partial \mu}{\partial \theta} - \frac{\partial \mu}{\partial \gamma}$ and $\pi_S \in \mathbb{R}^{S \times q}$ the projection matrix related to S, obtained by extracting the rows of the $q \times q$ identity matrix for which the row number is in S. Claeskens and Hjort (2008b, Sec. 6.7) show that for linear models the limiting MSE is in fact

the exact MSE when the focus takes the form $\mu = x^\top\beta + z^\top\gamma$. For a vector-valued focus one could use a one-dimensional summary of the corresponding mean squared error matrix to be minimized over the different models, such as the matrix' trace, determinant, a matrix norm, etc.

The current FIC formula (1) can not be applied in our framework for two reasons. First, and most importantly, the theory assumes that the dimension of γ_n is fixed. A diverging number of parameters is not supported by the theory. In this paper, we allow the dimension q_n of γ_n to grow with n and we make the sparsity assumption $s_n = o(n^{1/4})$. Secondly, the current version of the FIC formula is in many cases not available for high-dimensional data, even for low-dimensional submodels. Indeed, it requires to invert the Fisher information matrix J that is in many cases not invertible. For example, for a normal linear model, the Fisher information matrix $J = \sigma^{-2}\text{diag}\{2, n^{-1}X^\top X\}$. When $q_n > n$ the matrix J is by construction not invertible so that the expression (1) is not defined. These considerations motivate us to develop the FIC theory for a diverging number of parameters.

We distinguish two cases in the model selection search: (i) the submodel is low-dimensional such that regular least squares or maximum likelihood estimators can be computed, and (ii) the submodel is high-dimensional, requiring a regularized estimator. In both cases, an adjustment of the existing focused selection approach is needed. These two cases are studied in the next two sections.

We now give some notations. For two random variables A and B, the notation $A \doteq_d B$ means that $A - B \xrightarrow{p} 0$. Furthermore, we write $f_{\text{true}}(y|x) = f(y|x, \theta_0, \gamma_0 + \delta/\sqrt{n})$ the true density function, $f_0(y|x) = f(y|x, \theta_0, \gamma_0)$ the density function in the narrow model and $U(y|x) = \frac{\partial}{\partial\theta}\log f(y|x,\theta,\gamma)_{|(\theta_0,\gamma_0)}$ and $V(y|x) = \frac{\partial}{\partial\gamma}\log f(y|x,\theta,\gamma)_{|(\theta_0,\gamma_0)}$ the derivatives of the log-density evaluated in the narrow model. We define in the regression model's context

$$J(x) = \int f_0(y|x)\begin{pmatrix} U(y|x) \\ V(y|x)\end{pmatrix}\begin{pmatrix} U(y|x) \\ V(y|x)\end{pmatrix}^\top dy, \quad J_n = \frac{1}{n}\sum_{i=1}^n J(x_i) = \begin{pmatrix} J_{n,00} & J_{n,01} \\ J_{n,10} & J_{n,11}\end{pmatrix},$$

the latter matrix is the empirical Fisher information matrix. For a fixed subset S of $\{1, \ldots, q_n\}$, we denote by π_S the $|S|\times q_n$ projection matrix related to S that, when multiplied to a matrix or row vector consisting of q_n rows, selects those rows corresponding to the elements in S. Further, define

$$\pi_S^* = \begin{pmatrix} I_p & 0_{p\times q_n} \\ 0_{|S|\times p} & \pi_S\end{pmatrix}, \quad J_S(x) = \pi_S^* J(x)\pi_S^{*\top} = \int f_0(y|x)\begin{pmatrix} U(y|x) \\ V_S(y|x)\end{pmatrix}\begin{pmatrix} U(y|x) \\ V_S(y|x)\end{pmatrix}^\top dy$$

and write $J_{n,S} = \frac{1}{n}\sum_{i=1}^n J_S(x_i)$ the empirical Fisher matrix in model S and $J_S = \lim_{n\to\infty} J_{n,S}$. Note that $J_{n,S}$ is of fixed dimension $(p+|S|)\times(p+|S|)$ while J_n is of diverging dimension $(p+q_n)\times(p+q_n)$. As a consequence, for an unbounded sequence $\{q_n, n\to\infty\}$, J_n does not converge to a fixed quantity J.

3 FIC for a low-dimensional submodel

We consider the local misspecification framework of Section 2. Let S be a fixed subset of $\{1, \ldots, q_n\}$ such that the number of parameters $p + |S|$ to estimate in the submodel S is smaller than the sample size n.

Let us consider the maximum likelihood estimator for $(\theta_0, \gamma_{n,S})$

$$(\hat{\theta}_S, \hat{\gamma}_S) = \underset{\theta\in\mathbb{R}^p \gamma_S\in\mathbb{R}^{|S|}}{\arg\max} \frac{1}{n}\sum_{i=1}^{n} \log\ f(y_i|x_i, \theta, \gamma_S, \gamma_{0,S^c}) \tag{2}$$

and define the estimator of the focus in this model by

$$\hat{\mu}_S = \mu(\hat{\theta}_S, \hat{\gamma}_S, \gamma_{0,S^c}). \tag{3}$$

Before presenting our theoretical result for the limiting distribution of $\sqrt{n}(\hat{\mu}_S - \mu_{\text{true}})$ we give the corresponding conditions. A Taylor expansion of $f_{\text{true}}(y|x)$ gives

$$f_{\text{true}}(y|x) = f_0(y|x)\left\{1 + V(y|x)^\top \delta/\sqrt{n} + R(y|x, \delta/\sqrt{n})\right\}. \tag{4}$$

We make the following conditions, that are similar to Hjort and Claeskens (2003), phrased here for the regression setting.

(C1) The two integrals $\int f_0(y|x)U(y|x)R(y,t)dy$ and $\int f_0(y|x)V(y|x)R(y,t)dy$ are both $O(\|t\|_1^2)$, with R defined in (4).

(C2) The variables $|U_l(y|x)^2 V_k(y|x)|$ and $|V_j(y|x)^2 V_k(y|x)|$ have finite mean under $f_0(y|x)$ for each $1 \leq l \leq p$ and $j, k \in S$.

(C3) The two integrals $\int f_0(y|x)\left\|U(y|x)\right\|^2 R(y,t)dy$ and $\int f_0(y|x)\left\|V_S(y|x)\right\|^2 R(y,t)dy$ are both $o(1)$.

(C4) The log density has three continuous derivatives w.r.t the $p+|S|$ parameters (θ, γ_S) in a neighbourhood around (θ_0, γ_0), and they are dominated by functions with finite means under f_0.

(C5) $s_n = o(n^{1/4})$.

Conditions (C1) to (C4) are similar to those of Hjort and Claeskens (2003) in the low-dimensional case, while condition (C5) is a sparsity condition to deal with high-dimensional vectors.

Lemma 1. *Under (C1), (C2), (C3) and (C5), we have*

$$\begin{pmatrix} \sqrt{n}\bar{U}_n \\ \sqrt{n}\bar{V}_{n,S} \end{pmatrix} - \begin{pmatrix} J_{n,01}\delta \\ \pi_S J_{n,11}\delta \end{pmatrix} \xrightarrow{d} \mathcal{N}_{p+|S|}(0, J_S)$$

with $\bar{U}_n = \frac{1}{n}\sum_{i=1}^n U(y_i|x_i)$, $\bar{V}_{n,S} = \frac{1}{n}\sum_{i=1}^n V_S(y_i|x_i)$.

Lemma 2. *Under (C1), (C2), (C3), (C4) and (C5), we have*

$$\begin{pmatrix} \sqrt{n}(\hat{\theta}_S - \theta_0) \\ \sqrt{n}(\hat{\gamma}_S - \gamma_{0,S}) \end{pmatrix} - J_S^{-1}\begin{pmatrix} J_{n,01}\delta \\ \pi_S J_{n,11}\delta \end{pmatrix} \xrightarrow{d} \mathcal{N}_{p+|S|}(0, J_S^{-1}).$$

The following theoretical result is an extension of Theorem 6.1 of Claeskens and Hjort (2008b) to the diverging number of parameters case. It covers the important $p + q > n$ case and can thus be applied on high-dimensional data. A proof is given in Section 9.

Theorem 1. *Under conditions (C1) to (C5) it holds for the estimator (3) of the focus in model S that*

$$\sqrt{n}(\hat{\mu}_S - \mu_{\text{true}}) \ \doteq_d \ \Lambda_{n,S}$$

with

$$\Lambda_{n,S} = \left(\frac{\partial \mu}{\partial \theta}\right)^\top C_S + \left(\frac{\partial \mu}{\partial \gamma_S}\right)^\top D_S - \left(\frac{\partial \mu}{\partial \gamma}\right)^\top \delta \tag{5}$$

$$= \begin{pmatrix} \frac{\partial \mu}{\partial \theta} \\ \frac{\partial \mu}{\partial \gamma} \end{pmatrix}^\top \left(B_S \delta + \pi_S^{*\top} J_S^{-1} \begin{pmatrix} U \\ V_S \end{pmatrix} \right) \tag{6}$$

where the partial derivatives are evaluated at the null point (θ_0, γ_0) and where

$$\begin{pmatrix} C_S \\ D_S \end{pmatrix} = J_S^{-1} \begin{pmatrix} J_{n,01}\delta + U \\ \pi_S J_{n,11}\delta + V_S \end{pmatrix} \ \textit{with} \ \begin{pmatrix} U \\ V_S \end{pmatrix} \sim \mathcal{N}_{p+|S|}(0, J_S),$$

and

$$B_S = \pi_S^{*\top} J_S^{-1} \begin{pmatrix} J_{n,01} \\ \pi_S J_{n,11} \end{pmatrix} - \begin{pmatrix} 0_{p \times q_n} \\ I_{q_n} \end{pmatrix}.$$

The sparsity condition $s_n = o(n^{1/4})$ is crucial in this high-dimensional framework. Note that $\Lambda_{n,S}$ depends on n through $\frac{\partial \mu}{\partial \gamma}$, B_S and δ. While (5) leads to the original FIC formula, (6) turns out to be more useful in the high-dimensional case. From Theorem 1, for a model S, the limiting distribution of $\sqrt{n}(\hat{\mu}_S - \mu_{\text{true}})$ is the same as the one of $\Lambda_{n,S}$ which is normally distributed with mean and variance given by

$$\text{E}(\Lambda_{n,S}) = \begin{pmatrix} \frac{\partial \mu}{\partial \theta} \\ \frac{\partial \mu}{\partial \gamma} \end{pmatrix}^\top B_S \delta \ \text{ and } \ \text{Var}(\Lambda_{n,S}) = \begin{pmatrix} \frac{\partial \mu}{\partial \theta} \\ \frac{\partial \mu}{\partial \gamma} \end{pmatrix}^\top \pi_S^{*\top} J_S^{-1} \pi_S^* \begin{pmatrix} \frac{\partial \mu}{\partial \theta} \\ \frac{\partial \mu}{\partial \gamma} \end{pmatrix}.$$

We thus have

$$\text{MSE}(S, \delta) = \begin{pmatrix} \frac{\partial \mu}{\partial \theta} \\ \frac{\partial \mu}{\partial \gamma} \end{pmatrix}^\top \left(B_S \delta \delta^\top B_S^\top + \pi_S^{*\top} J_S^{-1} \pi_S^* \right) \begin{pmatrix} \frac{\partial \mu}{\partial \theta} \\ \frac{\partial \mu}{\partial \gamma} \end{pmatrix}, \tag{7}$$

and $\text{FIC}(S, \delta) = \widehat{\text{MSE}(S, \delta)}$ which defines the FIC in the high-dimensional setting for a low-dimensional submodel S. Interestingly, this formulation does not require the inverse of the Fisher matrix in the full model but only in the submodel S. Thus this expression may be used in the high-dimensional setting with $p + q_n > n$ if the considered submodel S is of low dimension, that is if $p + |S| \leq n$.

In fact, the formula (7) could also be used to compute the FIC in the classical fixed low dimensional case. Indeed, it is possible to show that for fixed q with $p + q < n$ the expressions (1) and (7) are equal, with $J_{n,01}$ and $J_{n,11}$ replaced by their limiting versions J_{01} and J_{11}, this is that

$$\omega^\top (I_q - G_S)\delta\delta^\top (I_q - G_S)^\top \omega + \left(\frac{\partial \mu}{\partial \theta}\right)^\top J_{00}^{-1} \frac{\partial \mu}{\partial \theta} + \omega^\top G_S Q_S G_S^\top \omega$$

is equal to

$$\begin{pmatrix} \frac{\partial \mu}{\partial \theta} \\ \frac{\partial \mu}{\partial \gamma} \end{pmatrix}^\top \left(B_S \delta \delta^\top B_S^\top + \pi_S^{*\top} J_S^{-1} \pi_S^* \right) \begin{pmatrix} \frac{\partial \mu}{\partial \theta} \\ \frac{\partial \mu}{\partial \gamma} \end{pmatrix}.$$

This can be obtained using $Q^{-1} = J_{11} - J_{10}J_{00}^{-1}J_{01}$ and $G_S Q_S G_S^\top = \pi_S^\top Q_S \pi_S$. The main novel contribution of the low-dimensional submodel case, though, is that the theory takes the presence of the high-dimensional vector δ into account.

To conclude this section, we note that if we wish to not protect any variable in the model selection procedure, our theory is still valid. In that case the Fisher information matrix is a $q_n \times q_n$ matrix and Theorem 1 and expression (7) are still valid with slight adjustments. The partial derivative $\frac{\partial \mu}{\partial \theta}$ disappears, B_S becomes $\pi_S^\top J_S \pi_S J_n - I_q$ and π_S^* becomes π_S. This remark also holds for the high-dimensional submodel case in the next section.

4 FIC for a high-dimensional submodel of a linear model

Let S be a subset of $\{1, \ldots, q_n\}$ of size larger than $n - p$. The maximum likelihood estimator (or least-squares estimator) is not available anymore and the results from Section 3 are not applicable. We propose to first use a ℓ_1-penalized estimator and then to desparsify it to obtain an estimator of $(\theta_0, \gamma_{n,S})$ whose distribution can be tracked. The idea to desparsify a penalized estimator has been introduced by several authors, including van de Geer et al. (2014), Javanmard and Montanari (2014) and Zhang and Zhang (2014). In this section, we restrict to linear models but extensions to generalized linear models and convex loss functions are expected to be feasible.

The desparsification is needed because the ℓ_1-based penalties have the property of setting some of the coefficients exactly equal to zero, one can show asymptotic consistency of such selection under some conditions. The remaining non-zero coefficients are estimated by an estimator which can asymptotically be normally distributed. This is the case for the adaptive Lasso (see Zou, 2006) and the SCAD (see Fan and Li, 2001). Since the focus might be a function of both types of coefficients, those that will be estimated by zero and those that will not, the asymptotic distribution of the focus estimator is not tractable due to this mixture containing a point-mass at zero.

Let us assume that for $i = 1, \ldots, n$, the response Y_i is generated by a linear model

$$Y_i = x_{\beta,i}^\top \beta_0 + x_{\gamma,i}^\top \gamma_n + \sigma \epsilon_i \tag{8}$$

with $\epsilon_i \sim \mathcal{N}(0,1)$, where $\beta_0 \in \mathbb{R}^p$ corresponds to the protected variables, $x_{\beta,i}$ is a $p \times 1$ vector of protected covariates, $\gamma_n \in \mathbb{R}^{q_n}$ corresponds to the unprotected parameters with corresponding covariate vector $x_{\gamma,i}$ on which variable selection is performed.

As in most of the high-dimensional literature, we assume that the noise variance σ^2 is known. Reid et al. (2016) describe strategies for estimating σ^2 and their empirical comparison suggests that using the estimator based on the residual sum of squares of cross-validated Lasso solution might yield a good estimator. For theoretical properties we refer to this paper. With σ^2 assumed to be known, the protected parameter θ_0 is thus β_0 and we note that for a linear model, $\gamma_0 = 0_{q_n}$ so that we have $\gamma_n = \delta/\sqrt{n}$ in this

section. We write

$$X_\beta = \begin{bmatrix} x_{\beta,1}^\top \\ \vdots \\ x_{\beta,n}^\top \end{bmatrix} \in \mathbb{R}^{n\times p},\ X_\gamma = \begin{bmatrix} x_{\gamma,1}^\top \\ \vdots \\ x_{\gamma,n}^\top \end{bmatrix} \in \mathbb{R}^{n\times q_n},\ X_{\gamma,S} = X_\gamma \pi_S^\top \text{ and } X_S^* = [X_\beta, X_{\gamma,S}] \in \mathbb{R}^{n\times(p+|S|)}.$$

The matrix X_S^* corresponds to the design matrix in the submodel S. Denoting by Y the vector of responses and ϵ the vector of the errors, we have $Y = X_\beta\beta_0 + X_\gamma\gamma_n + \epsilon$.

This section proceeds as follows. First, in section 4.1, we derive a desparsified estimator that can be interpreted as a generalization of the ordinary least-sqaures estimator. In section 4.2, we describe how to construct a relaxed inverse of the sample covariance matrix. Next, in section 4.3, we consider the case that a submodel S contains the true active set and derive theoretical results. In section 4.4, we derive a FIC formula for a general submodel S.

4.1 Desparsified estimator

Let us consider the following Lasso estimator (Tibshirani, 1996) where we do not penalize the intercept parameter (or take a model without intercept by centering the variables),

$$(\hat\beta_S^{\text{Lasso}}, \hat\gamma_S^{\text{Lasso}}) = \underset{\beta\in\mathbb{R}^p, \gamma_S\in\mathbb{R}^{|S|}}{\arg\min} \frac{1}{2n}\left\| Y - X_S^* \begin{pmatrix}\beta \\ \gamma_S\end{pmatrix}\right\|_2^2 + \lambda \left\| \begin{pmatrix}\beta \\ \gamma_S\end{pmatrix}\right\|_1. \tag{9}$$

We describe how to construct a desparsified estimator. The derivation presented herebelow is based on van de Geer et al. (2014). We write the Karush-Kuhn-Tucker condition

$$\frac{1}{n} X_S^{*\top}\left(Y - X_S^* \begin{pmatrix}\hat\beta_S^{\text{Lasso}} \\ \hat\gamma_S^{\text{Lasso}}\end{pmatrix}\right) = \lambda\hat\kappa_S; \quad \text{with } \hat\kappa_{S,j} = \text{sign}\left(\begin{pmatrix}\hat\beta_S^{\text{Lasso}} \\ \hat\gamma_S^{\text{Lasso}}\end{pmatrix}_j\right) \text{ if } \begin{pmatrix}\hat\beta_S^{\text{Lasso}} \\ \hat\gamma_S^{\text{Lasso}}\end{pmatrix}_j \neq 0, \tag{10}$$

where $\|\hat\kappa_S\|_\infty \leq 1$.

The matrix $J_S = \frac{1}{n\sigma^2} X_S^{*\top} X_S^*$ is by construction not invertible because $p+|S| > n$. We construct a relaxed inverse M_S of J_S by using the Lasso nodewise regression technique, as presented in van de Geer et al. (2014) and in section 4.2, and we define the following desparsified estimator:

$$\begin{aligned} \begin{pmatrix}\hat\beta_S^{\text{desp}} \\ \hat\gamma_S^{\text{desp}}\end{pmatrix} &= \begin{pmatrix}\hat\beta_S^{\text{Lasso}} \\ \hat\gamma_S^{\text{Lasso}}\end{pmatrix} + M_S \tfrac{1}{n\sigma^2} X_S^{*\top}\left(Y - X_S^* \begin{pmatrix}\hat\beta_S^{\text{Lasso}} \\ \hat\gamma_S^{\text{Lasso}}\end{pmatrix}\right) \\ &= M_S \tfrac{1}{n\sigma^2} X_S^{*\top} Y + \left(I_{p+|S|} - M_S J_S\right) \begin{pmatrix}\hat\beta_S^{\text{Lasso}} \\ \hat\gamma_S^{\text{Lasso}}\end{pmatrix}. \end{aligned} \tag{11}$$

We now give some intuition of the desparsfying estimator defined in (11). It can be seen as a bias-corrected version of the Lasso (first line) or as what we could call a pseudo-least-squares estimator in a high-dimensional framework (second line). We focus on the second interpretation. Since J_S is not invertible and M_S is used as a relaxed inverse, we could use the estimator $M_S \frac{1}{n\sigma^2} X_S^{*\top} Y$. This estimator

has mean $M_S J_S(\beta_0^\top, \gamma_{0,S}^\top + \delta_S^\top/\sqrt{n})^\top$ and variance $\frac{1}{n} M_S J_S M_S^\top$. We aim to correct this bias by adding $(I_{p+|S|} - M_S J_S)(\hat{\beta}^\top, \hat{\gamma}_S^\top)^\top$ using a reasonable estimator of the parameter vector. Here and in several referenced papers, the lasso estimator is taken.

By plugging $Y = X_\beta \beta_0 + X_\gamma \delta/\sqrt{n}$ into (11), we obtain the following equalities.

$$\begin{aligned}
\begin{pmatrix} \sqrt{n}(\hat{\beta}_S^{\text{desp}} - \beta_0) \\ \sqrt{n}(\hat{\gamma}_S^{\text{desp}} - \gamma_0) \end{pmatrix} &= M_S \begin{pmatrix} J_{01}\delta \\ \pi_S J_{11}\delta \end{pmatrix} + \left(I_{p+|S|} - M_S J_S\right) \begin{pmatrix} 0_p \\ \delta_S \end{pmatrix} \\
&\quad + M_S \tfrac{1}{\sqrt{n}\sigma^2} X_S^{*\top} \epsilon - \sqrt{n}\left(I_{p+|S|} - M_S J_S\right) \begin{pmatrix} \hat{\beta}_S^{\text{Lasso}} - \beta_0 \\ \hat{\gamma}_S^{\text{Lasso}} - \frac{\delta_S}{\sqrt{n}} \end{pmatrix} \\
&= \begin{pmatrix} 0_p \\ \delta_S \end{pmatrix} + M_S \tfrac{1}{n\sigma^2} X_S^{*\top} X_{\gamma, S^c} \delta_{S^c} \\
&\quad + M_S \tfrac{1}{\sqrt{n}\sigma^2} X_S^{*\top} \epsilon - \sqrt{n}\left(I_{p+|S|} - M_S J_S\right) \begin{pmatrix} \hat{\beta}_S^{\text{Lasso}} - \beta_0 \\ \hat{\gamma}_S^{\text{Lasso}} - \frac{\delta_S}{\sqrt{n}} \end{pmatrix}.
\end{aligned} \tag{12}$$

The right hand side of the second line of equation (12) has a very clear interpretation. It consists of a sum of four elements. The first two are related to the local misspecification, the third one is a variance term and the fourth one is a bias term that is shown in Theorem 2 to be $o_p(1)$ if $S_{0,n} \subseteq S$. Before stating our theoretical results and defining the FIC, we describe how to construct the relaxed inverse M_S.

4.2 Nodewise regression

Before stating our theoretical result we briefly describe how we construct the matrix M_S which acts as a relaxed inverse of J_S. We follow the methodology of van de Geer et al. (2014). For each $j \in \{1, \ldots, p + |S|\}$ we compute

$$\hat{\eta}_j = \underset{\eta \in \mathbb{R}^{p+|S|-1}}{\arg\min} \frac{1}{2n} \left\| X_{S,j}^* - X_{S,-j}^* \eta \right\|_2^2 + \lambda_j \|\eta\|_1,$$

where $X_{S,j}^*$ is the j-th column of X_S^* and $X_{S,-j}^* \in \mathbb{R}^{n \times (p+|S|-1)}$ is X_S^* without its j-th column, and we form

$$\hat{A}_S = \begin{bmatrix} 1 & -\hat{\eta}_{1,2} & \cdots & \hat{\eta}_{1,p+|S|} \\ -\hat{\eta}_{2,1} & 1 & \cdots & \hat{\eta}_{2,p+|S|} \\ \vdots & \vdots & \ddots & \vdots \\ -\hat{\eta}_{p+|S|,1} & -\hat{\eta}_{p+|S|,2} & \cdots & \hat{\eta}_{p+|S|,p+|S|} \end{bmatrix}$$

with components of $\hat{\eta}_j$ indexed by $k \in \{1, \ldots, j-1, j+1, \ldots, p+|S|\}$. We define

$$M_S = \hat{T}_S^{-2} \hat{A}_S$$

with $\hat{T}_S^2 = \text{diag}(\hat{\tau}_1^2, \ldots, \hat{\tau}_{p+|S|}^2)$ and $\hat{\tau}_j^2 = \frac{1}{n} \left\| X_{S,j}^* - X_{S,-j}^* \hat{\eta}_j \right\|_2^2 + \lambda_j \|\hat{\eta}_j\|_1$.

4.3 Submodel containing the true active set: theoretical results

In this section, we assume that the submodel S contains the true active $S_{0,n}$ of γ_n. We state the following conditions.

(A1) For the true active set $\{1,\dots,p\} \cup \{p+j : j \in S_{0,n}\}$, the compatibility condition holds for $\hat{\Sigma}_S = \frac{1}{n} X_S^{*\top} X_S^*$ with compatibility constant $\phi_0^2 > 0$, this is for all β and γ satisfying $\|\gamma_{S_{0,n}^c}\|_1 \leq 3(\|\beta\|_1 + \|\gamma_{S_{0,n}}\|_1)$, it holds that $(\|\beta\|_1 + \|\gamma_{S_{0,n}}\|_1)^2 \leq \begin{pmatrix} \beta \\ \gamma_S \end{pmatrix}^\top \hat{\Sigma}_S \begin{pmatrix} \beta \\ \gamma_S \end{pmatrix} (p+s_n)/\phi_0^2$. Furthermore, $\max_j \hat{\Sigma}_{S,j,j} \leq N^2$ for some $0 < N < \infty$.

(A2) For each j we take $\lambda_j = O(\sqrt{\log(p+|S|)/n}$ in the nodewise regression procedure and we have $\hat{\tau}_j \geq C > 0$.

(A3) $s_n = o(n^{1/4})$.

Assumption (A1) is common in the high-dimensional literature, see for example Bühlmann and van de Geer (2011), and assumption (A2) corresponds to (B1) in Bühlmann and van de Geer (2015). (A3) is a sparsity condition, which is the same as assumption (C5).

The following lemma follows from Theorem 2.1 of van de Geer et al. (2014) applied to the model $Y = X_\beta \beta_0 + X_{\gamma,S}\gamma_{n,S} + \epsilon$ (which holds if $S_{0,n} \subseteq S$) under the local misspecification framework $\gamma_n = \gamma_0 + \delta/\sqrt{n}$.

Lemma 3. *Let us consider the linear model (8). Let S be a subset of $\{1,\dots,q_n\}$ such that $S_{0,n} \subseteq S$ and let $t > 0$ be arbitrary. Under conditions (A1), if $\lambda \geq 2N\sigma\sqrt{2(t^2+\log(p+|S|))/n}$ we have:*

$$\begin{pmatrix} \sqrt{n}(\hat{\beta}_S^{\text{desp}} - \beta_0) \\ \sqrt{n}\hat{\gamma}_S^{\text{desp}} \end{pmatrix} \doteq_d \begin{pmatrix} C_S \\ D_S \end{pmatrix} + \Delta_1, \tag{13}$$

with

$$\begin{pmatrix} C_S \\ D_S \end{pmatrix} \sim \mathcal{N}_{p+|S|}\left(\begin{pmatrix} 0_p \\ \delta_S \end{pmatrix}, M_S J_S M_S^\top \right)$$

and

$$\mathbb{P}\left[\|\Delta_1\|_\infty \geq 8\sqrt{n}\left(\max_j \frac{\lambda_j}{\hat{\tau}_j^2}\right) \frac{\lambda(p+s_n)}{\phi_0^2} \right] \leq 2\exp(-t^2),$$

with λ_j and $\hat{\tau}_j^2$ being the tuning parameter and the residual sum of squares of the regression of $X_{S,j}^*$ on $X_{S,-j}^*$ in the nodewise regression procedure.

Using Lemma 3, we can obtain the distribution of the focus estimator.

Theorem 2. *Let consider the linear model (8). Let S be a subset of $\{1,\dots,q\}$ such that $S_{0,n} \subseteq S$ and let $t > 0$ be arbitrary. Under conditions (A1), (A2) and (A3), if $\lambda \geq 2N\sigma\sqrt{2(t^2+\log(p+|S|))/n}$ we have for $\hat{\mu}_S = \mu(\hat{\beta}_S^{\text{desp}}, \hat{\gamma}_S^{\text{desp}}, 0_{|S^c|}^\top)$ and $\mu_{\text{true}} = \mu(\beta_0, \gamma_n)$*

$$\sqrt{n}(\hat{\mu}_S - \mu_{\text{true}}) \doteq_d \Lambda_S + \Delta_2,$$

where

$$\Lambda_S = \left(\frac{\partial \mu}{\partial \theta}\right)^\top C_S + \left(\frac{\partial \mu}{\partial \gamma_S}\right)^\top D_S - \left(\frac{\partial \mu}{\partial \gamma}\right)^\top \delta = \begin{pmatrix} \frac{\partial \mu}{\partial \theta} \\ \frac{\partial \mu}{\partial \gamma} \end{pmatrix}^\top \pi_S^{*\top} M_S \begin{pmatrix} U \\ V_S \end{pmatrix}$$

with $(U^\top, V_S^\top) \sim \mathcal{N}_{p+|S|}(0, J_S)$

and

$$\mathbb{P}\left[\Delta_2 \geq 8K(p+|S|)\sqrt{n}\left(\max_j \frac{\lambda_j}{\hat{\tau}_j^2}\right)\frac{\lambda(p+s_n)}{\phi_0^2}\right] \leq 2\exp(-t^2).$$

Using a regularization parameter λ of order $\sqrt{\log(p+|S|)/n}$ and under assumption (A2), Δ_2 can be neglected if $s_n = o(\sqrt{n}/\{(p+|S|)\log(p+|S|)\})$ which holds thanks to (A3) and the fact that p and S are fixed.

For a model S, under conditions of Theorem 2 and with adequate tuning parameters and sparsity assumption, the limiting distribution Λ_S of $\sqrt{n}(\hat{\mu}_S - \mu_{\text{true}})$ is normal with mean $E(\Lambda_S) = 0$ and variance

$$\text{Var}(\Lambda_S) \quad = \quad \begin{pmatrix} \frac{\partial \mu}{\partial \theta} \\ \frac{\partial \mu}{\partial \gamma} \end{pmatrix}^\top \pi_S^{*\top} M_S J_S M_S^\top \pi_S^* \begin{pmatrix} \frac{\partial \mu}{\partial \theta} \\ \frac{\partial \mu}{\partial \gamma} \end{pmatrix}.$$

It is logical to observe a null bias because we make the assumption that the true active set is included in the considered submodel. The limiting mean squared error is thus

$$\text{MSE}(S,\delta) = \begin{pmatrix} \frac{\partial \mu}{\partial \theta} \\ \frac{\partial \mu}{\partial \gamma} \end{pmatrix}^\top \pi_S^{*\top} M_S J_S M_S^\top \pi_S^* \begin{pmatrix} \frac{\partial \mu}{\partial \theta} \\ \frac{\partial \mu}{\partial \gamma} \end{pmatrix}$$

and the FIC is defined as $\text{FIC}(S,\delta) = \widehat{\text{MSE}(S,\delta)}$.

4.4 Arbitrary submodel

For an arbitrary submodel indexed by S, Lemma 3 does not necessarily hold because we cannot guarantee that all active variables are in the chosen submodel. For the purpose of model selection, an estimator of the mean squared error of the focus is needed. We propose to use the approximations

$$\text{E}\begin{bmatrix} \sqrt{n}(\hat{\beta}_S^{\text{desp}} - \beta_0) \\ \sqrt{n}\hat{\gamma}_S^{\text{desp}} \end{bmatrix} \approx \begin{pmatrix} 0_p \\ \delta_S \end{pmatrix} + M_S \frac{1}{n\sigma^2} X_S^{*\top} X_{\gamma,S^c} \delta_{S^c}; \quad \text{Var}\begin{bmatrix} \sqrt{n}(\hat{\beta}_S^{\text{desp}} - \beta_0) \\ \sqrt{n}\hat{\gamma}_S^{\text{desp}} \end{bmatrix} \approx M_S J_S M_S^\top,$$

based on (12). This leads to the following definition of a high-dimensional FIC for a general submodel S:

$$\text{FIC}(S) = \widehat{\text{MSE}(S)}$$

with

$$\text{MSE}(S) = \begin{pmatrix} \frac{\partial \mu}{\partial \theta} \\ \frac{\partial \mu}{\partial \gamma} \end{pmatrix}^\top \left(B_S' \delta\delta^\top B_S'^t + \pi_S^{*\top} M_S J_S M_S^\top \pi_S^*\right) \begin{pmatrix} \frac{\partial \mu}{\partial \theta} \\ \frac{\partial \mu}{\partial \gamma} \end{pmatrix}$$

and

$$B_S' = \left(\pi_S^{*\top} M_S \begin{pmatrix} J_{01} \\ \pi_S J_{11} \end{pmatrix} - \begin{pmatrix} 0_{p\times q_n} \\ I_{q_n} \end{pmatrix}\right)\left(I_q - \pi_S^\top \pi_S\right).$$

This formula corresponds to (7) if $M_S = J_S^{-1}$.

Note that this corresponds to approximate the distribution of $\begin{pmatrix} \hat{\beta}_S^{\text{desp}} \\ \hat{\gamma}_S^{\text{desp}} \end{pmatrix}$ by the one of

$$\begin{pmatrix} \hat{\beta}_S^{\text{desp}} \\ \hat{\gamma}_S^{\text{desp}} \end{pmatrix} - \left(I_{p+|S|} - M_S J_S\right)\begin{pmatrix} \hat{\beta}_S^{\text{Lasso}} \\ \hat{\gamma}_S^{\text{Lasso}} \end{pmatrix} = M_S \frac{1}{n\sigma^2} X_S^{*\top} Y + \left(I_{p+|S|} - M_S J_S\right)\begin{pmatrix} \beta_0 \\ \delta/\sqrt{n} \end{pmatrix},$$

which can be seen as an *oracle desparsified estimator*.

5 Practical considerations

The results of last two sections for linear models are summarized in Table 1. We observe that the expressions giving the limiting variance of $\hat{\mu}_S$ are very similar. In the high-dimensional submodel case, J_S^{-1} is not available and is thus replaced by $M_S J_S M_S$. Regarding the bias, it is possible to show in both cases that if $S_{0,n} \subset S$ then the bias expression reduces to 0.

In practice, when computing the squared bias, we need to estimate $\delta\delta^\top$. There are several possibilities to do it but none of them produces a consistent estimator. We list here four possibilities. A first natural choice is to use the Lasso estimator $\hat{\delta}^{\text{Lasso}} = \sqrt{n}\hat{\gamma}^{\text{Lasso}}$ where

$$(\hat{\beta}^{\text{Lasso}}, \hat{\gamma}^{\text{Lasso}}) = \arg\min_{\beta,\gamma} \frac{1}{2n} \left\| Y - X^* \begin{pmatrix} \beta \\ \gamma \end{pmatrix} \right\|_2^2 + \lambda \|\beta\|_1 + \lambda \|\gamma\|_1 . \tag{14}$$

A second possibility is to use the more sophisticated adaptive Lasso $\hat{\delta}^{\text{adap}} = \sqrt{n}\hat{\gamma}^{\text{adap}}$ (see Zou, 2006) where

$$(\hat{\beta}^{\text{adap}}, \hat{\gamma}^{\text{adap}}) = \arg\min_{\beta,\gamma} \frac{1}{2n} \left\| Y - X^* \begin{pmatrix} \beta \\ \gamma \end{pmatrix} \right\|_2^2 + \lambda \sum_{j=1}^{p} w_j \beta_j + \lambda \sum_{j=p+1}^{p+q} w_j \gamma_j. \tag{15}$$

with $w_j = \begin{cases} \frac{1}{n^{-1/2}+|\hat{\beta}_j^{\text{Lasso}}|} & \text{for } 1 \le j \le p \\ \frac{1}{n^{-1/2}+|\hat{\gamma}_j^{\text{Lasso}}|} & \text{for } p+1 \le j \le p+q_n. \end{cases}$ This can provide a better estimator of δ in view of asymptotic results of Zou (2006). A third possibility is to use the desparsified estimator of the full model $\hat{\delta}^{\text{desp}} = \sqrt{n}\hat{\gamma}^{\text{desp}}$ defined as

$$\begin{pmatrix} \hat{\beta}^{\text{desp}} \\ \hat{\gamma}^{\text{desp}} \end{pmatrix} = \begin{pmatrix} \hat{\beta}^{\text{Lasso}} \\ \hat{\gamma}^{\text{Lasso}} \end{pmatrix} + M \frac{1}{n\sigma^2} X^{*\top} \left(Y - X^t \begin{pmatrix} \hat{\beta}^{\text{Lasso}} \\ \hat{\gamma}^{\text{Lasso}} \end{pmatrix} \right),$$

where M is a relaxed inverse of the Fisher information matrix $J = \frac{1}{n\sigma^2} X^{*\top} X^*$ obtained by the nodewise regression technique. The fourth possibility follows from Lemma 3 applied to $S = (1, \ldots, q_n)$. Under suitable conditions we have $\hat{\delta}^{\text{desp}} \doteq_d \mathcal{N}_q(\delta, \hat{\Omega}) + o_P(1)$ where $\hat{\Omega} = (MJM)_{-p,-p}$ is obtained by deleting the first p rows and the first p columns of MJM. Thus $\delta^{\text{desp}}\delta^{\text{desp},\top}$ has mean $\delta\delta^\top + \hat{\Omega}$. This leads to a fourth possibility for estimating $\delta\delta^\top$: to use $\hat{\delta}^{\text{desp}}\hat{\delta}^{\text{desp},t} - \hat{\Omega}$. In case this quantity would be negative, it can be truncated to zero. To summarize, we propose the four following ways to estimate $\delta\delta^\top$ in the FIC formula: (1) $\hat{\delta}^{\text{Lasso}}(\hat{\delta}^{\text{Lasso}})^\top$, (2) $\hat{\delta}^{\text{adap}}(\hat{\delta}^{\text{adap}})^\top$, (3) $\hat{\delta}^{\text{desp}}(\hat{\delta}^{\text{desp}})^\top$, (4) $\hat{\delta}^{\text{desp}}(\hat{\delta}^{\text{desp}})^\top - \hat{\Omega}$.

6 Simulation study

We perform a simulation study to illustrate the benefits of the high-dimensional FIC. We consider the linear model $Y_i = X_i\gamma_n + \sigma_\epsilon \epsilon_i$ with $\epsilon_i \sim \mathcal{N}(0,1)$ for $i = 1, \ldots, n$. We consider sample sizes $n = 100$ and

	Low-dimensional submodel	High-dimensional submodel
Estimator of $(\beta_0, \gamma_{n,S})$	least-squares estimator: $\begin{pmatrix}\hat\beta_S^{LS}\\ \hat\gamma_S^{LS}\end{pmatrix} = (X_S^{*\top}X_S^*)^{-1}X_S^{*\top}Y$	desparsified estimator: $\begin{pmatrix}\hat\beta_S^{\text{desp}}\\ \hat\gamma_S^{\text{desp}}\end{pmatrix} = \begin{pmatrix}\hat\beta_S^{\text{Lasso}}\\ \hat\gamma_S^{\text{Lasso}}\end{pmatrix} + M_S\frac{1}{n\sigma^2}X_S^{*\top}\left(Y - X_S^*\begin{pmatrix}\hat\beta_S^{\text{Lasso}}\\ \hat\gamma_S^{\text{Lasso}}\end{pmatrix}\right)$
Estimator $\hat\mu_S$ of μ_{true}	$\mu(\hat\beta_S^{LS}, \hat\gamma_S^{LS}, 0_{\mid S^c})$	$\mu(\hat\beta_S^{\text{desp}}, \hat\gamma_S^{\text{desp}}, 0_{\mid S^c})$
Bias of $\sqrt{n}\hat\mu_S$	$\left(\frac{\partial\mu}{\partial(\beta,\gamma)}\right)^\top\left(\pi_S^{*\top}J_S^{-1}\begin{pmatrix}J_{01}\\ \pi_S J_{11}\end{pmatrix} - \begin{bmatrix}0_{p\times q}\\ I_q\end{bmatrix}\right)\delta$	$\left(\frac{\partial\mu}{\partial(\beta,\gamma)}\right)^\top\left(\pi_S^{*\top}M_S\begin{pmatrix}J_{01}\\ \pi_S J_{11}\end{pmatrix} - \begin{bmatrix}0_{p\times q}\\ I_q\end{bmatrix}\right)(I_q - \pi_S^\top\pi_S)\,\delta$
Variance of $\sqrt{n}\hat\mu_S$	$\left(\frac{\partial\mu}{\partial(\beta,\gamma)}\right)^\top\pi_S^{*\top}J_S^{-1}\pi_S^*\left(\frac{\partial\mu}{\partial(\beta,\gamma)}\right)$	$\left(\frac{\partial\mu}{\partial(\beta,\gamma)}\right)^\top\pi_S^{*\top}M_S J_S M_S^\top\pi_S^*\left(\frac{\partial\mu}{\partial(\beta,\gamma)}\right)$

Table 1: Estimator $\hat\mu_S$ of the focus and its bias and variance for a low-dimensional and a high-dimensional submodel in the context of a linear model.

$n = 200$ and two different possibilities for the dimension q of the paramater γ_n: $q = 80$ and $q = 200$. The case $q = 200$ corresponds to high-dimensional data for which the classical FIC can not be used. We generate the true model according to four scenarios:

- Case 1: $\gamma_n = 10c(1, -1, 1, -1, 1, 0, \ldots, 0)/\sqrt{n}$ and X_i from $\mathcal{N}_q(0, I_q)$ for $i = 1, \ldots, n$.
- Case 2: γ_n as in case 1 and X_i from $\mathcal{N}_q(0, \Sigma)$ for $i = 1, \ldots, n$ with $\Sigma_{jj} = 1$ and $\Sigma_{jk} = 0.5$ for $j \neq k$.
- Case 3: $\gamma_n = 10c(1, -\frac{1}{2}, \frac{1}{3}, -\frac{1}{4}, \ldots, \pm\frac{1}{q})/\sqrt{n}$ and X_i as in case 1
- Case 4: γ_n as in case 3 and X_i as in case 2.

Cases 1 and 2 correspond to sparse models and cases 2 and 4 correspond to models with correlation between variables. The parameter c controls the amplitude of the components of γ_n. We consider three different focuses. The first focus is the prediction $\mu_1(\gamma_n) = X_0\gamma_n$ for a new value X_0 of the covariate vector with the components of X_0 randomly generated from $\mathbb{U}[-1, 1]$. The second focus is the first coefficient of γ_n, that is $\mu_2(\gamma_n) = \gamma_{n,1}$ and the third focus is the last coefficient of γ_n, that is $\mu_3(\gamma_n) = \gamma_{n,q}$. Note that the true value of the last focus is 0 for the sparse settings (cases 1 and 2).

We compare predictions of the focus μ_j for two types of methods: (i) we compute a penalized estimator of γ_n in the full model and make prediction based on this parameter estimate and (ii) we use the high-dimensional FIC as described in Sections 3 and 4. We consider two penalized estimators, the Lasso and the adaptive Lasso, with the tuning paramaters chosen by 10-fold cross-validation. Other tuning parameter choices are possible too. These two penalized estimators are also used to estimate δ in the FIC procedure. We thus obtain four different predictions of μ_j. For the estimation of σ_ϵ^2, we follow the recommendation of Reid et al. (2016) and use $\hat\sigma_\epsilon^2 = \text{RSS}/(n - \widehat{\text{df}})$ with $\widehat{\text{df}}$ the number of non-zero coefficients of the penalized estimator of γ_n.

Because the number of covariates is large, it is computationally impossible to obtain the FIC of every possible submodel. Instead, we propose to use a backward-forward stepwise procedure with two possible starting sets: the empty set and the set $\{j : \hat\delta_j \neq 0\}$ of active components of the estimator of δ. The

two procedures usually converge to two different subsets S_1 and S_2 and we keep the one that gives the smallest FIC value. More refined procedures can be used to improve the selection search. It is for example possible to do some pseudo-exhaustive search by computing the FIC of all submodels upto a certain size d.

In Tables 2 to 4, we report the averaged squared errors of the estimators for the three different focuses over 1000 simulated datasets for different settings. In Table 2 we consider settings with 80 covariates and in Table 3 we increase the number of covariates to 200, obtaining high-dimensional data for which the traditional FIC could not be used. Results for these two tables are similar. We observe that all methods perform well for the third focus $\mu_3 = \gamma_q$. Regarding focuses 1 and 2 we observe that the FIC procedures outperform the penalized estimators for the sparse settings (cases 1 and 2), the ones that are supported by the theory. For non-sparse settings (cases 3 and 4), the different methods are equally competitive. The presence of correlation makes things slightly more complicated.

In Table 4, we compare the sensitivity of the different methods to the standard noise level σ_ϵ. We observe that in the sparse cases, the FIC takes much more advantage of the decrease of the noise level. For $\sigma_\epsilon = 0.25$ the FIC largely outperforms the penalized methods while for $\sigma_\epsilon = 1$ the methods are equally competitive.

We conclude this simulation study by a remark on the size of the models selected by the FIC. We observed in our simulations that the models selected by the FIC procedure are very often of size smaller than 5. It turns out that it is often possible to find a small submodel S whose FIC is smaller than the FIC of S_{true}, the active set of the model having generated the data. On Figure 1, we illustrate this by giving the scatter plot of FIC(S) versus $\hat{\mu}_S$ for every possible submodel of size smaller or equal to 3. The setting is chosen to have many true non-zero coefficients (20) so that we expect the bias to be large for models of size only 3. We also choose a small value of the standard noise ($\sigma_\epsilon = 0.1$) to increase the weight of the squared bias in the FIC expression. We see on the left figure that many of the submodels exhibit large values of FIC but more importantly we also notice on the right figure that for some of the small models (about 3% of them), the FIC value is smaller than the FIC of the true model. For such submodels, the estimator $\hat{\mu}_S$ is very close to the true value (the grey horizontal line). This should be considered one of the strong features of the FIC.

7 Real data example: the riboflavin data

We apply the high-dimensional FIC procedure on the riboflavin data that can be found in the R package *hdi* (Meier et al., 2014). The data contains 71 observations, 4088 predictors (gene expressions) and a response variable measuring the riboflavin production of the Bacillus subtilis bacteria. This dataset has been used by many authors in the high-dimensional literature including van de Geer et al. (2014) and Javanmard and Montanari (2014). We center the response variable and randomly split the data into a training set $(X_{\text{train}}, Y_{\text{train}})$ of size 50 and a testing set $(X_{\text{test}}, Y_{\text{test}})$ of size 21. We then consider the linear model $Y_{\text{train}} = X_{\text{train}}\beta + \epsilon$ with $\epsilon \sim N(0, \sigma_\epsilon^2)$ and the 21 focuses $\mu_j = X_{\text{test}}^j\beta$ for $j = 1, \ldots, 21$.

	$n = 100$, $q = 80$						$n = 200$, $q = 80$					
	$c = 1$			$c = 2$			$c = 1$			$c = 2$		
Focus:	μ_1	μ_2	μ_3	μ_1	μ_2	μ_3	μ_1	μ_2	μ_3	μ_1	μ_2	μ_3
Case 1: sparsity and no correlation												
Lasso	0.023	0.016	0.000	0.125	0.087	0.000	0.003	0.002	0.000	0.021	0.015	0.000
Adap. Lasso	0.022	0.015	0.000	0.132	0.088	0.000	0.002	0.001	0.000	0.020	0.015	0.000
FIC Lasso	0.003	0.001	0.000	0.006	0.001	0.000	0.001	0.000	0.000	0.001	0.000	0.000
FIC Adap. Lasso	0.003	0.001	0.000	0.009	0.001	0.000	0.001	0.000	0.000	0.001	0.000	0.000
Case 2: sparsity and correlation												
Lasso	0.026	0.013	0.000	0.150	0.074	0.000	0.005	0.002	0.000	0.019	0.010	0.000
Adap. Lasso	0.021	0.010	0.000	0.150	0.072	0.000	0.002	0.001	0.000	0.017	0.010	0.000
FIC Lasso	0.008	0.003	0.000	0.028	0.013	0.000	0.003	0.001	0.000	0.004	0.002	0.000
FIC Adap. Lasso	0.007	0.003	0.000	0.030	0.017	0.000	0.002	0.001	0.000	0.003	0.002	0.000
Case 3: no sparsity and no correlation												
Lasso	0.028	0.003	0.000	0.091	0.010	0.001	0.009	0.001	0.000	0.017	0.001	0.000
Adap. Lasso	0.028	0.002	0.000	0.089	0.003	0.001	0.009	0.000	0.000	0.017	0.001	0.000
FIC Lasso	0.026	0.002	0.001	0.080	0.004	0.001	0.009	0.000	0.000	0.016	0.001	0.000
FIC Adap. Lasso	0.028	0.002	0.000	0.088	0.003	0.001	0.010	0.000	0.000	0.017	0.001	0.000
Case 4: no sparsity and correlation												
Lasso	0.038	0.005	0.001	0.114	0.013	0.001	0.015	0.001	0.000	0.024	0.001	0.001
Adap. Lasso	0.039	0.003	0.001	0.112	0.005	0.001	0.015	0.001	0.000	0.025	0.001	0.000
FIC Lasso	0.037	0.003	0.001	0.105	0.005	0.001	0.015	0.001	0.000	0.024	0.001	0.001
FIC Adap. Lasso	0.039	0.002	0.001	0.111	0.005	0.001	0.015	0.001	0.000	0.025	0.001	0.001

Table 2: Averaged squared errors of the estimators for the three different focuses over 1000 simulated datasets and for different settings. μ_1 is a random new observation, $\mu_2 = \gamma_1$ and $\mu_3 = \gamma_q$. The number of covariates is $q = 80$, the standard noise is $\sigma_\epsilon = 0.25$ and c is a parameter controlling the amplitude of the components of γ_n.

	$n = 100$, $q = 200$						$n = 200$, $q = 200$					
	$c = 1$			$c = 2$			$c = 1$			$c = 2$		
Focus:	μ_1	μ_2	μ_3	μ_1	μ_2	μ_3	μ_1	μ_2	μ_3	μ_1	μ_2	μ_3
Case 1: sparsity and no correlation												
Lasso	0.026	0.016	0.000	0.141	0.088	0.000	0.004	0.002	0.000	0.022	0.015	0.000
Adap. Lasso	0.024	0.016	0.000	0.141	0.091	0.000	0.002	0.001	0.000	0.022	0.015	0.000
FIC Lasso	0.003	0.001	0.000	0.006	0.001	0.000	0.001	0.000	0.000	0.001	0.000	0.000
FIC Adap. Lasso	0.003	0.001	0.000	0.010	0.001	0.000	0.001	0.000	0.000	0.001	0.000	0.000
Case 2: sparsity and correlation												
Lasso	0.031	0.014	0.000	0.161	0.074	0.000	0.006	0.003	0.000	0.021	0.011	0.000
Adap. Lasso	0.023	0.011	0.000	0.159	0.074	0.000	0.002	0.001	0.000	0.020	0.010	0.000
FIC Lasso	0.009	0.004	0.000	0.027	0.014	0.000	0.003	0.001	0.000	0.004	0.002	0.000
FIC Adap. Lasso	0.006	0.003	0.000	0.033	0.016	0.000	0.001	0.001	0.000	0.003	0.002	0.000
Case 3: no sparsity and no correlation												
Lasso	0.050	0.006	0.000	0.150	0.018	0.000	0.017	0.001	0.000	0.035	0.002	0.000
Adap. Lasso	0.048	0.003	0.000	0.193	0.008	0.000	0.020	0.001	0.000	0.086	0.002	0.000
FIC Lasso	0.047	0.003	0.000	0.137	0.008	0.000	0.016	0.001	0.000	0.034	0.001	0.000
FIC Adap. Lasso	0.048	0.002	0.000	0.188	0.007	0.000	0.020	0.001	0.000	0.084	0.002	0.000
Case 4: no sparsity and correlation												
Lasso	0.065	0.010	0.000	0.176	0.023	0.000	0.024	0.003	0.000	0.047	0.003	0.000
Adap. Lasso	0.063	0.004	0.000	0.252	0.010	0.000	0.028	0.001	0.000	0.115	0.003	0.000
FIC Lasso	0.062	0.005	0.000	0.162	0.011	0.000	0.023	0.001	0.000	0.046	0.002	0.000
FIC Adap. Lasso	0.062	0.004	0.000	0.242	0.010	0.000	0.028	0.001	0.000	0.111	0.003	0.000

Table 3: Averaged squared errors of the estimators for the three different focuses over 1000 simulated datasets and for different settings. μ_1 is a random new observation, $\mu_2 = \gamma_1$ and $\mu_3 = \gamma_q$. The number of covariates is $q = 200$, the standard noise is $\sigma_\epsilon = 0.25$ and c is a parameter controlling the amplitude of the components of γ_n.

Focus:	μ_1			μ_2			μ_3		
Standard noise: σ_ϵ	1	0.5	0.25	1	0.5	0.25	1	0.5	0.25
Case 1: sparsity and no correlation									
Lasso	0.086	0.024	0.023	0.036	0.013	0.016	0.001	0.000	0.000
Adap. Lasso	0.060	0.015	0.022	0.014	0.007	0.015	0.001	0.000	0.000
FIC Lasso	0.063	0.013	0.003	0.013	0.003	0.001	0.002	0.000	0.000
FIC Adap. Lasso	0.063	0.011	0.003	0.012	0.003	0.001	0.002	0.000	0.000
Case 2: sparsity and correlation									
Lasso	0.166	0.042	0.026	0.066	0.018	0.013	0.002	0.000	0.000
Adap. Lasso	0.135	0.024	0.021	0.029	0.007	0.010	0.003	0.000	0.000
FIC Lasso	0.140	0.029	0.008	0.039	0.009	0.003	0.003	0.001	0.000
FIC Adap. Lasso	0.142	0.024	0.007	0.027	0.006	0.003	0.004	0.001	0.000
Case 3: no sparsity and no correlation									
Lasso	0.125	0.056	0.028	0.040	0.009	0.003	0.001	0.001	0.000
Adap. Lasso	0.137	0.057	0.028	0.016	0.004	0.002	0.002	0.001	0.000
FIC Lasso	0.126	0.054	0.026	0.016	0.005	0.002	0.002	0.001	0.001
FIC Adap. Lasso	0.149	0.059	0.028	0.015	0.004	0.002	0.003	0.001	0.000
Case 4: no sparsity and correlation									
Lasso	0.188	0.086	0.038	0.084	0.018	0.005	0.002	0.001	0.001
Adap. Lasso	0.204	0.088	0.039	0.035	0.008	0.003	0.003	0.001	0.001
FIC Lasso	0.195	0.085	0.037	0.045	0.011	0.003	0.003	0.001	0.001
FIC Adap. Lasso	0.223	0.091	0.039	0.034	0.008	0.002	0.005	0.002	0.001

Table 4: Averaged squared errors of the estimators for the three different focuses over 1000 simulated datasets and for different settings. μ_1 is a random new observation, $\mu_2 = \gamma_1$ and $\mu_3 = \gamma_q$. Parameters are $q = 80$, $n = 100$ and $c = 1$. Results are given for different values of the standard noise: $\sigma_\epsilon = 1, 0.5, 0.25$.

In a first step we compute a Lasso estimator $\hat{\beta}^{\text{Lasso}}$ of β with the tuning paramaters chosen by 10-fold cross-validation and we obtain estimators $\hat{\mu}_j^{\text{Lasso}} = X_{\text{test}}^j \hat{\beta}^{\text{Lasso}}$ of the 21 focuses. In a second step we apply our FIC procedure. For each of the 21 focuses, we search for a submodel that provides a small FIC value. As in the simulation study, we apply a backward-forward stepwise procedure with two possible starting sets: the empty set and the set selected by the Lasso. We denote by S_1^j and S_2^j the sets obtained for the focus j with these two choices of starting sets. We then keep the best of the two by defining $S^j = \arg\min_{S \in \{S_1^j, S_2^j\}} \text{FIC}_j(S)$. We compute the corresponding estimator $\hat{\beta}_{S^j}$ and obtain an estimator $\hat{\mu}_j^{\text{FIC}} = X_{\text{test}}^j \hat{\beta}_{S^j}$ of the focus μ_j. We then compute the mean squared prediction errors $1/21 \sum_{j=1}^{21} (Y_{\text{test}}^j - \hat{\mu}_j)^2$ for $\hat{\mu}_j = \hat{\mu}_j^{\text{Lasso}}$ and $\hat{\mu}_j = \hat{\mu}_j^{\text{FIC}}$. For comparison purpose, we also compute estimators of the focuses with S_1^j and S_2^j. Note that each computation of the FIC takes about one millisecond on a regular computer. Thus, each step of the stepwise procedure takes about four seconds.

The results are reported in Table 5. We observe that the three strategies for performing the FIC optimization are very competitive and all outperform the Lasso (0.180, 0.177 and 0.182 versus 0.235 for

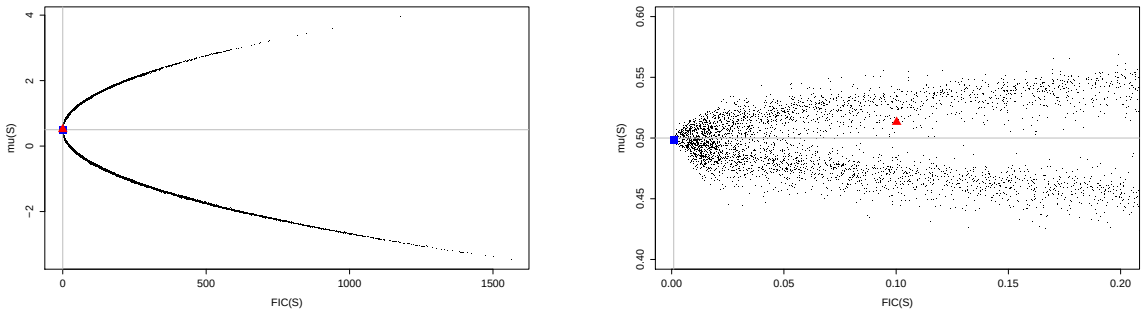
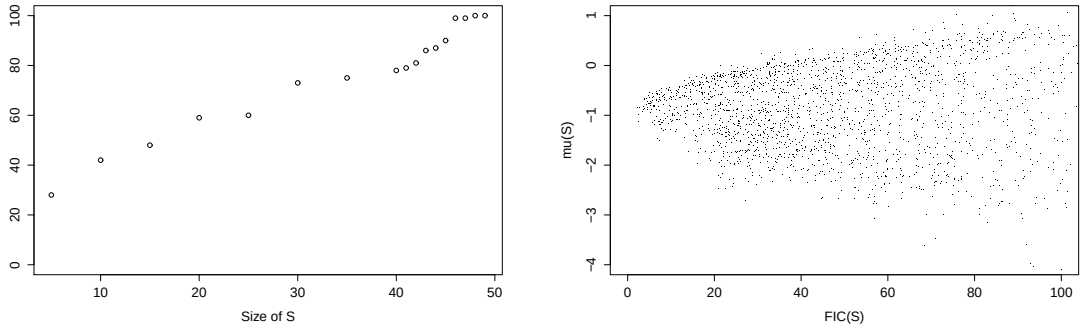


Figure 1: Parameters: $n = 100$, $q = 80$, $p = 0$, $s_0 = 20$, $\sigma_\epsilon = 0.1$, $c = 1$, μ_1. Scatterplots of $\hat{\mu}(S)$ versus FIC(S) for all the 85401 possible models of size smaller or equal to 3. The true δ and the true σ_ϵ are used in the FIC computations. The red triangle corresponds to the true model of size 20 and the blue square corresponds to the model minimizing the FIC amongst the models of size smaller than 3. The right figure is a zoom of the left figure.

the Lasso). We also observe that for two third of the focuses (14 out of 21), the set S_1 was chosen, corresponding to work with the empty set as starting set. In Table 6, we report information about the variables selected by the different procedures. As expected, the set S_1^j is generally smaller (4.7) than the set S_2^j (10.7). Furthermore, we note that only two variables are selected at least three times by FIC 1 and none of them is also selected by the Lasso. Conversely, all the 10 variables selected at least three times by FIC 2 are also selected by the Lasso. To conclude, we observe that the FIC uses for each prediction much fewer variables than the Lasso (6.7 versus 27) but in total the number of different variables used by the FIC for the 21 predictions is much larger than for the Lasso (120 versus 27). This is a key feature of the FIC.



(a) (b)

Figure 2: (a) For different subset sizes, the number of times that the desparsified FIC of Section 4 is smaller than the OLS FIC of Section 3 for 100 random subsets and for the first focus. (b) Scatterplot of $\hat{\mu}(S)$ versus FIC(S) for the desparsified FIC of Section 4 for models going from size 5 to 100.

We end this section by numerically comparing the behaviour of the low-dimensional FIC (OLS) introduced in Section 3 to the high-dimensional FIC (desparsified) presented in Section 4. Summary of the formulas can be found back in Table 1. We refer to these two FIC as OLS FIC and desparsified FIC. We consider the first focus $\mu_1 = X^1_{\text{test}}\beta$ of the real data example and compute the two FIC values for different subsets. Note that these two FIC aim at estimating the mean squared error of two different estimators (see Table 1). More concretely, we consider subsets of the 4088 covariates of size going from 5 to 49. Recall that the training sample size is 50. For each of the considered sizes we consider 100 different subsets.

In Table 7 we give the two FIC values and the two predictions for the first considered subset of each size. Note that the true value of the focus is unknown but we can expect it to be close to $y^1_{\text{test}} = -1.13$. We see that for small subsets the results are very close to each other. When $|S|$ increases, M_S gets further away from J_S^{-1} so that the estimators become more and more different. When $|S|$ is close to the sample size 50 the quality of the OLS estimator deteriorates and it is better to use the desparsified estimator. In Figure 2(a) we count how many times out of 100 random choices of the subsets the desparsified FIC is smaller than the OLS FIC. We observe that from size 20 the desparsified FIC tends to outperfom the OLS FIC. This gets more and more pronounced when $|S|$ gets closer to the sample size. In any case for each S we can always compute both FIC and keep the smaller one. Recall that they refer to different estimators. In Figure 2(b), we give the scatterplot of FIC(S) versus $\hat{\mu}_S$ for 2300 submodels of size $5, 10, \ldots, 40, 41, \ldots, 49, 50, 60, \ldots, 100$ (100 submodels of each size). We observe that the FIC obtained with the desparsified procedure behaves as it is supposed to: the FIC aims at estimating the expected value of $50(\hat{\mu}_S - \mu_{\text{true}})^2$ and we do observe a quadratic shape, which is slightly altered due to the difficulty to estimate δ in this high-dimensional example.

8 Extensions and discussion

8.1 Focused selection for high-dimensional generalized linear models

The results of Section 4 can be extended to high-dimensional generalized linear models (GLM). Let us consider observations $Y_1, \ldots, Y_n$ where Y_i has density $f(y, X_i, \theta_0, \gamma_0 + \delta/\sqrt{n})$ for $i = 1, \ldots, n$ with f from the exponential family of distributions. Consider a high-dimensional submodel S containing the true active set $S_{0,n}$ Let us write $\beta_S = \begin{pmatrix} \theta \\ \gamma_S \end{pmatrix}$ and denote the loss function for an observation (y, x) by $\rho_{\beta_S}(y, x) = -\log f(y, x, \theta, \gamma_S, \gamma_{0,S^c})$. We define the first an second partial derivatives of the loss function as $\dot{\rho}_{\beta_S} = \frac{\partial}{\partial \beta_S}\rho_{\beta_S}$ and $\ddot{\rho}_{\beta_S} = \frac{\partial^2}{\partial \beta_S \partial \beta_S^\top}\rho_{\beta_S}$ and use the following notation: for a function g we write $P_n g = \frac{1}{n}\sum_{i=1}^n g(Y_i, X_i)$. We use the penalized estimator

$$\begin{pmatrix} \hat{\theta}^L_S \\ \hat{\gamma}^L_S \end{pmatrix} = \hat{\beta}^L_S = \underset{\beta_S = (\theta, \gamma_S)}{\arg\min}\ P_n \rho_{\beta_S} + \lambda \left\|\beta_S\right\|_1 .$$

Similarly to van de Geer et al. (2014), we define $\hat{\Sigma}_S = P_n \ddot{\rho}_{\hat{\beta}^L_S}$ and $\hat{M}_S$ as a relaxed inverse of $\hat{\Sigma}_S$ obtained by the Lasso nodewise regression. Note that $\hat{\Sigma}_S$ corresponds to the empirical Fisher matrix estimated in

$(\hat{\theta}_S^L, \hat{\gamma}_S^L, \gamma_{0,S^c})$. We define the following desparsified estimator

$$\begin{pmatrix} \hat{\theta}_S^{\text{desp}} \\ \hat{\gamma}_S^{\text{desp}} \end{pmatrix} = \begin{pmatrix} \hat{\theta}_S^L \\ \hat{\gamma}_S^L \end{pmatrix} - \hat{M}_S P_n \dot{\rho}_{\hat{\theta}_S^L, \hat{\gamma}_S^L}. \tag{16}$$

Using results from van de Geer et al. (2014), we can show that

$$\begin{pmatrix} \hat{\theta}_S^{\text{desp}} - \theta_0 \\ \hat{\gamma}_S^{\text{desp}} - \gamma_{0,S} \end{pmatrix} = \begin{pmatrix} 0 \\ \delta_S \end{pmatrix} - \hat{M}_S P_n \dot{\rho}_{\hat{\beta}_S^{\text{true}}} + o_P(n^{-1/2}) \tag{17}$$

and that $\hat{M}_S P_n \dot{\rho}_{\hat{\beta}_S^L} \dot{\rho}_{\hat{\beta}_S^L}^\top \hat{M}_S^\top$ is a consistent estimator of the variance of $\sqrt{n}(\hat{\theta}_S^{\text{desp},t} - \theta_0^\top, \hat{\gamma}_S^{\text{desp},t} - \gamma_{0,S}^\top)^\top$. This leads to a result similar to Theorem 2 where $M_S J_S M_S$ is replaced by $\hat{M}_S P_n \dot{\rho}_{\hat{\beta}_S^L} \dot{\rho}_{\hat{\beta}_S^L}^\top \hat{M}_S^\top$.

8.2 Model averaging in high-dimensional models

Averaging estimators across several good models is another interesting route, also in the high-dimensional setting. Since estimators in a selected model can be written as model averaged estimators assigning weight one to the estimator in the selected model, and weight zero to all other models, the tool of model averaging is important to study proper post-selection inference.

Let the weighted estimator be obtained in the following way

$$\hat{\mu}_{\text{avg}} = \sum_{S \in \mathcal{A}} w_S(\hat{\delta}) \hat{\mu}_S,$$

where $\hat{\mu}_S = \mu(\hat{\theta}_S, \hat{\gamma}_S, \gamma_{0,S^c})$ and $\mathcal{A}$ is the set of models under consideration for averaging, this does not need to be the set of all possible submodels of the largest available model. The weight for each S, may be deterministic, e.g., assigning equal weight to each of the models, or a predetermined weight that is not data-dependent, or the weights may be data-driven, e.g., $w_S(\hat{\delta}) = I\{S = \arg\min_{S' \in \mathcal{A}} \text{FIC}(S', \hat{\delta})\}$ in the FIC selection case.

Using Lemma 3, or equation (17) in the GLM case, we obtain that the desparsified estimator $\hat{\delta}^{\text{desp}} \doteq_d \tilde{\delta} \sim \mathcal{N}_q(\delta, \Omega)$. Using the joint convergence of the random weights $w_S(\hat{\delta})$ and $\sqrt{n}(\hat{\mu}_S - \mu_{\text{true}})$ for $S \in \mathcal{A}$ to their respective limits, we obtain the following Corollary to Theorem 2.

Corollary 1. *Assume the local misspecification setting. Under the assumptions of Theorem 2, for a set of weights such that* $\sum_{S \in \mathcal{A}} w_S(d) = 1$ *for all* d *and with at most a countable number of discontinuities,*

$$\sqrt{n}\left\{\sum_{S \in \mathcal{A}} w_S(\hat{\delta}) \hat{\mu}_S - \mu_{\text{true}}\right\} \to_d \Lambda_{\text{avg}} = \sum_{S \in \mathcal{A}} w_S(\tilde{\delta}) \Lambda_S.$$

The limiting variable is in case of deterministic weights again normal. For random weights the limit distribution is a sum of products of the random weights and the normal limits Λ_S, which is in general not longer normally distributed. The mean of the limit random variable depends on the random weights $w_S(\tilde{\delta})$ and its correlation with the random variables C_S, D_S,

$$E(\Lambda_{\text{avg}}) = -\left(\frac{\partial \mu}{\partial \gamma}\right)^\top \delta + \sum_{S \in \mathcal{A}} E\left[w_S(\tilde{\delta})\left\{\left(\frac{\partial \mu}{\partial \theta}\right)^\top C_S + \left(\frac{\partial \mu}{\partial \gamma_S}\right)^\top D_S\right\}\right].$$

See also Claeskens et al. (2016) for a calculation of the first two moments of weighted forecasts under weak assumptions avoiding normality.

The use of the proper limiting distribution for model averaged estimators and for post-selection estimators in general, to obtain inference on the averaged estimator is a worthwhile topic of further research, as is the determination of the choice of the weights in the high-dimensional setting.

8.3 Other extensions

Since in high-dimensional models the desparsification is crucial to work with the asymptotic distribution of the focus estimator, which may depend on parameters that are finally set to zero as well as on parameters that are estimated non-zero, extensions to other types of models require first a study of the asymptotic distribution of a desparsified estimator. Gueuning and Claeskens (2016) obtained such a result for partially linear single-index models of the form $Y = \eta(z^\top \alpha) + x^\top \beta + \varepsilon$. In such model both model parts may contain high-dimensional variables, though for theoretical reasons only the dimension of the vector x is allowed to grow with the sample size. An extension of the FIC to such models with focuses that depend on the model parameters α, β, will go along the same lines. In the setting of weighted composite quantile estimation, this would be a generalization of the work on focused selection and model averaging by Xu et al. (2014) to the high-dimensional case.

The present paper also paves the way for extensions of some variations of the FIC to the high-dimensional framework. For example, the *weighted FIC* introduced by Claeskens and Hjort (2008a) aims at selecting a model that performs well for handling a range of similar tasks. Rather than minimizing a mean squared error we could consider selecting a model that minimizes another expected loss function, such as the expected value of a weighted version of the squared error loss,

$$\int n\{\hat{\mu}_S(x) - \mu_{\text{true}}(x)\}^2 d\nu(x),$$

where the dependence of the focus on, say, a covariate vector x is explicitly introduced in the notation. The choice of the weight function ν allows to specify a domain in the covariate space for which a good estimator of μ_{true} is sought. Another use could be to downweight certain regions in order to obtain a more outlier resistant estimator. An estimator of the limit version of this expected weighted loss version leads to the average-focused information criterion for the high-dimensional setting.

To conclude, this paper is the first one to obtain the focused information criterion for high-dimensional data where the parameter length is allowed to grow and even exceed the sample size. Due to the use of a desparsified estimator, the criterion is able to deal with high-dimensional submodels. In addition, we have obtained an alternative formula for FIC in the low-dimensional case that not only deals with the high-dimensionality of the model, but that also is of interest in low-dimensional models by its avoidance to invert the information matrix in a largest model. This paper may pave the way for other applications, estimation methods and models where there is a high-dimensional parameter and where focused selection could bring its benefits of better, targeted, estimators.

9 Proofs

9.1 Proofs for Theorem 1

The proof requires adjustments of the proofs of Lemmas 1, 2 and 3 of Hjort and Claeskens (2003).

Proof of Lemma 1. We use the univariate Lindeberg-Feller theorem (see Serfling (1980) section 1.9) for independant but not i.i.d. random variables and the Cramér-Wold device to obtain the appropriate multivariate normality. To ease the notations, we define the following $p+|S|-$dimensional vectors:

$$\begin{aligned}
W_i &= \begin{pmatrix} U(y_i|x_i) \\ V_S(y_i|x_i) \end{pmatrix} \text{ for } i = 1, \ldots, n, \\
\bar{W}_n &= \frac{1}{n}\sum_{i=1}^{n} W_i, \\
T_i &= W_i - \begin{pmatrix} J_{01}(x_i)\delta/\sqrt{n} \\ \pi_S J_{11}(x_i)\delta/\sqrt{n} \end{pmatrix} \text{ for } i = 1, \ldots, n, \\
\bar{T}_n &= \frac{1}{n}\sum_{i=1}^{n} T_i = \bar{W}_n - \begin{pmatrix} J_{n,01}\delta/\sqrt{n} \\ \pi_S J_{n,11}\delta/\sqrt{n} \end{pmatrix}.
\end{aligned}$$

We want to prove that

$$\sqrt{n}\bar{T}_n \xrightarrow{d} T = \mathcal{N}_{p+|S|}(0, J_S). \tag{18}$$

Let $j \in \{1, \ldots, p+|S|\}$ and let us show first that it holds that

$$\sqrt{n}\bar{T}_n^j \xrightarrow{d} T^j = N(0, (J_S)_{j,j}) \tag{19}$$

with $\bar{T}_n^j$ the j-th component of $\bar{T}_n$.

We have

$$\begin{aligned}
\mathrm{E}[W_i^j] &= \int W_i^j f_0(y|x_i) \left\{1 + V(y|x_i)^\top \delta/\sqrt{n} + R(y|x_i, \delta/\sqrt{n})\right\} dy \\
&= 0 + e_j^\top \begin{pmatrix} J_{01}(x_i)\delta/\sqrt{n} \\ \pi_S J_{11}(x_i)\delta/\sqrt{n} \end{pmatrix} + \int W_i^j f_0(y|x_i) R(y|x_i, \delta/\sqrt{n}) dy.
\end{aligned}$$

By conditions (C1) and (C5), the last term is $o(n^{-1/2})$ so that $\mathrm{E}[T_i^j] = o(n^{-1/2})$.

Furthermore

$$\begin{aligned}
\mathrm{E}[(W_i^j)^2] &= \int (W_i^j)^2 f_0(y|x_i) \left\{1 + V(y|x_i)^\top \delta/\sqrt{n} + R(y|x_i, \delta/\sqrt{n})\right\} dy \\
&= J_S(x_i)_{j,j} + \int (W_i^j)^2 f_0(y|x_i) V(y|x_i)^\top \delta_n/\sqrt{n}\, dy + \int (W_i^j)^2 f_0(y|x_i) R(y|x_i, \delta/\sqrt{n})\, dy \\
&= J_S(x_i)_{j,j} + o(1)
\end{aligned}$$

where the last equality comes from (C2), (C3) and (C5). This implies that $\mathrm{Var}(T_i^j) = J_S(x_i)_{j,j} + o(1)$. Now, applying the Lindeberg-Feller theorem to $\{T_i^j\}_i$ we have

$$\frac{\bar{T}_n^j - \frac{1}{n}\sum_{i=1}^n \mathrm{E}[T_i^j]}{\sqrt{\frac{1}{n^2}\sum_{i=1}^n \mathrm{Var}(T_i^j)}} \xrightarrow{d} \mathcal{N}(0,1)$$

which is equivalent to

$$\frac{\sqrt{n}\bar{T}_n^j}{\sqrt{(J_{n,S})_{j,j}}} \xrightarrow{d} \mathcal{N}(0,1)$$

and because $(J_{n,S})_{j,j}$ converges to $(J_S)_{j,j}$ this is also equivalent to (19). The Lindeberg condition for applying the Lindeberg-Feller theorem requires

$$\frac{1}{n}\sum_{i=1}^n \frac{1}{B_n^2/n} \int_{|t-E(T_i^j)|>\epsilon B_n} \left\{t - E(T_i^j)\right\}^2 dF_i(t) \to 0 \text{ for each } \epsilon > 0$$

with $B_n^2 = \sum_{i=1}^n \mathrm{Var}(T_i^j)$ and $F_i(t)$ the distribution function of T_i^j. This is satisfied if

$$\int_{|T_i^j|>\epsilon\sqrt{n}} (T_i^j)^2 f_{\text{true}}(y|x_i)dy \to 0 \text{ for each } \epsilon > 0$$

which holds thanks to conditions (C2) and (C3). Thus, (19) is proven.

We now apply the Cramér-Wold device that states that

$$\sqrt{n}\,\bar{T}_n \xrightarrow{d} T \quad \text{if and only if} \quad \sqrt{n}\,a^\top \bar{T}_n \xrightarrow{d} a^\top T \;\; \forall a \in \mathbb{R}^{p+|S|}.$$

Let consider an arbitrary $a \in \mathbb{R}^{p+|S|}$. Using (19), it is clear that $\sqrt{n}a^\top\bar{T}_n = \sum_{j=1}^{p+|S|}\sqrt{n}a_j\bar{T}_n^j$ tends to a normal distribution with mean 0 and variance given by $\sum_{j,k=1}^{p+|S|} a_j a_k (J_S)_{j,k} = a^\top J_S a$. Indeed,

$$\begin{aligned}\mathrm{Var}(\sqrt{n}a^\top\bar{T}_n) &= \sum_{j,k=1}^{p+|S|} n\, a_j a_k \mathrm{Cov}(\bar{T}_n^j, \bar{T}_n^k) \\ &= \sum_{j,k=1}^{p+|S|} \frac{1}{n}\sum_{i=1}^n a_j a_k \mathrm{Cov}(T_i^j, T_i^k) = \sum_{j,k=1}^{p+|S|} a_j a_k (J_{n,S})_{j,k} = a^\top (J_{n,S}) a\end{aligned}$$

which tends to $a^\top J_S a$. This implies that $\sqrt{n}\,a^\top\bar{T}_n \xrightarrow{d} a^\top T$ so that (18) holds. □

Proof of Lemma 2. By Lemma 1, it suffices to show that

$$\begin{pmatrix} \sqrt{n}(\hat{\theta}_S - \theta_0) \\ \sqrt{n}(\hat{\gamma}_S - \gamma_{0,S}) \end{pmatrix} \doteq_d J_S^{-1} \begin{pmatrix} \sqrt{n}\bar{U}_n \\ \sqrt{n}\bar{V}_{n,S} \end{pmatrix},$$

which can be done by using traditional arguments for maximum likelihood estimators (see for example Serfling (1980) section 4.2.2). We give here the explicit derivations.

Writing $\hat{\beta} = (\hat{\theta}_S, \hat{\gamma}_S)$ and $\beta_0 = (\theta_0, \gamma_{0,S})$, a Taylor expansion of $\frac{1}{n}\sum_{i=1}^n \frac{\partial}{\partial\beta}\log f(y_i|x_i,\hat{\beta})$ around β_0 gives

$$0 = \begin{pmatrix} \bar{U}_n \\ \bar{V}_{n,S} \end{pmatrix} + \frac{1}{n}\sum_{i=1}^n \frac{\partial^2}{\partial\beta\partial\beta^\top}\log f(y_i|x_i,\beta_0)(\hat{\beta}-\beta_0) + \frac{1}{2}\sum_{j=1}^{p+|S|}(\hat{\beta}-\beta_0)^\top \frac{1}{n}\sum_{i=1}^n \frac{\partial^3}{\partial\beta_j\partial\beta\partial\beta^\top}\log f(y_i|x_i,\tilde{\beta})(\hat{\beta}-\beta_0) \tag{20}$$

with $\tilde{\beta}$ between β_0 and $\hat{\beta}$. By condition (C4), there exists a function $H(x)$ with finite mean such that $\left|\frac{\partial^3}{\partial\beta_j\partial\beta_k\partial\beta_l}\log f(y|x,\beta)\right| \leq H(x)$ for each $1 \leq j,k,l \leq p+|S|$ in a neighbourhood of β_0. Furthermore, because $\tilde{\beta}$ lies between β_0 and $\hat{\beta}$, we can write $\frac{\partial^3}{\partial\beta_j\partial\beta_k\partial\beta_l}\log f(y|x,\beta) = H(x_i)\xi^i_{jkl}$ with $|\xi^i_{jkl}|\leq 1$. Denoting by ξ^i_j the $(p+|S|)\times(p+|S|)$ matrix whose element (k,l) is ξ^i_{jkl}, the last term of (20) can be expressed as

$$\frac{1}{2}(\hat{\beta}-\beta_0)^\top \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{p+|S|} H(x_i)\xi^i_j(\hat{\beta}-\beta_0).$$

We now define $C_n = \frac{1}{n}\sum_{i=1}^n H(x_i)$ and $\xi^* = \frac{1}{C_n}\frac{1}{n}\sum_{i=1}^n\sum_{j=1}^{p+|S|} H(x_i)\xi^i_j$ and see that this last term is also equal to $\frac{1}{2}(\hat{\beta}-\beta_0)^\top C_n\xi^*(\hat{\beta}-\beta_0)$. Note that each component of the matrix ξ^* is smaller than $p+|S|$ in absolute value. Defining $B_n = \frac{1}{n}\sum_{i=1}^n \frac{\partial^2}{\partial\beta\partial\beta^\top}\log f(y_i|x_i,\beta_0)$, the equation (20) can now be rewritten as

$$0 = \begin{pmatrix} \bar{U}_n \\ \bar{V}_{n,S} \end{pmatrix} + B_n(\hat{\beta}-\beta_0) + \frac{1}{2}(\hat{\beta}-\beta_0)^\top C_n\xi^*(\hat{\beta}-\beta_0)$$

or equivalently as

$$\begin{pmatrix} \bar{U}_n \\ \bar{V}_{n,S} \end{pmatrix} = -(B_n + \frac{1}{2}(\hat{\beta}-\beta_0)^\top C_n\xi^*)\,(\hat{\beta}-\beta_0).$$

We observe that B_n converges to J_S, that C_n converges to $\mathrm{E}[H(X)]$ which is finite by (C4), that all the elements of ξ^* are bounded by $p+|S|$ (which is finite) and that $\hat{\beta}$ tends to β_0. All of this implies that $-(B_n + \frac{1}{2}(\hat{\beta}-\beta_0)^\top C_n\xi^*)$ tends to J_S, which ends the proof of Lemma 2.

□

Proof of Theorem 1. Taylor expansions of $\hat{\mu}_S$ and μ_{true} around $\mu_0 = \mu(\theta_0,\gamma_0)$ give

$$\sqrt{n}(\hat{\mu}_S - \mu_{\text{true}}) = \left(\frac{\partial\mu}{\partial(\theta,\gamma_S)}\right)^\top \begin{pmatrix} \sqrt{n}(\hat{\theta}_S-\theta_0) \\ \sqrt{n}(\hat{\gamma}_S-\gamma_{0,S}) \end{pmatrix} - \left(\frac{\partial\mu}{\partial\gamma}\right)^\top \delta + R_1 - R_2$$

with

$$R_1 = \frac{1}{2}n^{-1/2}\delta^\top \frac{\partial^2\mu}{\partial\gamma\partial\gamma^\top}_{|(\theta_0,\tilde{\gamma}_1)}\delta$$

and

$$R_2 = \frac{1}{2}n^{-1/2}\begin{pmatrix} \sqrt{n}(\hat{\theta}_S-\theta_0) \\ \sqrt{n}(\hat{\gamma}_S-\gamma_{0,S}) \end{pmatrix}^\top \frac{\partial^2\mu}{\partial(\theta,\gamma_S)\partial(\theta,\gamma_S)^\top}_{|(\tilde{\theta},\tilde{\gamma}_S)} \begin{pmatrix} \sqrt{n}(\hat{\theta}_S-\theta_0) \\ \sqrt{n}(\hat{\gamma}_S-\gamma_{0,S}) \end{pmatrix}$$

with $\tilde{\gamma}_1$ between γ_0 and $\gamma_0+\delta/\sqrt{n}$ and $(\tilde{\theta},\tilde{\gamma}_S)$ between (θ_0,γ_0) and $(\hat{\theta}_S,\hat{\gamma}_S,\gamma_{0,S^c})$. By (C5) and Lemma 1 2, $R_1 = o(1)$ and $R_2 = o_P(1)$ which implies that

$$\sqrt{n}(\hat{\mu}_S - \mu_{\text{true}}) \;\doteq_d\; \left(\frac{\partial\mu}{\partial(\theta,\gamma_S)}\right)^\top \begin{pmatrix} \sqrt{n}(\hat{\theta}_S-\theta_0) \\ \sqrt{n}(\hat{\gamma}_S-\gamma_{0,S}) \end{pmatrix} - \left(\frac{\partial\mu}{\partial\gamma}\right)^\top \delta.$$

Lemma 2 and algebraic manipulations end the proof. □

9.2 Proofs for Theorem 2

Proof of Lemma 3. Because $S_{0,n} \subseteq S$, it holds that $Y = X_\beta \beta_0 + X_{\gamma,S}\gamma_{n,S} + \epsilon$ with $\gamma_{n,S} = \delta_S/\sqrt{n}$. As conditions of Theorem 2.1 of van de Geer et al. (2014) hold for this linear model we have

$$\begin{pmatrix} \sqrt{n}(\hat{\beta}_S^{\text{desp}} - \beta_0) \\ \sqrt{n}(\hat{\gamma}_S^{\text{desp}} - \delta_S/\sqrt{n}) \end{pmatrix} \doteq_d \; W + \Delta_1,$$

with

$$W \sim \mathcal{N}_{p+|S|}\left(\begin{pmatrix} 0_p \\ 0_{|S|} \end{pmatrix} \; , \; M_S J_S M_S^\top \right)$$

and

$$\mathbb{P}\left[\|\Delta_1\|_\infty \geq 8\sqrt{n}\left(\max_j \frac{\lambda_j}{\hat{\tau}_j^2} \right) \frac{\lambda(p+s_n)}{\phi_0^2} \right] \leq 2\exp(-t^2)$$

which ends the proof of Lemma 3.

Proof of Theorem 2. The proof is straightforward using Lemma 3 and the same reasoning as for Theorem 1.

Acknowledgements

We acknowledge the support of the Fund for Scientific Research Flanders, KU Leuven grant GOA/12/14 and of the IAP Research Network P7/06 of the Belgian Science Policy. The computational resources and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by the Hercules Foundation and the Flemish Government - department EWI.

References

Behl, P., Claeskens, G., and Dette, H. (2014). Focussed model selection in quantile regression. *Statistica Sinica*, 24(2):601–624.

Bühlmann, P. and van de Geer, S. (2011). *Statistics for high-dimensional data*. Springer Series in Statistics. Springer, Heidelberg. Methods, theory and applications.

Bühlmann, P. and van de Geer, S. (2015). High-dimensional inference in misspecified linear models. *Electronic Journal of Statistics*, 9(1):1449–1473.

Claeskens, G. (2012). Focused estimation and model averaging with penalization methods: an overview. *Statistica Neerlandica*, 66(3):272–287.

Claeskens, G., Croux, C., and Van Kerckhoven, J. (2007). Prediction-focused model selection for autoregressive models. *Australian & New Zealand Journal of Statistics*, 49(4):359–379.

Claeskens, G. and Hjort, N. L. (2003). The focused information criterion. *Journal of the American Statistical Association*, 98:900–916.

Claeskens, G. and Hjort, N. L. (2008a). Minimizing average risk in regression models. *Econometric Theory*, 24(02):493–527.

Claeskens, G. and Hjort, N. L. (2008b). *Model selection and model averaging*. Cambridge series in statistical and probabilistic mathematics. Cambridge University Press, Cambridge, New York.

Claeskens, G., Magnus, J., Vasnev, A., and Wang, W. (2016). "The forecast combination puzzle: A simple theoretical explanation". *International Journal of Forecasting*, 32:754 – 762.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360.

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

Gueuning, T. and Claeskens, G. (2016). Confidence intervals for high-dimensional partially linear single-index models. *Journal of Multivariate Analysis*, 149:13–29.

Hjort, N. L. and Claeskens, G. (2003). Frequentist model average estimators. *Journal of the American Statistical Association*, 98(464):879–899.

Hjort, N. L. and Claeskens, G. (2006). Focused information criteria and model averaging for the Cox hazard regression model. *Journal of the American Statistical Association*, 101(476):1449–1464.

Javanmard, A. and Montanari, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research*, 15:2869–2909.

Kim, Y., Kwon, S., and Choi, H. (2012). Consistent model selection criteria on high dimensions. *The Journal of Machine Learning Research*, 13(1):1037–1057.

Luo, S. and Chen, Z. (2013). Extended BIC for linear regression models with diverging number of relevant features and high or ultra-high feature spaces. *Journal of Statistical Planning and Inference*, 143(3):494–504.

Meier, L., Meinshausen, N., and Dezeure, R. (2014). *hdi: High-Dimensional Inference*. R package version 0.1-2.

Pircalabelu, E., Claeskens, G., Jahfari, S., and Waldorp, L. (2016). A focused information criterion for graphical models in fMRI connectivity with high-dimensional data. *Annals of Applied Statistics*, 9(4):2179–2214.

Pircalabelu, E., Claeskens, G., and Waldorp, L. (2015). A focused information criterion for graphical models. *Statistics and Computing*, 25(6):1071–1092.

Reid, S., Tibshirani, R., and Friedman, J. (2016). A study of error variance estimation in lasso regression. *Statistica Sinica*, 26(1):35–67.

Rohan, N. and Ramanathan, T. V. (2011). Order selection in arma models using the focused information criterion. *Australian & New Zealand Journal of Statistics*, 53(2):217–231.

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. John Wiley Sons, Inc.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288.

van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202.

Vansteelandt, S., Bekaert, M., and Claeskens, G. (2012). On model selection and model misspecification in causal inference. *Statistical Methods in Medical Research*, 21(1, SI):7–30.

Wang, H., Li, B., and Leng, C. (2009). Shrinkage tuning parameter selection with a diverging number of

parameters. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(3):671–683.

Wang, H., Li, Y., and Sun, J. (2015). Focused and model average estimation for regression analysis of panel count data. *Scandinavian Journal of Statistics*, 42(3):732–745.

Xu, G., Wang, S., and Huang, J. Z. (2014). Focused information criterion and model averaging based on weighted composite quantile regression. *Scandinavian Journal of Statistics*, 41(2):365–381.

Yang, H., Liu, Y., and Liang, H. (2015). Focused information criterion on predictive models in personalized medicine. *Biometrical Journal*, 57(3):422–440.

Zhang, C.-H. and Zhang, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):217–242.

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

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429.

	Squared prediction errors				FIC values	
Focus number	Lasso	Best FIC	FIC 1	FIC 2	Value FIC 1	Value FIC 2
1	0.189	0.057	0.057	0.076	**1.138**	1.248
2	0.102	0.304	0.324	0.304	1.312	**1.271**
3	0.380	0.241	0.241	0.326	**0.479**	1.983
4	0.036	0.047	0.047	0.035	**0.368**	0.907
5	0.017	0.039	0.039	0.031	**0.475**	1.308
6	0.288	0.118	0.118	0.107	**0.916**	1.008
7	0.003	0.182	0.101	0.182	3.064	**2.899**
8	0.816	0.733	0.733	0.639	**0.305**	0.594
9	0.061	0.047	0.047	0.050	**0.271**	0.525
10	1.706	0.951	0.915	0.951	2.643	**1.989**
11	0.011	0.003	0.003	0.000	**0.344**	0.454
12	0.044	0.001	0.001	0.001	1.191	**1.117**
13	0.635	0.370	0.399	0.370	1.230	**0.806**
14	0.081	0.009	0.009	0.009	**1.224**	1.358
15	0.188	0.130	0.130	0.114	**0.357**	0.500
16	0.049	0.001	0.000	0.001	1.558	**0.863**
17	0.045	0.034	0.034	0.045	**0.220**	0.715
18	0.003	0.002	0.005	0.002	0.670	**0.666**
19	0.021	0.016	0.016	0.028	**0.011**	0.358
20	0.002	0.000	0.000	0.001	**0.354**	0.394
21	0.254	0.502	0.502	0.549	**1.229**	1.470
Average	0.235	0.180	0.177	0.182		

Table 5: Squared prediction errors and FIC values for the 21 focuses of the riboflavin data. The FIC search is done trough a stepwise procedure with as starting set the empty set for FIC 1 and the active set of the Lasso for FIC 2. Best FIC consists is obtain by keeping the submodel that gives the smallest of the two FIC values.

	Lasso	Best FIC	FIC 1	FIC 2
Average number of selected variables	27	6.7	4.6	10.7
Number of variables selected at least once	27	120	77	177
Number of variables selected at least 3 times	27	5	2	10

Table 6: Information on the variables selected for the 21 focuses of the riboflavin data. The FIC search is done trough a stepwise procedure with as starting set the empty set for FIC 1 and the active set of the Lasso for FIC 2. Best FIC consists is obtained by keeping the submodel that gives the smallest of the two FIC values.

$\|S\|$	FIC OLS	FIC desp.	Pred. OLS	Pred. desp.
5	**12.46**	12.56	-0.42	-0.42
10	45.39	**45.01**	0.10	0.10
15	26.23	**26.21**	-0.22	-0.21
20	39.33	**39.18**	-0.18	-0.18
25	**21.43**	21.63	-0.44	-0.43
30	17.14	**15.47**	-0.73	-0.75
35	**21.81**	22.67	-0.67	-0.65
40	**31.90**	38.50	-0.32	-0.33
41	**32.10**	39.29	-0.32	-0.45
42	**32.85**	37.50	-0.29	-0.39
43	67.16	**58.84**	0.42	0.09
44	55.61	**46.93**	0.22	-0.43
45	**92.91**	130.30	0.57	0.70
46	114.14	**85.71**	0.83	0.46
47	121.62	**98.36**	0.47	0.52
48	416.60	**119.20**	1.44	0.77
49	924.01	**114.23**	4.51	0.73
50	n/a	89.99	n/a	0.55
60	n/a	79.35	n/a	0.15
70	n/a	66.49	n/a	-0.26
80	n/a	133.70	n/a	0.73
90	n/a	66.53	n/a	0.02
100	n/a	98.10	n/a	-0.43

Table 7: OLS FIC of Section 3 and desparsified FIC of Section 4 and their corresponding predictions for one random subset for each considered size. The focus is $\mu_1 = X^1_{\text{test}}\beta$ whose true value is unknown but should be close to $y^1_{\text{test}} = -1.13$.

FACULTY OF ECONOMICS AND BUSINESS
Naamsestraat 69 bus 3500
3000 LEUVEN, BELGIË
tel. + 32 16 32 66 12
fax + 32 16 32 67 91
info@econ.kuleuven.be
www.econ.kuleuven.be