ESAIM: PS
March 2006, Vol. 10, p. 164–183
DOI: 10.1051/ps:2006004

ESAIM: Probability and Statistics
www.edpsciences.org/ps

MODEL SELECTION FOR ESTIMATING THE NON ZERO COMPONENTS OF A GAUSSIAN VECTOR

Sylvie Huet[1]

Abstract. We propose a method based on a penalised likelihood criterion, for estimating the number on non-zero components of the mean of a Gaussian vector. Following the work of Birgé and Massart in Gaussian model selection, we choose the penalty function such that the resulting estimator minimises the Kullback risk.

Mathematics Subject Classification. 62G05, 62G09.

Received January 13, 2004. Revised September 28, 2005.

Introduction

The following regression model is considered:

$$\mathbf{X} = \mathbf{m} + \tau\varepsilon, \;\; \varepsilon \sim \mathcal{N}_n(0, I_n),$$

where $\mathbf{X} = (X_1, \dots X_n)^T$ is the vector of observations. The expectation of $\mathbf{X}$, say $\mathbf{m} = (m_1, \dots, m_n)^T$, and the variance τ^2 are unknown. Assuming that some of the components of $\mathbf{m}$ are equal to zero, our objective is to estimate the number of zero components as well as their positions. We propose an estimation method based on a model choice procedure.

We denote by J a subset of $J_n = \{1, 2, \dots, n\}$ with dimension k_J, and we consider the collection $\mathcal{J}$ of all subsets of J_n with dimension less than k_n for some k_n less than n:

$$\mathcal{J} = \{J \subset J_n, k_J \leq k_n\}.$$

Let $\mathbf{x} = (x_1, \dots, x_n)^T$, then for each subset $J \in \mathcal{J}$ we denote by $\mathbf{x}_J$ the vector in $\mathbb{R}^n$ whose component i equals x_i if i belongs to J and 0 if not. We denote by $\|\mathbf{x}\|^2$ the Euclidean distance in $\mathbb{R}^n$ and we set $\|\mathbf{x}\|_n^2 = \|\mathbf{x}\|^2/n$.

For each subset J in the collection $\mathcal{J}$, assuming that $\mathbf{m} = \mathbf{m}_J$, the maximum likelihood estimators of the parameters $(\mathbf{m}, \tau^2)$ are $(\mathbf{X}_J, \|\mathbf{X}_{J^c}\|_n^2)$, where J^c denotes the complement of J in J_n. We thus define a collection of estimators, $(\mathbf{X}_J, \|\mathbf{X}_{J^c}\|_n^2)$ and the problem is now to choose an estimator of $(\mathbf{m}, \tau^2)$ in this collection, or equivalently, to choose the best J in $\mathcal{J}$, say $\widehat{J}$, and to take $(\widehat{\mathbf{m}}, \widehat{\tau}^2) = (\mathbf{X}_{\widehat{J}}, \|\mathbf{X}_{\widehat{J}^c}\|_n^2)$. We associate to each estimator in the collection a risk defined as

$$R(J) = \mathrm{E}\left\{\mathcal{K}_{(\mathbf{m},\tau^2)}\left(\mathbf{X}_J, \|\mathbf{X}_{J^c}\|_n^2\right)\right\},$$

Keywords and phrases. Kullback risk, model selection, penalised likelihood criteria.

[1] INRA, MIA, 78352 Jouy-en-Josas Cedex, France; huet@banian.jouy.inra.fr

where for all $g \in \mathbb{R}^n$ and σ positive, $\mathcal{K}_{(\mathbf{m},\tau^2)}(\mathbf{g},\sigma^2)$ denotes the Kullback-Leibler divergence:

$$\mathcal{K}_{(\mathbf{m},\tau^2)}\left(\mathbf{g},\sigma^2\right) = \frac{n}{2}\left\{\log\left(\frac{\sigma^2}{\tau^2}\right) - 1 + \frac{\tau^2 + \|\mathbf{m}-\mathbf{g}\|_n^2}{\sigma^2}\right\}.$$

The ideal subset J^*, defined as the minimiser of the risk over all the subsets in the collection,

$$R(J^*) = \inf_{J\in\mathcal{J}} R(J),$$

is estimated by a model selection procedure. Namely, we propose to estimate J^* by $\widehat{J}$ that minimises a penalised likelihood criterion defined as follows:

$$\operatorname{crit}(J,\operatorname{pen}) = \frac{n}{2}\log\left(\|\mathbf{X}_{J^c}\|_n^2\right) + \operatorname{pen}(k_J), \tag{1}$$

where pen is a penalty function that depends on k_J. The calculation of the penalty function is based on the following equality:

$$R(J) = \mathcal{K}_{(\mathbf{m},\tau^2)}\left(\mathbf{m}_J, \mathrm{E}\|\mathbf{X}-\mathbf{m}_J\|_n^2\right) + \mathrm{E}\left\{\mathcal{K}_{(\mathbf{m}_J,\mathrm{E}\|\mathbf{X}-\mathbf{m}_J\|_n^2)}\left(\mathbf{X}_J, \|\mathbf{X}_{J^c}\|_n^2\right)\right\}. \tag{2}$$

The first term on the right hand side is analogous to a bias term: it represents the distance between the expectation and the variance of $\mathbf{X}$ under the model J and the parameters $(\mathbf{m},\tau^2)$. It is equal, up to some terms that do not depend on J, to $(n/2)\log(\mathrm{E}\|\mathbf{X}-\mathbf{m}_J\|_n^2)$. The quantity $\mathrm{E}\|\mathbf{X}-\mathbf{m}_J\|_n^2$ being estimated by $\|\mathbf{X}_{J^c}\|_n^2$, the second term on the right hand side is analogous to a *variance term.* The penalty function is calculated so that it compensates both this *variance term* and the bias due to the estimation of $\log(\mathrm{E}\|\mathbf{X}-\mathbf{m}_J\|_n^2)$ by $\log(\|\mathbf{X}_{J^c}\|_n^2)$.

In theorem 2.1 we show that, with probability close to one, $\mathcal{K}_{(\mathbf{m},\tau^2)}\left(\widehat{\mathbf{m}},\widehat{\tau}^2\right)$ is smaller than the minimum over the sets J of

$$K(J) = \mathcal{K}_{(\mathbf{m},\tau^2)}\left(\mathbf{m}_J, \mathrm{E}\|\mathbf{X}-\mathbf{m}_J\|_n^2\right) + \operatorname{pen}(k_J)$$

as soon as the penalty function is larger than a function that can be written as follows:

$$\operatorname{pen}(k) = n\left\{c_1\log\left(\frac{n}{k}\right) + c_2\right\}\frac{k}{n-k} \tag{3}$$

for some constants c_1, c_2. From this result, we deduce in Corollary 2.2 that the expectation of $\mathcal{K}_{(\mathbf{m},\tau^2)}\left(\widehat{\mathbf{m}},\widehat{\tau}^2\right)$ restricted to a set with probability close to 1, is smaller than the minimum of $K(J)$ over the sets J.

This approach has already been proposed by Birgé and Massart [8] for the problem of variable selection in the Gaussian regression model with known variance. In that context, the Kullback risk is the quadratic risk and equals

$$Q(J) = \mathrm{E}\left(\frac{1}{2}\|\mathbf{X}_J - \mathbf{m}\|^2\right) = \frac{1}{2}\left(k\tau^2 + \|\mathbf{m}_J - \mathbf{m}\|^2\right). \tag{4}$$

Minimising the Kullback risk comes to realise the best compromise between the bias term $\|\mathbf{m}-\mathbf{m}_J\|^2$ and the variance term $\tau^2 k_J$. The difference with our work is that we consider the case of an unknown variance, and that we propose a penalty function independent of the error variance. In Theorem 2.3 we show that the quadratic risk associated with the penalised estimator $\widehat{J}$ that minimises $\operatorname{crit}(J,\operatorname{pen})$ defined at equation (1) is bounded above as soon as the penalty is larger than a function of the form given by equation (3).

We compare our penalty function to others for which asymptotic properties have been established, as the AIC procedure and its generalisations. We also compare our estimating procedure to procedures based on the penalisation of the residual sum of squares: the method proposed by Birgé and Massart [8] and the threshold methods.

1. The method

Let $\widehat{J}(\text{pen})$ be the subset in $\mathcal{J}$ that minimises $\text{crit}(J, \text{pen})$ defined in equation (1),

$$\widehat{J}(\text{pen}) = \arg\min_{J\in\mathcal{J}} \text{crit}(J, \text{pen}). \tag{5}$$

The components of $\mathbf{m}$ that are not estimated by 0, correspond to the greatest absolute values of the components of $\mathbf{X}$. Let $\{\ell_1, \ldots, \ell_n\}$ be the order statistics, such that the $X^2_{\ell_i}$ are sortered in the descending order: $X^2_{\ell_1} \geq \ldots \geq X^2_{\ell_n}$, and let J_k be the subset of $\mathcal{J}$ corresponding to the k first order statistics, $J_k = \{\ell_1, \ldots, \ell_k\}$. Then for all subset J in $\mathcal{J}$ with dimension $k_J = k$, we have

$$\frac{n}{2}\log\left(\|\mathbf{X}_{J_k^c}\|^2_n\right) + \text{pen}(k) \leq \frac{n}{2}\log\left(\|\mathbf{X}_{J^c}\|^2_n\right) + \text{pen}(k_J),$$

and the problem reduces to choose k, less than n, that minimises $\text{crit}(J_k, \text{pen})$, say $\widehat{k}$, and to take $\widehat{J} = J_{\widehat{k}}$.

2. Theorems

2.1. Control of the Kullback risk

We denote by $\mathcal{X}_k$ the distribution function of a $\mathcal{X}^2$-variable with k degrees of freedom and by $\mathcal{X}_k^{-1}$ the quantile function.

For $1 < k_n < n$, let us define two collections of positive scalars say $L_0, L_1, \ldots, L_{k_n}$ and $\zeta_0, \zeta_1, \ldots, \zeta_{k_n}$ satisfying the following equalities:

$$\Sigma = \sum_{k=0}^{k_n} \binom{n}{k} \exp(-kL_k) \tag{6}$$

$$\epsilon = \sum_{k=0}^{k_n} \binom{n}{k} \exp(-\zeta_k), \tag{7}$$

and such that

$$\kappa = \epsilon + \frac{3+4\Sigma}{n} < 1, \quad \zeta_0 < \frac{n}{4}, \quad \zeta_k \geq -\log\left\{\mathcal{X}_{n-k}(n)\right\}. \tag{8}$$

Theorem 2.1. *Let $\widehat{J}$ be defined at equation (5), and $\mathcal{K}_{(\mathbf{m},\tau^2)}\left(\widehat{\mathbf{m}}, \widehat{\tau}^2\right)$ the Kullback-Leibler divergence in $(\widehat{\mathbf{m}}, \widehat{\tau}^2) = \left(\mathbf{X}_{\widehat{J}}, \|\mathbf{X}_{\widehat{J}^c}\|^2_n\right)$. If for some constant $C > 1$, and each $k = 0, \ldots, k_n$,*

$$\text{pen}(k) \geq \frac{C}{2} k\left(1 + 2\sqrt{2L_k} + 6L_k\right) \frac{n}{\mathcal{X}^{-1}_{n-k}(\exp(-\zeta_k))}, \tag{9}$$

then

$$\text{pr}\left[\mathcal{K}_{(\mathbf{m},\tau^2)}\left(\widehat{\mathbf{m}}, \widehat{\tau}^2\right) \leq \frac{2C}{C-1} \inf_{J\in\mathcal{J}} \left\{\mathcal{K}_{(\mathbf{m},\tau^2)}\left(\mathbf{m}_J, \tau^2 + \|\mathbf{m}_{J^c}\|^2_n\right) + \text{pen}(k_J)\right\} + K_n(C)\right] \geq 1 - \kappa$$

where

$$K_n(C) = \frac{(3C-1)^2}{2(C-1)^2}\left(\frac{2C(C+1)}{(C-1)^2}\frac{n}{\mathcal{X}^{-1}_{n-k_n}(\exp(-\zeta_{k_n}))} + 1\right)\left(\sqrt{5} + \sqrt{\frac{\log(n)}{n}}\right)^2 \log(n).$$

Let us comment the theorem and choose suitable values for the L_k's and ζ_k's such that κ is small and $K_n(C)$ of smaller order that the penalty.

1-. Choice of the L_k's, $k = 0, \ldots, k_n$. The L_k's must be chosen large enough such that Σ is not too large. Following the discussion in the paper by Birgé and Massart [8] when they consider the variable selection problem in the Gaussian model, we can choose the L_k's independent or dependent of k. This is what they call constant or variable weights strategy.

- Constant weights strategy. If $L_k = L$ for all k, then $\{1 + \exp(-L)\}^{k_n} < \Sigma < \{1 + \exp(-L)\}^n$. Taking $L > \log(n/c)$ for a positive constant c, we get $\Sigma \leq \exp(c)$. If L does not increase with n, then Σ explodes.
- Variable weights strategy. Using Lemma 6 of Barron *et al.* [6], we get

$$\sum_{k=0}^{k_n} \binom{n}{k} \exp(-kL_k) \leq \sum_{k=0}^{k_n} \exp\left[k\left\{1 + \log\left(\frac{n}{k}\right)\right\} - kL_k\right]. \tag{10}$$

Then, if we choose $L_k = 1 + \log(2) + \log(n/k)$, we get that $\Sigma \leq 2$.

2-. Choice of the ζ_k's, $k = 1, \ldots, k_n$. The ζ_k's must be chosen large enough such that ϵ is small. From Inequality (10) (with kL_k replaced by ζ_k), we deduce that if for each $k \geq 1$,

$$\zeta_k = k\left\{1 + 2\log\left(\frac{n}{k}\right)\right\}, \tag{11}$$

then ζ_k satisfies the inequality (8) and

$$\sum_{k=1}^{k_n} \binom{n}{k} \exp(-\zeta_k) \leq \sum_{k=1}^{k_n} \left(\frac{k}{n}\right)^k \leq \frac{\kappa_1}{n} \text{ (see Lem. 7.5)},$$

for some constant κ_1. Therefore, if we choose $\zeta_0 = \kappa_2 \log(n)$ for some constant κ_2, we get $\epsilon \leq (\kappa_1 + \kappa_2)/n$.

3-. The terms $\mathcal{X}_{n-k}^{-1}\{\exp(-\zeta_k)\}$. If we choose ζ_k as above and $k_n = n/2$, thanks to Lemma 7.3, we know that $\mathcal{X}_{n-k}^{-1}\{\exp(-\zeta_k)\} \geq c_1(n-k)$ for some constant c_1. Therefore we can replace in the penalty function $\mathcal{X}_{n-k}^{-1}\{\exp(-\zeta_k)\}$ by $(n-k)$ and the penalty is of order $k\log(n)$. In the same way, the term $n/\mathcal{X}_{n-k_n}^{-1}\{\exp(-\zeta_{k_n})\}$ appearing in $K_n(C)$ is smaller than $2/c_1$. It follows that $K_n(C)$ is of order $\log(n)$.

As explained in the introduction, see equation (2), the proof of the theorem lies on the control of the quantities $\mathcal{K}_{(\mathbf{m}_J, \mathrm{E}\|\mathbf{X}-\mathbf{m}_J\|_n^2)}\left(\mathbf{X}_J, \|\mathbf{X}_{J^c}\|_n^2\right)$ for all $J \in \mathcal{J}$. In fact, we are able to control these quantities (see Lem. 4.3) if we restrict the calculations to the set Ω defined as follows: $\Omega = \Omega_0 \cap \Omega_1$, where

$$\left.\begin{array}{lcl} \Omega_0 & = & \left\{\|X\|^2 \geq n\tau^2\left(1 - 2\sqrt{\zeta_0/n}\right)\right\} \\ \Omega_1 & = & \left\{\forall J \in \mathcal{J}, k_J \geq 1, \|\mathbf{X}_{J^c}\|^2 \geq \tau^2 \mathcal{X}_{n-k_J}^{-1}\left(\exp(-\zeta_{k_J})\right)\right\} \end{array}\right\}. \tag{12}$$

In other words the theorem is shown when, simultaneously over all the subsets $J \in \mathcal{J}$, the quantities $\tau^2/\|\mathbf{X}_{J^c}\|_n^2$ are bounded below by $n/\mathcal{X}_{n-k_J}^{-1}\left\{\exp(-\zeta_{k_J})\right\}$ that appear in the penalty function. It follows that we easily deduce from Theorem 2.1 an upper bound for the expectation of $\mathcal{K}_{(\mathbf{m},\tau^2)}\left(\widehat{\mathbf{m}}, \widehat{\tau}^2\right)$ restricted to the set Ω. Let us remark that if the ζ_k's are chosen as in equation (11) then $\mathrm{pr}(\Omega^c) \leq \epsilon \leq (\kappa_1 + \kappa_2)/n$. Indeed, we have the following inequalities:

$$\begin{array}{rcl} \mathrm{pr}\left(\Omega_1^c\right) & \leq & \displaystyle\sum_{J \in \mathcal{J}} \mathrm{pr}\left\{\|\mathbf{X}_{J^c}\|^2 \leq \tau^2 \mathcal{X}_{n-k_J}^{-1}\left(\exp(-\zeta_{k_J})\right)\right\} \\ & \leq & \displaystyle\sum_{k=1}^{k_n} \binom{n}{k} \mathrm{pr}\left\{\|\mathbf{X}_{J_k^c}\|^2 \leq \tau^2 \mathcal{X}_{n-k}^{-1}\left\{\exp(-\zeta_k)\right\}\right\}. \end{array}$$

Thanks to Lemma 7.2 and to the inequality (29) in Lemma 7.1, we can show that $\mathrm{pr}(\Omega_1^c) \leq \sum_{k=1}^{k_n} \binom{n}{k} \exp(-\zeta_k)$, and $\mathrm{pr}(\Omega_0^c) \leq \exp(-\zeta_0)$. Therefore, $\mathrm{pr}(\Omega^c) \leq \epsilon$. We can now state the corollary.

Corollary 2.2. *Let Ω be defined at equation (12) with ζ_k as in equation(11). Under the assumptions of Theorem* 2.1*,*

$$\mathrm{E}\left[\mathcal{K}_{(\mathbf{m},\tau^2)}\left(\widehat{\mathbf{m}},\widehat{\tau}^2\right)\mathbb{1}_\Omega\right] \quad\leq\quad \frac{2C}{C-1}\left(\inf_{J\in\mathcal{J}}\left\{\mathcal{K}_{(\mathbf{m},\tau^2)}\left(\mathbf{m}_J,\tau^2+\|\mathbf{m}_{J^c}\|_n^2\right)+\mathrm{pen}(k_J)\right\}\right)+r_n(C,\Sigma),$$

where

$$r_n(C,\Sigma)=\frac{\kappa C}{C-1}\left[\frac{(1+\Sigma)C}{C-1}\frac{n}{n-k_n}+1\right]$$

for some constant κ.

From a practical point of view, we want to have in hand a penalty function such that the risk associated with the corresponding estimator is as close as possible to the minimum of the risks associated with the sets J in the collection $\mathcal{J}$. We will use the lower bound for the penalty function given at equation (9) in Theorem 2.1 as the penalty function, taking $L_k = \log(n/k)$, neglecting the term $\sqrt{L_k}$ and replacing $\mathcal{X}_{n-k}^{-1}(\exp(-\zeta_k))$ by $n-k$. Though the results given in Corollary 2.2 are restricted to the set Ω we calculated the best constants c_1, c_2 that occur in the penalty function of the following form:

$$\mathrm{pen}(c_1,c_2)=\frac{n}{2}\left\{c_1\log\left(\frac{n}{k}\right)+c_2\right\}\frac{k}{n-k},$$

such that $\mathrm{E}\left[\mathcal{K}_{(\mathbf{m},\tau^2)}\left(\widehat{\mathbf{m}},\widehat{\tau}^2\right)\right]$ is as close as possible to $R(J^*)$. We find $c_1 = 2$ and $c_2 = 4$, see [14].

2.2. Control of the quadratic risk

Birgé and Massart [8] provided a general approach to model selection *via* penalisation for Gaussian regression with known variance. If we assume that τ is known their result applies to our model. For each $J \in \mathcal{J}$, the Kullback risk for the maximum likelihood estimator of $\mathbf{m}$ when $\mathbf{m} = \mathbf{m}_J$ equals $Q(J)$ defined at equation (4) and the likelihood penalised estimator is defined by $\widehat{J}(\underline{\mathrm{pen}}) = \arg\min_{J\in\mathcal{J}} \underline{\mathrm{crit}}(J, \underline{\mathrm{pen}})$, where

$$\underline{\mathrm{crit}}(J,\underline{\mathrm{pen}})=\frac{n}{2}\|\mathbf{X}_{J^c}\|_n^2+\underline{\mathrm{pen}}(k_J). \tag{13}$$

We call this estimator a RSS-penalised estimator and we denote the penalty function by $\underline{\mathrm{pen}}$ to highlight that the residual sum-of-squares is penalised, not its logarithm.

They showed that if

$$\underline{\mathrm{pen}}(k)\geq C\tau^2(1+\sqrt{2L_k})^2\frac{k}{2},\ \text{for } C>1$$

then

$$\mathrm{E}\left(\frac{1}{2}\left\|\mathbf{m}-\mathbf{X}_{\widehat{J}(\underline{\mathrm{pen}})}\right\|^2\right)\leq C_1\inf_{J\in\mathcal{J}}\left\{\frac{1}{2}\|\mathbf{m}_{J^c}\|^2+\underline{\mathrm{pen}}(k_J)\right\}+r_n$$

where C_1 is some constant that depends on C and r_n a remainder term that is smaller than some constant. Moreover they calculated an upper bound for the penalty term. Namely they showed [9] that, when n tends to infinity and k is of order n^α for $0 < \alpha < 1$, if

$$\underline{\mathrm{pen}}(k)\leq\tau^2k\left[\frac{1}{2}+\log\left(\frac{n}{k}\right)\right],$$

then $\mathrm{E}\left(\left\|\mathbf{m}-\mathbf{X}_{\widehat{J}(\underline{\mathrm{pen}})}\right\|^2\right) \geq C\tau^2 k\log(n)$. The differences between their estimator and ours lie in the value of the maximum likelihood and in the presence of the coefficient τ^2.

The link between the penalised likelihood criterion estimator and the RSS-penalised estimator can be easily done: minimising $\mathrm{crit}(J,\mathrm{pen})$ is equivalent to minimising $\underline{\mathrm{crit}}(J,\underline{\mathrm{pen}})$ with

$$\underline{\mathrm{pen}}(J) = \frac{1}{2}\|X_{J^c}\|^2\left(\exp\frac{2}{n}\mathrm{pen}(k)-1\right). \tag{14}$$

Using this relation, we calculate an upper bound for the quadratic risk associated with our penalised likelihood criterion.

Let $L_0,\ldots,L_{k_n}$ and $\zeta_0,\ldots,\zeta_{k_n}$ satisfying equations (6) and (7).

Theorem 2.3. *Let* $\widehat{J}(\mathrm{pen})$ *be defined at equation (5) and let us assume that the penalty function* pen *satisfies equation (9). Then,*

$$\mathrm{E}\left(\left\|\mathbf{m}-\mathbf{X}_{\widehat{J}(\mathrm{pen})}\right\|^2\right) \leq \frac{4C(C+1)^2}{(C-1)^3}\left(\inf_{J\in\mathcal{J}}\left\{\|m_{J^c}\|^2+2\mathrm{E}(\underline{\mathrm{pen}}(k_J))\right\}+(C+1)\tau^2\Sigma\right)+r_n(\mathbf{m},\epsilon),$$

where $\underline{\mathrm{pen}}$ *is given at equation (14) and*

$$r_n(\mathbf{m},\epsilon) = \|\mathbf{m}\|^2\epsilon + \alpha\tau^2 k_n n^{1/4}\sqrt{\epsilon},$$

for some constant α.

If the ζ_k's satisfy equation (11), and $k_n < \beta n^{1/4}$, then $r_n(\mathbf{m},\epsilon)$ is smaller than some constant.

3. Comparison with other criteria

3.1. The Akaike criterion and its generalisations

Several criteria have been proposed in the literature, the most famous ones being the Akaike and the Schwarz criteria. Their properties have been studied in an asymptotic framework, namely, the number of non zero components of $\mathbf{m}$, denoted by k_0, is fixed and n tends to infinity.

The Akaike criterion [2] with

$$\mathrm{pen}_{AIC}(k) = k$$

is based on an asymptotic expansion of the Kullback-Leibler divergence calculated in the maximum likelihood estimator of $(\mathbf{m},\tau^2)$ on one subset J with dimension k. It can be shown that $(n/2)\log\left(\|\mathbf{X}_{\widehat{J}^c}\|_n^2\right)+k$ is an asymptotically unbiased estimator of $\mathrm{E}\left[\mathcal{K}_{(\mathbf{m},\tau^2)}(\widehat{\mathbf{X}}_J,\|\mathbf{X}_{\widehat{J}^c}\|_n^2)\right]$ (up to some terms that do not depend on k).

It appears that the penalty function proposed by Akaike can be deduced from the penalty function introduced in Theorem 2.1 by neglecting the term $n/\mathcal{X}_{n-k}\{\exp(-\zeta_k)\}$ that is of order $n/(n-k)$ and by choosing the L_k's independent of n and k, say $L_k = L$ for a small value of L. As we have seen before, such a choice of the L_k's leads to very high values of the upper bound of the risk (Σ explodes). At the same time, the penalty function is too small, leading to choose $\widehat{J}$ equal to J_{k_n}, where k_n is the maximum dimension of the sets $J\in\mathcal{J}$.

Several authors tried to correct the over-fitting tendency of the Akaike criterion. The corrected Akaike criterion [15],

$$\mathrm{pen}_{AIC_c} = \frac{n}{2}\frac{n+k}{n-k-2},$$

is based on an approximation of the risk function on one subset J. Let J_0 be the subset of k_0 indices on which the components of $\mathbf{m}$ are different from 0, if J contains J_0, the corrected Akaike criterion is an unbiased estimator of $R(J)$ (up to some terms that do not depend on J).

Rewriting pen_{AIC_c} as follows

$$\mathrm{pen}_{AIC_c}(k) = \frac{n}{n-k-2}(k+1) + \frac{n}{2},$$

and noting that constants do not play any role in the penalty functions, it appears that the corrected Akaike criterion intends to correct the over-fitting tendency of the Akaike criterion, at least for small sample.

The SIC criterion,

$$\mathrm{pen}_{SIC}(k) = \frac{1}{2}k\log(n), \tag{15}$$

was proposed by Schwartz [22] and Akaike [3]. Schwartz derived its penalty function from Bayesian arguments and asymptotic expansions. It is shown by Nishii [18] that if the penalty function is written as $\mathrm{pen}(k) = c_n k$ such that $c_n \to \infty$ and $c_n/n \to 0$, then $\widehat{J}_{c_n} = \widehat{J}(\mathrm{pen})$ converges to J_0 in probability. Moreover the quadratic risk function satisfies the following convergence property: $\mathrm{E}(\|\mathbf{m} - \mathbf{X}_{\widehat{J}_{c_n}}\|_n^2)$ tends to $k_0\tau^2$.

The *AMDL* (for approximate minimum description length) criterion proposed by Rissanen [20]

$$\mathrm{pen}_{AMDL}(k) = \frac{3}{2}k\log(n) \tag{16}$$

was studied by Antoniadis *et al.* [4] for determining the number of nonzero coefficients in the vector of wavelet coefficients.

More generally, the SIC criterion can be generalised by choosing a penalty function of the form $\mathrm{pen}_{a_1,a_2}(k) = k(a_1\log(n) + a_2)$ for some constants a_1, a_2. This criterion corresponds to a minimum Kullback risk criterion, in the case where we adopt a constant weight strategy (namely $L_k = \log(n)$) and where we neglect the correction term $n/(n-k)$.

3.2. Threshold estimators

Another important class of estimators is the class of threshold estimators. The estimator of $\mathbf{m}$ equals $\mathbf{X}_{\widehat{J}}$ where $\widehat{J}$ is defined as the set of indices i in $\{1,\ldots,n\}$ such that the absolute value of X_i is greater than a threshold t. This method is applied for detecting the non-zero coefficients in a vector of independent variables. Precisely, it consists in choosing a decreasing sequence $t(k)$ of positive numbers and comparing the order statistics $X^2_{\ell_1},\ldots,X^2_{\ell_n}$ to $t^2(1),\ldots,t^2(n)$. Then define

$$\left.\begin{array}{lcl} \widehat{k} & = & 0 \text{ if } X^2_{\ell_k} < t^2(k) \;\; \forall k \geq 1 \\ \widehat{k} & = & \max_k \left\{X^2_{\ell_k} \geq t^2(k)\right\} \;\text{ if not} \end{array}\right\}, \tag{17}$$

and choose $\widehat{t} = t(\widehat{k})$. The link between threshold and penalised estimators is done as follows: $\widehat{k}$ is the location of the right most local minimum of the quantity $\underline{\mathrm{crit}}(J_k, \underline{\mathrm{pen}})$ defined at equation (13) by taking the following penalty function:

$$\begin{array}{lcl} \underline{\mathrm{pen}}(0) & = & 0 \\ \underline{\mathrm{pen}}(k) & = & \displaystyle\frac{1}{2}\sum_{l=1}^{k} t^2(l), \text{ if } k \geq 1. \end{array}$$

Conversely the RSS-penalised estimator defined a threshold estimator by setting $t^2(k) = 2\left(\underline{\mathrm{pen}}(k) - \underline{\mathrm{pen}}(k-1)\right)$. The link between the threshold estimator and a logRSS-penalised is done in the same way: the logRSS-penalised estimator is a threshold estimator with

$$t^2(k) = \|\mathbf{X}_{J_k^c}\|^2 \left(\exp\left\{\frac{2}{n}\left[\mathrm{pen}(k) - \mathrm{pen}(k-1)\right]\right\} - 1\right).$$

Conversely the threshold estimator defined a logRSS-penalised estimator by setting

$$\begin{aligned}\mathrm{pen}(0) &= 0\\ \mathrm{pen}(k) &= \frac{n}{2}\sum_{l=1}^{k}\log\left(1+\frac{t^2(l)}{\|\mathbf{X}_{J_l^c}\|^2}\right), \text{ if } k\geq 1\end{aligned}$$

in $\mathrm{crit}(J_k, \mathrm{pen})$ defined at equation (1).

For analysing un-replicated factorial and fractional factorial designs, several authors proposed threshold estimators. See for example Box and Meyer [12], Lenth [21], and Haaland and O'Connell [19]. The idea is to choose a threshold that should provide a powerful testing procedure for identifying non-null effects. Lenth [21] proposed a threshold estimator based on constant $t(k)$'s.

Firstly, he proposed to estimate τ as follows:

$$\widehat{\tau} = 1.5\times \mathrm{median}\{|X_i| \text{ for } |X_i| < 2.5 s_0\},$$

where $s_0 = 1.5\times\mathrm{median}|X_i|$. He showed that, when the effects are sparse (the number of non-zero components is small, and n is large), $\widehat{\tau}$ is a fairly good estimate of τ, that slightly overestimates τ. For small samples, a simulation study was conducted to validate this choice. Several other estimators of τ have been proposed and compared by Haaland and O'Connell [19].

Secondly he defined a simultaneous margin of error with approximately 95% confidence by taking

$$t_{SME} = \widehat{\tau}\, T^{-1}\left(\gamma_n, \frac{n}{3}\right), \tag{18}$$

where $\gamma_n = (1+0.95^{1/n})/2$ and $T^{-1}(\gamma, d)$ denotes the γ-quantile of a student variable with d degrees of freedom. The choice $d = n/3$ comes from the comparison of the empirical distribution of $\widehat{\tau}^2$ to chi-squared distribution. Haaland and O'Connell [19] calculated empirically the critical values of the test of $H_0 : m_i = 0, i = 1,\dots,n$ and showed that the Lenth's procedure is conservative. The penalty function in the RSS-penalised estimator associated with this threshold estimator is defined by $\underline{\mathrm{pen}}_{SME}(k) = \frac{k}{2}t^2_{SME}$. When n is large, $\underline{\mathrm{pen}}_{SME}(k)$ is of order $k\log(n)$.

For the problem of testing simultaneously several hypotheses, Benjamini and Hochberg [7] proposed a procedure that controls the false discovery rate. Precisely, the procedure seeks to ensure that at most a fraction q of the rejected null hypotheses corresponds to false rejections. It corresponds to a threshold estimator with

$$t(k) = \widehat{\tau}\Phi^{-1}\left(1-\frac{qk}{2n}\right) \text{ and } t_{FDR} = t(\widehat{k}), \tag{19}$$

where $\widehat{k}$ is defined by equation (17). Abramovich *et al.* [1] showed that in case of sparse data, and letting $q = q_n$ tending to zero when n tends to infinity, the procedure is asymptotically minimax. It is easy to see that for large n, and small q, $t(k)$ is of order $\sqrt{\log(n/k)}$. Therefore the penalty function $\underline{\mathrm{pen}}_{FDR}(k)$ associated with t_{FDR} is of order $k\log(n/k)$. More precisely, using the well known inequality $\Phi^{-1}(1-u)\leq\sqrt{-2\log(2u)}$, it can be shown that

$$\underline{\mathrm{pen}}_{FDR}(k) \leq \widehat{\tau}^2 k\left\{\log\left(\frac{n}{k}\right) - \log(q) + \frac{1}{k}\sum_{l=1}^{k}\log\left(\frac{k}{l}\right)\right\}.$$

Foster and Stine [13] compared the performances of adaptive variable selection to that obtained by Bayes expert and proposed an approximate empirical Bayes estimator defined as a threshold estimator with

$$t(k) = \widehat{\tau}\sqrt{2\log\left(\frac{n}{k}\right)} \text{ and } t_{FS} = t(\widehat{k}), \tag{20}$$

where $\widehat{k}$ is defined by equation (17). The penalty function $\underline{\text{pen}}_{FS}(k)$ associated with t_{FS} is the following:

$$\underline{\text{pen}}_{FS}(k) = \widehat{\tau}^2 k \left\{ \log\left(\frac{n}{k}\right) + \frac{1}{k}\sum_{l=1}^{k} \log\left(\frac{k}{l}\right) \right\}.$$

In a recent paper, Johnston and Silverman [16] propose a method for estimating the non-zero coefficients of the wavelet transform of an unknown function f. Because the wavelet coefficients of a signal are generally sparse at fine resolution scales and dense at the coarser scales, their method adapts the threshold level by level. It is based on an empirical Bayes approach where the prior distribution of the wavelet coefficient estimators is a mixture of a Dirac distribution at zero and an heavy-tailed distribution with an unimodal symmetric density. They show that the procedure gives good theoretical and practical performances for estimating f.

3.3. Practical issues

The performances of all these criteria, when they are applied to real data set, are compared in an extensive simulation study reported in [14]. We considered the following criteria:

- Criteria based on penalised logarithm of the residual sum of squares
  - The SIC criterion defined at equation (15).
  - The AMDL criterion defined at equation (16).
  - The MKR criterion, that aims at minimising the Kullback risk, defined by

$$\underline{\text{pen}}_{BM}(k) = \widehat{\tau}^2 \left\{ \log\left(\frac{n}{k}\right) + 2 \right\} k.$$

- Threshold estimators or criteria based on penalised residual sum of squares. For these estimators we chose to estimate τ using the estimator given by Lenth [21] that should generally perform well for moderate to large numbers of non-null effects.
  - The SME estimator defined at equation (18).
  - The FDR estimator defined at equation (19) with $q = 0.05$.
  - The FS estimator defined at equation (20).
  - The estimator proposed by Birgé and Massart using the penalty function given in [14]:

$$\underline{\text{pen}}_{BM}(k) = \widehat{\tau}^2 \left\{ \log\left(\frac{n}{k}\right) + 2 \right\} k.$$

The conclusions are the following: the behaviour of AMDL and SIC methods depends on k_0, and is very bad in some cases. The FS method gives good results only when $k_0 = 0$ and overestimates k_0 in other cases. This was already noticed by several authors, see for example [19]. The MKR, FDR and BM methods give similar results. We note that the BM method tends to overestimate k_0. When n is small (for example $n = 20$ or $n = 100$) and k_0 is small, the FDR method tends to overestimate k_0.

4. Proof of Theorem 2.1

Proposition 4.1. *Let $\widehat{k}$ be the dimension of $\widehat{\mathcal{J}}_{\text{pen}}$. Then for all $J \in \mathcal{J}$ we get the following inequality: for all $\xi, \eta > 0$, and $0 < \theta < 1$, with probability greater than $1 - \varepsilon - (3 + 4\Sigma)\exp(-\xi)$*

$$\begin{aligned}(1-\theta)\mathcal{K}_{(\mathbf{m},\tau^2)}\left(\mathbf{X}_{\widehat{\mathcal{J}}_{\text{pen}}}, \|\mathbf{X}_{\widehat{\mathcal{J}}^c(\text{pen})}\|_n^2\right) \leq\ & (1+\theta)\mathcal{K}_{(\mathbf{m},\tau^2)}\left(\mathbf{m}_J, \tau^2 + \|\mathbf{m}_{J^c}\|_n^2\right) \\ & + \left\{ \text{pen}(k_J) + \frac{1}{\theta} k_J L_{k_J} - \text{pen}(\widehat{k}) + q(\widehat{k}, \eta, \theta) \right\} \\ & + \frac{1}{\theta} c(\xi) C(k_n, \eta),\end{aligned}$$

where

$$\begin{aligned} q(k,\eta,\theta) &= \frac{n}{2}\frac{1+\eta}{\theta}\frac{k}{\mathcal{X}^{-1}_{n-k}(\exp(-\zeta_k)}\left(1+2\sqrt{2L_k}+6L_k\right) \\ C(k_n,\eta) &= 2(1+\eta^{-1})\frac{n}{\mathcal{X}^{-1}_{n-k_n}(\exp(-\zeta_{k_n}))}+1 \\ c(\xi) &= \xi\left(5+\sqrt{\frac{\xi}{n}}+2\frac{\xi}{n}\right). \end{aligned}$$

Using this proposition we show the theorem as follows. Taking $\mathrm{pen}(k) \geq q(k,\theta,\eta)$, it follows immediately that $q(\widehat{k},\eta,\theta) - \mathrm{pen}(\widehat{k})$ is negative and that

$$\frac{1}{\theta}kL_k < \frac{\mathrm{pen}(k)}{3(1+\eta)}$$

thanks to Lemma 7.3. Therefore

$$\begin{aligned} \mathcal{K}_{(\mathbf{m},\tau^2)}\left(\mathbf{X}_{\widehat{\mathcal{J}}_{\mathrm{pen}}}, \|\mathbf{X}_{\widehat{\mathcal{J}}^c(\mathrm{pen})}\|_n^2\right) &\leq \frac{1+\theta}{1-\theta}\left\{\mathcal{K}_{(\mathbf{m},\tau^2)}\left(\mathbf{m}_J, \tau^2+\|\mathbf{m}_{J^c}\|_n^2\right)\right. \\ &\quad \left. + \frac{3\eta+4}{3(1+\eta)}\frac{1}{1+\theta}\mathrm{pen}(k_J) + \frac{c(\xi)C(k_n,\eta)}{\theta(1+\theta)}\right\}, \end{aligned}$$

is true with probability greater than $\varepsilon + (3+4\Sigma)\exp(-\xi)$.

Finally, the theorem is proved by taking $\theta = (C+1)/(3C-1)$, $\eta = \theta C - 1$, for some $C > 1$ and $\xi = \log(n)$.

4.1. Proof of Proposition 4.1

Let $L(\mathbf{X},\mathbf{g},\sigma^2)$ be the likelihood of $\mathbf{X}$ calculated in $(\mathbf{g},\sigma^2)$. By definition of $\widehat{\mathcal{J}}_{\mathrm{pen}}$, we have the following inequality:

$$\forall J \in \mathcal{J},\ \dim(J) = k_J,\ \ -L(X, \mathbf{X}_{\widehat{\mathcal{J}}_{\mathrm{pen}}}, \|\mathbf{X}_{\widehat{\mathcal{J}}^c(\mathrm{pen})}\|_n^2) + \mathrm{pen}(\widehat{k}) \leq -L(X,\mathbf{X}_J, \|\mathbf{X}_{J^c}\|_n^2) + \mathrm{pen}(k_J).$$

Because $L(X,\mathbf{X}_J,\|\mathbf{X}_{J^c}\|_n^2)$ is the maximum of the log-likelihood when $\mathbf{m} = \mathbf{m}_J$, we get

$$-L(X, \mathbf{X}_{\widehat{\mathcal{J}}_{\mathrm{pen}}}, \|\mathbf{X}_{\widehat{\mathcal{J}}^c(\mathrm{pen})}\|_n^2) + \mathrm{pen}(\widehat{k}) \leq -L(X,\mathbf{m}_J,\tau^2+\|\mathbf{m}_{J^c}\|_n^2) + \mathrm{pen}(k_J). \tag{21}$$

Let us define the function Φ as

$$\Phi(g,\sigma^2) = \frac{\|X-g\|_n^2}{\sigma^2} - \mathrm{E}\left(\frac{\|X-g\|_n^2}{\sigma^2}\right).$$

By simple calculations, we get

$$-L(X,g,\sigma^2) = \frac{n}{2}\Phi(g,\sigma^2) + \mathcal{K}_{(\mathbf{m},\tau^2)}(g,\sigma^2) - \mathrm{E}\{L(X,\mathbf{m},\tau^2)\},$$

and the inequality (21) is equivalent to

$$\left\{\begin{aligned} \mathcal{K}_{(\mathbf{m},\tau^2)}(\mathbf{X}_{\widehat{\mathcal{J}}_{\mathrm{pen}}}, \|\mathbf{X}_{\widehat{\mathcal{J}}^c(\mathrm{pen})}\|_n^2) \leq\ & \mathcal{K}_{(\mathbf{m},\tau^2)}(\mathbf{m}_J,\tau^2+\|\mathbf{m}_{J^c}\|_n^2) + \mathrm{pen}(k_J) - \mathrm{pen}(\widehat{k}) \\ & + \tfrac{n}{2}\Phi(\mathbf{m}_J,\tau^2+\|\mathbf{m}_{J^c}\|_n^2) - \tfrac{n}{2}\Phi(\mathbf{X}_{\widehat{\mathcal{J}}_{\mathrm{pen}}}, \|\mathbf{X}_{\widehat{\mathcal{J}}^c(\mathrm{pen})}\|_n^2). \end{aligned}\right. \tag{22}$$

Now, $\Phi(\mathbf{m}_J, \tau^2 + \|\mathbf{m}_{J^c}\|_n^2) - \Phi(\mathbf{X}_{\widehat{J}_{\text{pen}}}, \|\mathbf{X}_{\widehat{J}^c(\text{pen})}\|_n^2)$ is split up into three parts:

$$\Phi(\mathbf{m}_J, \tau^2 + \|\mathbf{m}_{J^c}\|_n^2) - \Phi(\mathbf{X}_{\widehat{J}_{\text{pen}}}, \|\mathbf{X}_{\widehat{J}^c(\text{pen})}\|_n^2) = D_1(\widehat{J}_{\text{pen}}) + D_2(\widehat{J}_{\text{pen}}) - D_2(J),$$

where

$$\begin{array}{rcl} D_1(J) & = & \Phi(\mathbf{m}_J, \tau^2 + \|\mathbf{m}_{J^c}\|_n^2) - \Phi(\mathbf{X}_J, \|\mathbf{X}_{J^c}\|_n^2) \\ D_2(J) & = & \Phi(\mathbf{m}, \tau^2) - \Phi(\mathbf{m}_J, \tau^2 + \|\mathbf{m}_{\overline{J}}\|_n^2). \end{array}$$

The two following lemmas give upper bounds for $D_1(\widehat{J}_{\text{pen}})$ and $D_2(\widehat{J}_{\text{pen}})$. They are shown in Sections 4.2 and 4.3.

Lemma 4.2. *Let $0 < \theta < 1$, let $(L_k, k = 0, \ldots, k_n)$ and $\Sigma > 0$ satisfying equations (6), then for all $\xi > 0$*

$$\text{pr}\left[\sup_{J \in \mathcal{J}} \left\{ \frac{n}{2} D_2(J) - \theta \mathcal{K}_{(\mathbf{m},\tau^2)}(\mathbf{m}_J, \tau^2 + \|\mathbf{m}_{J^c}\|_n^2) - \frac{1}{\theta} k_J L_{k_J} \right\} \geq \frac{C_2(\xi)}{\theta}\right] \leq (1+\Sigma)\exp(-\xi),$$

where

$$C_2(\xi) = \xi\left(2 + \frac{\xi}{n} + 2\sqrt{\frac{\xi}{n}}\right).$$

Lemma 4.3. *Let $0 < \theta < 1$, $\eta > 0$, let $(L_k, 0 = 1, \ldots, k_n)$ and $\Sigma > 0$ satisfying equations (6), let $(\zeta_k, k = 0, \ldots, k_n)$ and $\varepsilon > 0$ satisfying (7), and let Ω defined at equation (12). For each $k = 0, \ldots, k_n$, let $q_1(k)$ and $C_1(\xi)$ be defined as*

$$\begin{array}{rcl} q_1(0) & = & 0 \\ q_1(k) & = & \dfrac{n}{2} \dfrac{k}{\mathcal{X}_{n-k}^{-1}(\exp(-\zeta_k))} \left(1 + 2\sqrt{2L_k} + 4L_k\right) \quad \textit{if } k \geq 1 \\ C_1(\xi) & = & \xi\left(5 + \dfrac{\xi}{n} + 2\sqrt{\dfrac{\xi}{n}}\right). \end{array}$$

For all $J \in \mathcal{J}$, let $Z(J)$ be defined as follows:

$$Z(J) = \frac{n}{2} D_1(J) - \theta \mathcal{K}_{(\mathbf{m}_J, \tau^2 + \|\mathbf{m}_{J^c}\|_n^2)}(\mathbf{X}_J, \|\mathbf{X}_{J^c}\|_n^2) - \frac{1+\eta}{\theta} q_1(k_J).$$

Then for all $\xi > 0$,

$$\text{pr}\left[\sup_{J \in \mathcal{J}} \{Z(J)\} \mathbb{1}_\Omega \geq \frac{1+\eta^{-1}}{\theta} \frac{n}{\mathcal{X}_{n-k_n}^{-1}\{\exp(-\zeta_{k_n})\}} C_1(\xi)\right] \leq (1+2\Sigma)\exp(-\xi).$$

Now, applying these lemmas to equation (22), and noting that

$$\mathcal{K}_{(\mathbf{m},\tau^2)}(\mathbf{X}_J, \|\mathbf{X}_{J^c}\|_n^2) = \mathcal{K}_{(\mathbf{m},\tau^2)}(\mathbf{m}_J, \tau^2 + \|\mathbf{m}_{J^c}\|_n^2) + \mathcal{K}_{(\mathbf{m}_J, \tau^2 + \|\mathbf{m}_{J^c}\|_n^2)}(\mathbf{X}_J, \|\mathbf{X}_{J^c}\|_n^2),$$

we get that with probability greater than $1-(2+3\Sigma)\exp(-\xi)$,

$$\begin{aligned}
\mathcal{K}_{(\mathbf{m},\tau^2)}\left(\mathbf{X}_{\widehat{J}_{\text{pen}}},\|\mathbf{X}_{\widehat{J^c}_{\text{pen}}}\|_n^2\right)\mathbb{1}_\Omega \quad \leq \quad & \mathcal{K}_{(\mathbf{m},\tau^2)}\left(\mathbf{m}_J,\tau^2+\|\mathbf{m}_{J^c}\|_n^2\right)\mathbb{1}_\Omega \\
&+\left\{\text{pen}(k_J)-\text{pen}(\widehat{k})\right\}\mathbb{1}_\Omega \\
&+\theta\mathcal{K}_{(\mathbf{m},\tau^2)}\left(\mathbf{X}_{\widehat{J}_{\text{pen}}},\|\mathbf{X}_{\widehat{J^c}(\text{pen})}\|_n^2\right)\mathbb{1}_\Omega \\
&+\left\{\frac{1+\eta}{\theta}q_1(\widehat{k})+\frac{1}{\theta}\widehat{k}L_{\widehat{k}}\right\}\mathbb{1}_\Omega \\
&+\left\{\frac{\eta+1}{\eta}\frac{nC_1(\xi)}{\theta\mathcal{X}_{n-k_n}^{-1}\{\exp(-\zeta_{k_n})\}}+\frac{C_2(\xi)}{\theta}\right\}\mathbb{1}_\Omega \\
&-\frac{n}{2}D_2(J)\mathbb{1}_\Omega.
\end{aligned}$$

Let us note that thanks to Assumption (8), $\mathcal{X}_{n-k}^{-1}\{\exp(-\zeta_k)\}\leq n$. Therefore, we have the following inequalities:

$$\begin{aligned}
q(k,\eta,\theta) \quad &\geq \quad \frac{1+\eta}{\theta}q_1(k)+\frac{1}{\theta}kL_k \\
C(k_n,\eta)c(\xi) \quad &\geq \quad (1+\eta^{-1})\frac{n}{\mathcal{X}_{n-k_n}^{-1}\{\exp(-\zeta_{k_n})\}}C_1(\xi)+C_2(\xi),
\end{aligned}$$

and Proposition 4.1 is shown.

4.2. **Proof of lemma 4.2**

$$\begin{aligned}
D_2(J) \quad &= \quad \Phi(\mathbf{m},\tau^2)-\Phi(\mathbf{m}_J,\tau^2+\|\mathbf{m}_{J^c}\|_n^2) \\
&= \quad \frac{\|\varepsilon\|_n^2}{\tau^2}-\frac{\|\mathbf{m}_{J^c}+\varepsilon\|_n^2}{\tau^2+\|\mathbf{m}_{J^c}\|_n^2} \\
&= \quad \frac{1}{n}\frac{\|\mathbf{m}_{J^c}\|_n^2}{\tau^2+\|\mathbf{m}_{J^c}\|_n^2}\left(\frac{n}{\tau^2}\|\varepsilon\|_n^2-n\right)-\frac{\tau\|\mathbf{m}_{J^c}\|_n}{\sqrt{n}}\frac{2}{\tau^2+\|\mathbf{m}_{J^c}\|_n^2}\frac{\sqrt{n}<\mathbf{m}_{J^c},\varepsilon>_n}{\tau\|\mathbf{m}_{J^c}\|_n}. \qquad (23)
\end{aligned}$$

The proof is divided into two steps.

First step.

Let

$$d(J)=2(\sqrt{n\xi}+\xi)\frac{1}{n}\frac{\|\mathbf{m}_{J^c}\|_n^2}{\tau^2+\|\mathbf{m}_{J^c}\|_n^2}+\frac{\tau\|\mathbf{m}_{J^c}\|_n}{\sqrt{n}}\frac{2}{\tau^2+\|\mathbf{m}_{J^c}\|_n^2}\sqrt{2(\xi+k_JL_{k_J})},$$

we show that

$$\text{pr}\left(\exists J/D_2(J)\geq d(J)\right)\leq\exp(-\xi)(1+\Sigma). \qquad (24)$$

Using equations (28) of Lemma 7.1, we have that for all $\xi>0$

$$\text{pr}\left(\frac{n}{\tau^2}\|\varepsilon\|_n^2-n\geq 2\sqrt{n\xi}+2\xi\right)\leq\exp(-\xi).$$

Using the well-known inequality for a Gaussian variable:

$$\forall x>0,\text{pr}(Z\geq x)\leq\exp(-x^2/2)\text{ where }Z\sim\mathcal{N}(0,1), \qquad (25)$$

we get that

$$\mathrm{pr}\left\{ \frac{\sqrt{n} < \mathbf{m}_{J^c}, \varepsilon >_n}{\tau \|\mathbf{m}_{J^c}\|_n} \leq -\sqrt{2(\xi + L_{k_J} k_J)} \right\} \leq \exp(-\xi - L_{k_J} k_J).$$

Equation (24) follows immediately.

Second step.

It remains to make the link with $\mathcal{K}_{(\mathbf{m},\tau^2)}(\mathbf{m}_J, \tau^2 + \|\mathbf{m}_{J^c}\|_n^2)$. Firstly we calculate an upper bound for $\mathcal{K}$ using equation (33) in Lemma 7.4:

$$\begin{aligned}
\frac{2}{n} K_{(\mathbf{m},\tau^2)}(m_J, \tau^2 + \|m_{J^c}\|_n^2) &= -\log\left(\frac{\tau^2}{\tau^2 + \|\mathbf{m}_{J^c}\|_n^2} \right) \\
&\geq 1 - \frac{1}{1 + \frac{\|\mathbf{m}_{J^c}\|_n^2}{\tau^2}} + \frac{1}{2}\left(\frac{\frac{\|\mathbf{m}_{J^c}\|_n^2}{\tau^2}}{1 + \frac{\|\mathbf{m}_{J^c}\|_n^2}{\tau^2}} \right)^2 \\
&\geq \frac{\|\mathbf{m}_{J^c}\|_n^2}{\tau^2 + \|\mathbf{m}_{J^c}\|_n^2} + \frac{1}{2}\left(\frac{\|\mathbf{m}_{J^c}\|_n^2}{\tau^2 + \|\mathbf{m}_{J^c}\|_n^2} \right)^2 \\
&\geq \frac{\tau^2 \|\mathbf{m}_{J^c}\|_n^2}{(\tau^2 + \|\mathbf{m}_{J^c}\|_n^2)^2} + \frac{1}{2}\left(\frac{\|\mathbf{m}_{J^c}\|_n^2}{\tau^2 + \|\mathbf{m}_{J^c}\|_n^2} \right)^2.
\end{aligned}$$

Secondly, using that for all $\theta > 0$ $2ab \leq a^2\theta + b^2/\theta$ we get

$$d(J) \leq \theta \frac{2}{n} K_{(\mathbf{m},\tau^2)}(\mathbf{m}_J, \tau^2 + \|\mathbf{m}_{J^c}\|_n^2) + \frac{2}{\theta}\frac{L_J k_J}{n} + \frac{1}{\theta}\frac{2\xi}{n}\left(2 + \frac{\xi}{n} + 2\sqrt{\frac{\xi}{n}} \right).$$

4.3. **Proof of Lemma 4.3**

For any vector $g \in \mathbb{R}^n$, and $\sigma^2 > 0$, let us define

$$Z_J(g, \sigma^2) = \Phi(\mathbf{m}_J, \tau^2 + \|\mathbf{m}_{J^c}\|_n^2) - \Phi(g_J, \sigma^2).$$

We have $D_1(J) = Z_J(\widehat{\mathbf{X}}_{\widehat{J}}, \|\mathbf{X}_{\widehat{J}^c}\|_n^2)$ and

$$Z_J(g, \sigma^2) = Z_{1,J}(\sigma^2) + Z_{2,J}(\sigma^2) + Z_{3,J}(g, \sigma^2)$$

where

$$\begin{aligned}
Z_{1,J}(\sigma^2) &= \tau^2\sqrt{\frac{2}{n}}\left(\frac{1}{\tau^2 + \|\mathbf{m}_{J^c}\|_n^2} - \frac{1}{\sigma^2} \right) \frac{n\|\varepsilon\|_n^2 \tau^2 - n}{\sqrt{2n}} \\
Z_{2,J}(\sigma^2) &= 2\frac{\tau \|\mathbf{m}_{J^c}\|_n}{\sqrt{n}}\left(\frac{1}{\tau^2 + \|\mathbf{m}_{J^c}\|_n^2} - \frac{1}{\sigma^2} \right) \frac{\sqrt{n}\langle \varepsilon, \mathbf{m}_{J^c} \rangle_n}{\tau \|\mathbf{m}_{J^c}\|_n} \\
Z_{3,J}(g, \sigma^2) &= -2\frac{\tau \|\mathbf{m}_J - g_J\|_n}{\sqrt{n}\sigma^2} \frac{\sqrt{n}\langle \varepsilon, \mathbf{m}_J - g_J \rangle_n}{\tau \|\mathbf{m}_J - g_J\|_n}.
\end{aligned}$$

The proof is divided into three steps.

First step.

We calculate an upper bound for $D_1(J)$. Let

$$\begin{aligned} d_1(J) &= 2\left|\frac{1}{\tau^2+\|\mathbf{m}_{J^c}\|_n^2}-\frac{1}{\|\mathbf{X}_{J^c}\|_n^2}\right|\left\{\left(\sqrt{\frac{\xi}{n}}+\frac{\xi}{n}\right)\tau^2+\sqrt{\frac{2(\xi+k_JL_{k_J})}{n}}\tau\|\mathbf{m}_{J^c}\|_n\right\} \\ &\quad +2\tau\frac{\|\mathbf{m}_J-\mathbf{X}_J\|_n}{\|\mathbf{X}_{J^c}\|_n^2}\left\{\sqrt{\frac{k_J}{n}}+\sqrt{\frac{2(\xi+k_JL_{k_J})}{n}}\right\}, \end{aligned}$$

we show that for all $\xi > 0$

$$\mathrm{pr}\left\{\exists J/D_1(J)\geq d_1(J)\right\}\leq(1+2\Sigma)\exp(-\xi). \tag{26}$$

Let us study the processes $Z_{1,J}(\sigma^2)$, $Z_{2,J}(\sigma^2)$ and $Z_{3,J}(g,\sigma^2)$. Using equation (28) in Lemma 7.1, we have that for all $\xi > 0$

$$Z_{1,J}(\|\mathbf{X}_{J^c}\|_n^2)\leq\tau^2\sqrt{\frac{2}{n}}\left|\frac{1}{\tau^2+\|\mathbf{m}_{J^c}\|_n^2}-\frac{1}{\|\mathbf{X}_{J^c}\|_n^2}\right|\left(\sqrt{2\xi}+\sqrt{\frac{2}{n}}\xi\right),$$

with probability greater than $1-\exp(-\xi)$. Using equation (25), we have that

$$Z_{2,J}(\|\mathbf{X}_{J^c}\|_n^2)\leq 2\frac{\tau\|\mathbf{m}_{J^c}\|_n}{\sqrt{n}}\left|\frac{1}{\tau^2+\|\mathbf{m}_{J^c}\|_n^2}-\frac{1}{\|\mathbf{X}_{J^c}\|_n^2}\right|\sqrt{2(\xi+k_JL_{k_J})}$$

with probability greater than $1-\exp(-\xi-k_JL_{k_J})$. Let us define

$$Z(g)=\frac{\sqrt{n}\langle\varepsilon,\mathbf{m}_J-g_J\rangle_n}{\tau\|\mathbf{m}_J-g_J\|_n}.$$

Let us remark that if $J=\emptyset$, $Z(g)=0$. Using the same proof as Birgé and Massart [8], we use a classical inequality due to Cirel'son, Ibragimov and Sudakov [10] and we get that

$$\sup_{g\in\mathbb{R}^n,g=g_J}Z(g)\leq\sqrt{k_J}+\sqrt{2(\xi+k_JL_{k_J})},$$

with probability greater than $1-\exp(-\xi-k_JL_{k_J})$. Therefore, we get

$$Z_{3,J}(\mathbf{X}_J,\|\mathbf{X}_{J^c}\|_n^2)\leq 2\frac{\tau\|\mathbf{m}_J-\mathbf{X}_J\|_n}{\sqrt{n}\|\mathbf{X}_{J^c}\|_n^2}\left(\sqrt{k_J}+\sqrt{2(\xi+k_JL_{k_J})}\right).$$

Second step.

We calculate a lower bound for
$\mathcal{K}_{(\mathbf{m}_J,\tau^2+\|\mathbf{m}_{J^c}\|_n)}(\mathbf{X}_J,\|\mathbf{X}_{J^c}\|_n^2)$.

$$\frac{2}{n}K_{(\mathbf{m}_J,\tau^2+\|\mathbf{m}_{J^c}\|_n)}\left(\mathbf{X}_J,\|\mathbf{X}_{J^c}\|_n^2\right)=-\log\left(\frac{\tau^2+\|\mathbf{m}_{J^c}\|_n^2}{\|\mathbf{X}_{J^c}\|_n^2}\right)-1+\frac{\tau^2+\|\mathbf{m}_{J^c}\|_n^2}{\|\mathbf{X}_{J^c}\|_n^2}+\frac{\|\mathbf{m}_J-\mathbf{X}_J\|_n^2}{\|\mathbf{X}_{J^c}\|_n^2}.$$

Applying Lemma 7.4 with

$$v=\frac{\|\mathbf{X}_{J^c}\|_n^2}{\tau^2+\|\mathbf{m}_{J^c}\|_n^2}-1,$$

we get

$$\frac{2}{n}K_{(\mathbf{m}_J,\tau^2+\|\mathbf{m}_{J^c}\|_n)}\left(\mathbf{X}_J,\|\mathbf{X}_{J^c}\|_n^2\right) \geq$$

$$\frac{1}{2}\left(\frac{1}{\tau^2+\|\mathbf{m}_{J^c}\|_n^2}-\frac{1}{\|\mathbf{X}_{J^c}\|_n^2}\right)^2\left(\tau^2+\|\mathbf{m}_{J^c}\|_n^2\right)\min\left\{\|\mathbf{X}_{J^c}\|_n^2,\tau^2+\|\mathbf{m}_{J^c}\|_n^2\right\}+\frac{\|\mathbf{m}_J-\mathbf{X}_J\|_n^2}{\|\mathbf{X}_{J^c}\|_n^2}.$$

Thanks to Assumption (8) on the ζ_k's, it is easy to verify that for each $J \in \mathcal{J}$,

$$\tau^2+\|\mathbf{m}_{J^c}\|_n^2 \geq \frac{\tau^2}{n}\mathcal{X}_{n-k_J}^{-1}\{\exp(-\zeta_{k_J})\}\cdot$$

Moreover using Lemma 7.2,

$$\mathrm{pr}\left[\frac{n}{\tau^2}\|\mathbf{X}_{J^c}\|_n^2 \leq \mathcal{X}_{n-k_J}^{-1}\{\exp(-\zeta_{k_J})\}\right] \leq \exp(-\zeta_{k_J}).$$

Therefore we get that on the set Ω, for all $J \in \mathcal{J}$,

$$\begin{aligned}\min\left\{\|\mathbf{X}_{J^c}\|_n^2,\tau^2+\|\mathbf{m}_{J^c}\|_n^2\right\} &\geq \tau^2\frac{\mathcal{X}_{n-k_J}^{-1}\{\exp(-\zeta_{k_J})\}}{n}\\ \frac{\mathcal{X}_{n-k_J}^{-1}\{\exp(-\zeta_{k_J})\}}{n}\frac{\tau^2}{\|\mathbf{X}_{J^c}\|_n^2} &\leq 1\end{aligned}$$

and finally

$$\begin{aligned}\frac{2}{n}K_{(\mathbf{m}_J,\tau^2+\|\mathbf{m}_{J^c}\|_n^2)}\left(\mathbf{X}_J,\|\mathbf{X}_{J^c}\|_n^2\right) \geq\ & \frac{1}{2}\left(\frac{1}{\tau^2+\|\mathbf{m}_{J^c}\|_n^2}-\frac{1}{\|\mathbf{X}_{J^c}\|_n^2}\right)^2\left(\tau^4+\tau^2\|\mathbf{m}_{J^c}\|_n^2\right)\frac{\mathcal{X}_{n-k_J}^{-1}\{\exp(-\zeta_{k_J})\}}{n}\\ &+\frac{\tau^2\|\mathbf{m}_J-\mathbf{X}_J\|_n^2}{\|\mathbf{X}_{J^c}\|_n^4}\frac{\mathcal{X}_{n-k_J}^{-1}\{\exp(-\zeta_{k_J})\}}{n}\cdot\end{aligned}$$

Third step.

We make the link between the Kullback distance $\mathcal{K}_{(\mathbf{m}_J,\tau^2+\|\mathbf{m}_{J^c}\|_n)}(\mathbf{X}_J,\|\mathbf{X}_{J^c}\|_n^2)$ and $d_1(J)$. Let J be a non empty set in $\mathcal{J}$ and

$$\Pi(J)=\frac{n}{\mathcal{X}_{n-k_J}^{-1}(\exp(-\zeta_{k_J}))}\left\{\left(\sqrt{\frac{\xi}{n}}+\frac{\xi}{n}\right)^2+\frac{2(\xi+k_JL_{k_J})}{n}+\left(\sqrt{\frac{k_J}{n}}+\sqrt{\frac{2(\xi+k_JL_{k_J})}{n}}\right)^2\right\}.$$

Using the inequality $2ab \leq a^2\theta+b^2/\theta$ for all $\theta>0$, we get that on the set Ω, for all $J \in \mathcal{J}$,

$$d_1(J) \leq \theta\frac{2}{n}\mathcal{K}_{(\mathbf{m}_J,\tau^2+\|\mathbf{m}_{J^c}\|_n)}\left(\mathbf{X}_J,\|\mathbf{X}_{J^c}\|_n^2\right)+\frac{1}{\theta}\Pi(J). \tag{27}$$

Using successively that $\sqrt{a+b} \leq \sqrt{a}+\sqrt{b}$ and $(a+b)^2 \leq (1+\eta)a^2+(1+1/\eta)b^2$ for all $a,b,\eta>0$, we get that

$$\Pi(J) \leq \frac{n}{\mathcal{X}_{n-k_J}^{-1}\{\exp(-\zeta_{k_J})\}}\left\{\frac{\xi}{n}\left(5+\frac{\xi}{n}+2\sqrt{\frac{\xi}{n}}\right)(1+1/\eta)+\frac{k_J}{n}\left(1+4L_{k_J}+2\sqrt{2L_{k_J}}\right)(1+\eta)\right\}$$

and putting together the inequalities (26) and (27), we have the desired result.

If $J = \emptyset$, the computation are simplified. The term $Z_{3,J}$ disappears and the term $\|\mathbf{m}_J - \mathbf{X}_J\|_n^2/\|\mathbf{X}_{J^c}\|_n^2$ disappears in the expression of $\mathcal{K}_{(\mathbf{m}_J,\tau^2+\|\mathbf{m}_{J^c}\|_n)}(\mathbf{X}_J, \|\mathbf{X}_{J^c}\|_n^2)$. In place of equation (27) we get

$$d_1(J) \leq \theta \frac{2}{n} \mathcal{K}_{(\mathbf{0},\tau^2+\|\mathbf{m}\|_n)}\left(z, \|X\|_n^2\right) + \frac{1}{\theta} \frac{n}{\mathcal{X}_n^{-1}\{\exp(-\zeta_0)\}} \frac{\xi}{n} \left(3 + \frac{\xi}{n} + 2\sqrt{\frac{\xi}{n}}\right),$$

where $\mathbf{0}$ is the vector null.

5. Proof of Corollary 2.2

From the proof of Proposition 4.1 we get that

$$\begin{aligned} \mathcal{K}_{(\mathbf{m},\tau^2)}(\mathbf{X}_{\widehat{J}_{\text{pen}}}, \|\mathbf{X}_{\widehat{J^c}_{\text{pen}}}\|_n^2)\mathbb{1}_\Omega \quad &\leq \quad \frac{1}{1-\theta}\left\{\mathcal{K}_{(\mathbf{m},\tau^2)}\left(\mathbf{m}_J, \tau^2 + \|\mathbf{m}_{J^c}\|_n^2\right) + \text{pen}(k_J)\right. \\ &\qquad \left. + \frac{1}{\theta} C(k_n, \eta) c(\xi) - \frac{n}{2} D_2(J) \mathbb{1}_\Omega \right\}, \end{aligned}$$

is true except on a set with probability smaller than $(2 + 3\Sigma)\exp(-\xi)$. Consequently, there exists a positive variate, U, such that

$$\text{pr}\left\{U > c(\xi)\right\} \leq (2 + 3\Sigma)\exp(-\xi),$$

and

$$\begin{aligned} \mathcal{K}_{(\mathbf{m},\tau^2)}\left(\mathbf{X}_{\widehat{J}_{\text{pen}}}, \|\mathbf{X}_{\widehat{J^c}_{\text{pen}}}\|_n^2\right)\mathbb{1}_\Omega \quad &\leq \quad \frac{1}{1-\theta}\left\{\mathcal{K}_{(\mathbf{m},\tau^2)}\left(\mathbf{m}_J, \tau^2 + \|\mathbf{m}_{J^c}\|_n^2\right) + \text{pen}(k_J)\right. \\ &\qquad \left. + \frac{1}{\theta} C(k_n, \eta) U - \frac{n}{2} D_2(J) \mathbb{1}_\Omega \right\}. \end{aligned}$$

It remains to calculate $\text{E}(U)$ and $\text{E}|D_2(J)\mathbb{1}_\Omega|$.

$$\begin{aligned} \text{E}(U) \quad &= \quad \int_0^\infty \text{pr}(U > u)\text{d}u = \int_0^\infty \text{pr}(U > c(\xi))c'(\xi)\text{d}\xi \\ &\leq \quad \kappa_1(1 + \Sigma), \end{aligned}$$

where κ_1 is a positive constant. Noting that $D_2(J)$ has expectation 0 (see Eq. (23)), then

$$\text{E}\left\{D_2(J)\mathbb{1}_\Omega\right\} = -\text{E}\left\{D_2(J)\mathbb{1}_{\Omega^c}\right\},$$

and

$$\begin{aligned} \left\{\text{E}|D_2(J)\mathbb{1}_{\Omega^c}|\right\}^2 \quad &\leq \quad \text{E}\{D_2(J)\}^2 \text{pr}(\Omega^c) \\ &\leq \quad \frac{2\epsilon}{n} \frac{\|\mathbf{m}_{J^c}\|_n^2 \left(2\tau^2 + \|\mathbf{m}_{J^c}\|_n^2\right)}{\left(\tau^2 + \|\mathbf{m}_{J^c}\|_n^2\right)^2}. \end{aligned}$$

It follows that for all $J \in \mathcal{J}$,

$$\begin{aligned} \text{E}\left\{\mathcal{K}_{(\mathbf{m},\tau^2)}(\mathbf{X}_{\widehat{J}_{\text{pen}}}, \|\mathbf{X}_{\widehat{J^c}_{\text{pen}}}\|_n^2)\mathbb{1}_\Omega\right\} \quad &\leq \quad \frac{1}{1-\theta}\left\{\mathcal{K}_{(\mathbf{m},\tau^2)}\left(\mathbf{m}_J, \tau^2 + \|\mathbf{m}_{J^c}\|_n^2\right) + \text{pen}(k_J)\right\} \\ &\qquad + \frac{\kappa_1}{\theta(1-\theta)} C(k_n, \eta)(1 + \Sigma) + \frac{1}{1-\theta}\sqrt{\frac{n\epsilon}{2}}. \end{aligned}$$

Finally, the corollary is proved by taking $\theta = (1/C + 1)/2$, and $\eta = \theta C - 1$, for some $C > 1$.

6. Proof of Theorem 2.3

The proof is divided into two steps. The first step is an application of the theorem proved by Birgé and Massart in [8]. More precisely, under the assumptions of Theorem 2.3,

$$\mathrm{E}\left[\left\|\mathbf{m}-\mathbf{X}_{\widehat{J}(\text{pen})}\right\|^2 \mathbb{1}_\Omega\right] \leq \frac{4C(C+1)^2}{(C-1)^3}\left(\inf_{J\in\mathcal{J}}\left\{\|\mathbf{m}_{J^c}\|^2+2\mathrm{E}(\underline{\text{pen}}(k_J))\right\}+(C+1)\tau^2\Sigma\right),$$

where Ω is defined at equation (12). This result is shown by verifying that on the set Ω

$$\underline{\text{pen}}(k) > \frac{C}{2}\tau^2 k\left(1+\sqrt{2L_k}\right)^2,$$

and by following the lines of the proof of Theorem 1 in [8].

The second step consists in showing that the expectation of the quadratic risk on the set Ω^c is controlled. From the following equality

$$\left\|\mathbf{m}-\mathbf{X}_{\widehat{J}}\right\|^2 = \left\|\mathbf{m}-\mathbf{m}_{\widehat{J}}\right\|^2 + \left\|\mathbf{m}_{\widehat{J}}-\mathbf{X}_{\widehat{J}}\right\|^2$$

we deduce that

$$\mathrm{E}\left(\left\|\mathbf{m}-\mathbf{X}_{\widehat{J}}\right\|^2 \mathbb{1}_{\Omega^c}\right) \leq \|\mathbf{m}\|^2\,\mathrm{pr}\left(\Omega^c\right)+\tau^2\mathrm{E}\left(\left\|\varepsilon_{\widehat{J}}\right\|^2 \mathbb{1}_{\Omega^c}\right)\cdot$$

Calculation of $\mathrm{E}\left(\left\|\varepsilon_{\widehat{J}}\right\|^2 \mathbb{1}_{\Omega^c}\right)$. We first write that

$$\mathrm{E}\left(\left\|\varepsilon_{\widehat{J}}\right\|^2 \mathbb{1}_{\Omega^c}\right) \leq k_n\sqrt{\mathrm{E}\left(\max\{\varepsilon_i^2\}\right)^2}\sqrt{\mathrm{pr}\left(\Omega^c\right)}.$$

Using Formulae (4.2.6) given in [11],

$$\mathrm{E}\left(\left(\max\{\varepsilon_i^2\}\right)^2\right) \leq 3+4\sqrt{6}\frac{n-1}{\sqrt{2n-1}} \leq \alpha^2\sqrt{n}$$

for some α, leading to

$$\mathrm{E}\left(\left\|\varepsilon_{\widehat{J}}\right\|^2 \mathbb{1}_{\Omega^c}\right) \leq \alpha k_n n^{1/4}\sqrt{\mathrm{pr}\left(\Omega^c\right)}.$$

The theorem is shown because $\mathrm{pr}\left(\Omega^c\right) \leq \epsilon$, as it was shown in Section 2.1.

7. Useful lemmas

Lemma 7.1. *Let X be a $\mathcal{X}^2$ variable with D degrees of freedom, then for all $x>0$*

$$\mathrm{P}\left(X \geq D+2\sqrt{Dx}+2x\right) \leq \exp(-x), \tag{28}$$

$$\mathrm{P}\left(X \leq D-2\sqrt{Dx}\right) \leq \exp(-x), \tag{29}$$

and for all $0<\lambda<1$,

$$\mathrm{P}\left(X \leq \lambda D\right) \leq \left\{\lambda\exp(1-\lambda)\right\}^{D/2}. \tag{30}$$

Proof. Inequality (28) is shown by Laurent and Massart (Lem. 1) [17].

For proving Inequality (30), let us start with the Chernoff inequality:

$$\forall x>0, \mu>0,\ \ \mathrm{pr}\left(X \leq x\right) \leq \exp\left[\inf_{\mu>0}\left\{\mu x+\log\left\{\mathrm{E}\left(\mathrm{e}^{-\mu x}\right)\right\}\right\}\right].$$

If $x < D$, the function

$$g(\mu) = \mu x + \log\left\{\mathrm{E}\left(\mathrm{e}^{-\mu x}\right)\right\}$$

is minimum in $\mu = D/2x - 1/2$ and

$$g\left(\frac{D}{2x} - \frac{1}{2}\right) = \exp\left\{\frac{D - x - D\log(D) + D\log(x)}{2}\right\}.$$

□

Lemma 7.2. *Let T be a χ^2 variable with D degrees of freedom and let T' be a non central χ^2 variable with D degrees of freedom and non centrality parameter $a > 0$. Then for all $u > 0$*

$$P(T' \leq u) \leq P(T \leq u).$$

Proof. The proof of Lemma 7.2 can be found in [5], Lemma 1. □

Lemma 7.3. *Let $k_n \leq n/2$ and, for each $k = 0, \dots, k_n$ let*

$$\zeta_k = k\left\{1 + 2\log\left(\frac{n}{k}\right)\right\}.$$

Then there exist some constants, c_1, c_2, such that $c_1(n-k) \leq \mathcal{X}^{-1}_{n-k}\{\exp(-\zeta_k)\} \leq c_2(n-k)$.

Proof. From equation (28) in Lemma 7.1, we get that

$$\mathcal{X}^{-1}_{n-k}\{\exp(-\zeta_k)\} \leq (n-k)\left[1 + 2\sqrt{\frac{-\log(1-\exp(-\zeta_k))}{n-k}} - 2\frac{\log\{1-\exp(-\zeta_k)\}}{n-k}\right].$$

Some simple calculation show that $\zeta_k \geq 2\log(n)$ leading for $n \geq 2$ to

$$-\frac{\log\{1-\exp(-\zeta_k)\}}{n-k} \leq -\frac{2}{n}\log\left(1 - \frac{1}{n^2}\right) \leq 0.3.$$

From equation (30), we get that $\mathcal{X}^{-1}\{\exp(-\zeta_k)\} \geq \lambda(n-k)$ if

$$\{\lambda\exp(1-\lambda)\}^{(n-k)/2} \leq \exp(-\zeta_k). \qquad (31)$$

Let $u \in [0, 1/2]$ and $h(u)$ be defined by

$$h(u) = -2\frac{u}{1-u}\{1 - 2\log(u)\}.$$

Solving Inequality (31) is equivalent to find $0 < \lambda < 1$ such that

$$\log(\lambda) + 1 - \lambda \leq h(k/n).$$

It is easy to prove that such a λ exists: the function h decreases from 0 to $h(1/2)$ and the function $\lambda \mapsto \log(\lambda) + 1 - \lambda$ increases from minus infinity to 0. □

Lemma 7.4.

$$\log(1+v) - 1 + \frac{1}{1+v} \geq \frac{1}{2}\frac{v^2}{1+v} \text{ } \textit{if } -1 < v \leq 0 \qquad (32)$$

$$\geq \frac{1}{2}\frac{v^2}{(1+v)^2} \text{ } \textit{if } v \geq 0. \qquad (33)$$

Proof. Let $-1 < v \leq 0$ and

$$f_1(v) = \log(1+v) - 1 + \frac{1}{1+v}\left(1 - \frac{1}{2}v^2\right).$$

$f_1(0) = 0$ and $f_1(-1) = +\infty$.

$$f_1'(v) = -\frac{1}{2}\frac{v^2}{(1+v)^2}$$

is negative, showing thus equation (32).

Let $v \geq 0$ and

$$f_2(v) = \log(1+v) - 1 + \frac{1}{1+v} - \frac{1}{2}\frac{v^2}{(1+v)^2}.$$

$f_2(0) = 0$ et $f_2(+\infty) = +\infty$.

$$f_2'(v) = \frac{v^2}{(1+v)^3}$$

is positive, showing thus equation (33). □

Lemma 7.5. *For $n \geq 2$ we have the following property:*

$$\sum_{k=1}^{n}\left(\frac{k}{2n}\right)^k \leq \frac{\kappa}{n}$$

where κ is some constant.

Proof. Let g be defined as $g(x) = \exp(nx\log(x/2))$ for $1/n \leq x \leq 1$. Its derivative $g'(x) = n\log(ex/2)g(x)$ is negative for $x \leq 2/e$ and positive for $x \geq 2/e$. Let the integer k_0 be such that $k_0 \leq ne^{-1} \leq k_0 + 1$. We get

$$\begin{aligned}
\sum_{k=1}^{n}\left(\frac{k}{2n}\right)^k &\leq \sum_{k=1}^{k_0}\left(\frac{k}{2n}\right)^k + n\sup_{x\in[e^{-1},1]}\{g(x)\} \\
&\leq g\left(\frac{1}{n}\right) + n\sum_{k=2}^{k_0}\frac{1}{n}g\left(\frac{k}{n}\right) + n\max\left\{g\left(\frac{1}{e}\right), g(1)\right\} \\
&\leq g\left(\frac{1}{n}\right) + n\int_{1/n}^{1/e} g(x)\mathrm{d}x + ng\left(\frac{1}{e}\right).
\end{aligned}$$

Noting that

$$\int_{1/n}^{1/e} g(x)\mathrm{d}x = \int_{1/n}^{1/e}\frac{-g'(x)}{n\log(2/ex)}\mathrm{d}x \leq \frac{1}{n\log(2)}\int_{1/n}^{1/e} -g'(x)\mathrm{d}x \leq \frac{1}{n\log(2)}g\left(\frac{1}{n}\right),$$

and that

$$g\left(\frac{1}{n}\right) = \frac{1}{2n},\ \ g\left(\frac{1}{e}\right) = \exp\left\{-n\frac{\log(2e)}{e}\right\},$$

we get the desired result. □

References

[1] F. Abramovich, Y. Benjamini, D. Donoho and I. Johnston, *Adapting to unknown sparsity by controlloing the false discovery rate.* Technical Report 2000-19, Department of Statistics, Stanford University (2000).

[2] H. Akaike, Information theory and an extension of the maximum likelihood principle, in *2nd International Symposium on Information Theory*, B.N. Petrov and F. Csaki Eds., Budapest Akademia Kiado (1973) 267–281.

[3] H. Akaike, A bayesian analysis of the minimum aic procedure. *Ann. Inst. Statist. Math.* **30** (1978) 9–14.

[4] A. Antoniadis, I. Gijbels and G. Grégoire, Model selection using wavelet decomposition and applications. *Biometrika* **84** (1997) 751–763.

[5] Y. Baraud, S. Huet and B. Laurent, Adaptive tests of qualitative hypotheses. *ESAIM: PS* **7** (2003) 147–159.

[6] A. Barron, L. Birgé and P. Massart, Risk bounds for model selection via penalization. *Probab. Theory Rel. Fields* **113** (1999) 301–413.

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

[8] L. Birgé and P. Massart, Gaussian model selection. *J. Eur. Math. Soc. (JEMS)* **3** (2001) 203–268.

[9] L. Birgé and P. Massart, *A generalized cp criterion for gaussian model selection.* Technical report, Univ. Paris 6, Paris 7, Paris (2001).

[10] B.S. Cirel'son, I.A. Ibragimov and V.N. Sudakov, Norm of gaussian sample function, in *Proceedings of the 3rd Japan-URSS. Symposium on Probability Theory*, Berlin, Springer-Verlag. *Springer Lect. Notes Math.* **550** (1976) 20–41.

[11] H.A. David, *Order Statistics. Wiley series in Probability and mathematical Statistics.* John Wiley and Sons, NY (1981).

[12] E.P. Box and R.D. Meyer, An analysis for unreplicated fractional factorials. *Technometrics* **28** (1986) 11–18.

[13] D.P. Foster and R.A. Stine, Adaptive variable selection competes with bayes expert. Technical report, The Wharton School of the University of Pennsylvania, Philadelphia (2002).

[14] S. Huet, *Comparison of methods for estimating the non zero components of a gaussian vector.* Technical report, INRA, MIA-Jouy, `www.inra.fr/miaj/apps/cgi-bin/raptech.cgi` (2005).

[15] M.C. Hurvich and C.L. Tsai, Regression and time series model selection in small samples. *Biometrika* **76** (1989) 297–307.

[16] I. Johnston and B. Silverman, Empirical bayes selection of wavelet thresholds. Available from `www.stats.ox.ac.uk/ silverma/papers.html` (2003).

[17] B. Laurent and P. Massart, Adaptive estimation of a quadratic functional by model selection. *Ann. Statist.* **28** (2000) 1302–1338.

[18] R. Nishii, Maximum likelihood principle and model selection when the true model is unspecified. *J. Multivariate Anal.* **27** (1988) 392–403.

[19] P.D. Haaland and M.A. O'Connell, Inference for effect-saturated fractional factorials. *Technometrics* **37** (1995) 82–93.

[20] J. Rissanen, Universal coding, information, prediction and estimation. *IEEE Trans. Infor. Theory* **30** (1984) 629–636.

[21] R.V. Lenth, Quick and easy analysis of unreplicated factorials. *Technometrics* **31(4)** (1989) 469–473.

[22] G. Schwarz, Estimating the dimension of a model. *Ann. Statist.* **6** (1978) 461–464.