High-dimensional Gaussian model selection on a Gaussian design

Nicolas Verzelen

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

High-dimensional Gaussian model selection on a Gaussian design

Nicolas Verzelen

N° 6616 — version 2

initial version Aout 2008 — revised version Avril 2009

Thème COG

ISSN 0249-6399 ISRN INRIA/RR--6616--FR+ENG

INSTITUT NATIONAL
DE RECHERCHE
EN INFORMATIQUE
ET EN AUTOMATIQUE

INRIA

centre de recherche
SACLAY - ÎLE-DE-FRANCE

High-dimensional Gaussian model selection on a Gaussian design

Nicolas Verzelen * †

Thème COG — Systèmes cognitifs
Équipes-Projets Select

Rapport de recherche n° 6616 — version 2 — initial version Aout 2008 — revised version Avril 2009 — 54 pages

Abstract: We consider the problem of estimating the conditional mean of a real Gaussian variable $Y = \sum_{i=1}^{p} \theta_i X_i + \epsilon$ where the vector of the covariates $(X_i)_{1 \leq i \leq p}$ follows a joint Gaussian distribution. This issue often occurs when one aims at estimating the graph or the distribution of a Gaussian graphical model. We introduce a general model selection procedure which is based on the minimization of a penalized least-squares type criterion. It handles a variety of problems such as ordered and complete variable selection, allows to incorporate some prior knowledge on the model and applies when the number of covariates p is larger than the number of observations n. Moreover, it is shown to achieve a non-asymptotic oracle inequality independently of the correlation structure of the covariates. We also exhibit various minimax rates of estimation in the considered framework and hence derive adaptiveness properties of our procedure.

Key-words: Model selection, Linear regression, oracle inequalities, Gaussian graphical models, minimax rate of estimation

* Laboratoire de Mathématiques UMR 8628, Université Paris-Sud, 91405 Osay
† INRIA Saclay, Projet SELECT, Université Paris-Sud, 91405 Osay

Centre de recherche INRIA Saclay – Île-de-France
Parc Orsay Université
4, rue Jacques Monod, 91893 ORSAY Cedex
Téléphone : +33 1 72 92 59 00

Sélection de modèles en grande dimension pour des design gaussiens

Résumé : We consider the problem of estimating the conditional mean of a real Gaussian variable $Y = \sum_{i=1}^{p} \theta_i X_i + \epsilon$ where the vector of the covariates $(X_i)_{1 \leq i \leq p}$ follows a joint Gaussian distribution. This issue often occurs when one aims at estimating the graph or the distribution of a Gaussian graphical model. We introduce a general model selection procedure which is based on the minimization of a penalized least squares type criterion. It handles a variety of problems such as ordered and complete variable selection, allows to incorporate some prior knowledge on the model and applies when the number of covariates p is larger than the number of observations n. Moreover, it is shown to achieve a non-asymptotic oracle inequality independently of the correlation structure of the covariates. We also exhibit various minimax rates of estimation in the considered framework and hence derive adaptivity properties of our procedure.

Mots-clés : Sélection de modèles, régression linéaire, inégalités oracles, modèles graphiques gaussiens, vitesse minimax d'estimation

1 Introduction

1.1 Regression model

We consider the following regression model

$$Y \quad = \quad X\theta + \epsilon \ , \tag{1}$$

where θ is an unknown vector of $\mathbb{R}^p$. The row vector $X := (X_i)_{1\leq i\leq p}$ follows a real zero mean Gaussian distribution with non singular covariance matrix Σ and ϵ is a real zero mean Gaussian random variable independent of X with variance σ^2. The variance of ϵ corresponds to the conditional variance of Y given X, $\mathrm{Var}(Y|X)$. In the sequel, the parameters θ, Σ, and σ^2 are considered as unknown.

Suppose we are given n i.i.d. replications of the vector (Y, X). We respectively write $\mathbf{Y}$ and $\mathbf{X}$ for the vector of n observations of Y and the $n \times p$ matrix of observations of X. In the present work, we propose a new procedure to estimate the vector θ, when the matrix Σ and the variance σ^2 are both unknown. This corresponds to estimating the conditional expectation of the variable Y given the random vector X. Besides, we want to handle the difficult case of high-dimensional data, i.e. the number of covariates p is possibly much larger than n. This estimation problem is equivalent to building a suitable predictor of Y given the covariates $(X_i)_{1\leq i\leq p}$. Classically, we shall use the mean-squared prediction error to assess the quality of our estimation. For any $(\theta_1, \theta_2) \in \mathbb{R}^p$, it is defined by

$$l(\theta_1, \theta_2) := \mathbb{E}\left[(X\theta_1 - X\theta_2)^2\right] \ . \tag{2}$$

1.2 Applications to Gaussian graphical models (GGM)

Estimation in the regression model (1) is mainly motivated by the study of Gaussian graphical models (GGM). Let Z be a Gaussian random vector indexed by the elements of a finite set Γ. The vector Z is a GGM with respect to an undirected graph $\mathcal{G} = (\Gamma, E)$ if for any couple (i, j) which is not contained in the edge set E, Z_i and Z_j are independent, given the remaining variables. See Lauritzen [23] for definitions and main properties of GGM. Estimating the neighborhood of a given point $i \in \Gamma$ is equivalent to estimating the support of the regression of Z_i with respect to the covariates $(Z_j)_{j\in\Gamma\setminus\{i\}}$. Meinshausen and Bühlmann [26] have taken this point of view in order to estimate the graph of a GGM. Similarly, we can apply the model selection procedure we shall introduce in this paper to estimate the support of the regression and therefore the graph $\mathcal{G}$ of a GGM.

Interest in these models has grown since they allow the description of dependence structure of high-dimensional data. As such, they are widely used in spatial statistics [16, 27] or probabilistic expert systems [15]. More recently, they have been applied to the analysis of microarray data. The challenge is to infer the network regulating the expression of the genes using only a small sample of data, see for instance Schäfer and Strimmer [29], or Wille *et al.* [39].

This has motivated the search for new estimation procedures to handle the linear regression model (1) with Gaussian random design. Finally, let us mention that the model (1) is also of interest when estimating the distribution of directed graphical models or more generally the joint distribution of a large Gaussian random vector. Estimating the joint distribution of a Gaussian vector $(Z_i)_{1\leq i\leq p}$ indeed amounts to estimating the conditional expectations and variance of Z_i given $(Z_j)_{1\leq j\leq i-1}$ for any $1 \leq i \leq p$.

1.3 General oracle inequalities

Estimation of high-dimensional Gaussian linear models has now attracted a lot of attention. Various procedures have been proposed to perform the estimation of θ when $p > n$. The challenge at hand it to design estimators that are both computationally feasible and are proved to be efficient. The Lasso estimator has been introduced by Tibshirani [33]. Meinshausen and Bühlmann [26] have shown that this estimator is consistent under a neighborhood stability condition. These convergence results were refined in the works of Zhao and Yu [40], Bunea *et al.* [11], Bickel *et al.* [5], or Candès and Plan [12] in a slightly different framework. Candès and Tao [13] have also introduced the Dantzig-selector procedure which performs similarly as l_1 penalization methods. In the more specific context of GGM, Bühlmann and Kalisch [21] have analyzed the PC algorithm and have proven its consistency when the GGM follows a faithfulness assumption. All these methods share an attractive computational efficiency and most of them are proven to converge at the optimal rate when the covariates are nearly independent. However, they also share two main drawbacks. First, the l_1 estimators are known to behave poorly when the covariates are highly correlated and even for some covariance structures with small correlation (see e.g. [12]). Similarly, the PC algorithm is not consistent if the faithfulness assumption is not fulfilled. Second, these procedures do not allow to integrate some biological or physical prior knowledge. Let us provide two examples. Biologists sometimes have a strong preconception of the underlying biological network thanks to previous experimentations. For instance, Sachs *et al.* [28]) have produced multivariate flow cytometry data in order to study a human T cell signaling pathway. Since this pathway has important medical implications, it was already extensively studied and a network is conventionally accepted (see [28]). For this particular example, it could be more interesting to check whether some interactions were forgotten or some unnecessary interactions were added in the model than performing a complete graph estimation. Moreover, the covariates have in some situations a temporal or spatial interpretation. In such a case, it is natural to introduce an *order* between the covariates, by assuming that a covariate which is *close* (in space or time) to the response Y is more likely to be significant. Hence, an ordered variable selection method is here possibly more relevant than the complete variable selection methods previously mentioned.

Let us emphasize the main differences of our estimation setting with related studies in the literature. Birgé and Massart [8] consider model selection in a fixed design setting with known variance. Bunea *et al.* [10] also suppose that the variance is known. Yet, they consider a random design setting, but they assume that the regression functions are bounded (Assumption A.2 in their paper) which is not the case here. Moreover, they obtain risk bounds with respect to the empirical norm $\|\mathbf{X}(\widehat{\theta} - \theta)\|_n^2$ and not the integrated loss $l(.,.)$. Here, $\|.\|_n$ refers to the canonical norm in $\mathbb{R}^n$ reweighted by $\sqrt{n}$. As mentioned earlier, our objective is to infer the conditional expectation of Y given X. Hence, it is more significant to assess the risk with respect to the loss $l(.,.)$. Baraud *et al.* [4] consider fixed design regression but do not assume that the variance is known.

Our objective is twofold. First, we introduce a general model selection procedure that is very flexible and allows to integrate any prior knowledge on the regression. We prove non-asymptotic oracle inequalities that hold without any assumption on the correlation structure between the covariates. Second, we obtain non-asymptotic rates of estimation for our model (1) that help us to derive adaptive properties for our criterion.

In the sequel, a *model* m stands for a subset of $\{1, \ldots, p\}$. We note d_m the size of m whereas the linear space S_m refers to the set of vectors $\theta \in \mathbb{R}^p$ whose components outside m equal zero. If d_m is smaller than n, then we define $\widehat{\theta}_m$ as the least-square estimator of θ over S_m. In the

sequel, Π_m stands for the projection of $\mathbb{R}^n$ into the space generated by $(\mathbf{X}_i)_{i\in m}$. Hence, we have the relation $\mathbf{X}\widehat{\theta}_m = \Pi_m \mathbf{Y}$. Since the covariance matrix Σ is non singular, observe that almost surely the rank of Π_m is d_m. Given a collection $\mathcal{M}$ of models, our purpose is to select a model $\widehat{m} \in \mathcal{M}$ that exhibits a risk as small as possible with respect to the prediction loss function $l(.,.)$ defined in (2). The model m^* that minimizes the risks $\mathbb{E}[l(\widehat{\theta}_m, \theta)]$ over the whole collection $\mathcal{M}$ is called an oracle. Hence, we want to perform as well as the oracle $\widehat{\theta}_{m^*}$. However, we do not have access to m^* as it requires the knowledge of the true vector θ. A classical method to estimate a *good* model $\widehat{m}$ is achieved through *penalization* with respect to the complexity of models. In the sequel, we shall select the model $\widehat{m}$ as

$$\widehat{m} := \arg\min_{m\in\mathcal{M}} \mathrm{Crit}(m) := \arg\min_{m\in\mathcal{M}} \|\mathbf{Y} - \Pi_m \mathbf{Y}\|_n^2 \left[1 + pen(m)\right] \ , \tag{3}$$

where $pen(.)$ is a positive function defined on $\mathcal{M}$. Besides, we recall that $\|.\|_n$ refers to the canonical norm in $\mathbb{R}^n$ reweighted by $\sqrt{n}$. Observe that $Crit(m)$ is the sum of the least-square error $\|\mathbf{Y} - \Pi_m \mathbf{Y}\|_n^2$ and a penalty term $pen(m)$ rescaled by the least-square error in order to come up with the fact that the conditional variance σ^2 is unknown. We precise in Section 2 the heuristics underlying this model selection criterion. Baraud *et al.* [4] have extensively studied this penalization method in the fixed design Gaussian regression framework with unknown variance. In their introduction, they explain how one may retrieve classical criteria like AIC [2], BIC [30], and FPE [1] by choosing a suitable penalty function $pen(.)$.

This model selection procedure is really flexible through the choices of the collection $\mathcal{M}$ and of the penalty function $pen(.)$. Indeed, we may perform complete variable selection by taking the collection of subsets of $\{1,\ldots,p\}$ whose is smaller than some integer d. Otherwise, by taking a nested collection of models, one performs ordered variable selection. We give more details in Sections 2 and 3. If one has some prior idea on the true model m, then one could only consider the collection of models that are close in some sense to m. Moreover, one may also give a Bayesian flavor to the penalty function $pen(.)$ and hence specify some prior knowledge on the model.

First, we state a non-asymptotic oracle inequality when the complexity of the collection $\mathcal{M}$ is small and for penalty functions $pen(m)$ that are larger than $Kd_m/(n-d_m)$ with $K > 1$. Then, we prove that the FPE criterion of Akaike [1] which corresponds to the choice $K = 2$ achieves an asymptotic exact oracle inequality for the special case of ordered variable selection. For the sake of completeness, we prove that choosing K smaller than one yields to terrible performances.

In Section 3.2, we consider general collection of models $\mathcal{M}$. By introducing new penalties that take into account the complexity of $\mathcal{M}$ as in [9], we are able to state a non-asymptotic oracle inequality. In particular, we consider the problem of complete variable selection. In Section 3.4, we define penalties based on a prior distribution on $\mathcal{M}$. We then derive the corresponding risk bounds.

Interestingly, these rates of convergence do not depend on the covariance matrix Σ of the covariates, whereas known results on the Lasso or the Dantzig selector rely on some assumptions on Σ, as discussed in Section 3.2. We illustrate in Section 5 on simulated examples that for some covariance matrices Σ the Lasso performs poorly whereas our methods still behaves well. Besides, our penalization method does not require the knowledge of the conditional variance σ^2. In contrast, the Lasso and the Dantzig selector are constructed for known variance. Since σ^2 is unknown, one either has to estimate it or has to use a cross-validation method in order

to calibrate the penalty. In both cases, there is some room for improvements for the practical calibration of these estimators.

However, our model selection procedure suffers from a computational cost that depends linearly on the size of the collection $\mathcal{M}$. For instance, the complete variable selection problem is NP-hard. This makes it intractable when p becomes too large (i.e. more than 50). In contrast, our criterion applies for arbitrary p when considering ordered variable selection since the size of $\mathcal{M}$ is linear with n. We shall mention in the discussion some possible extensions that we hope can cope with the computational issues.

In a simultaneous and independent work to ours, Giraud [19] applies an analogous procedure to estimate the graph of a GGM. Using slightly different techniques, he obtains non-asymptotic results that are complementary to ours. However, he performs an unnecessary thresholding to derive an upper bound of the risk. Moreover, he does not consider the case of nested collections of models as we do in Section 3.1. Finally, he does not derive minimax rates of estimation.

1.4 Minimax rates of estimation

In order to assess the optimality of our procedure, we investigate in Section 4 the minimax rates of estimation for ordered and complete variable selection. For ordered variable selection, we compute the minimax rate of estimation over ellipsoids which is analogous to the rate obtained in the fixed design framework. We derive that our penalized estimator is adaptive to the collection of ellipsoids independently of the covariance matrix Σ. For complete variable selection, we prove that the minimax rates of estimator of vectors θ with at most k non-zero components is of order $\frac{k \log p}{n}$ when the covariates are independent. This is again coherent with the situation observed in the fixed design setting. Then, the estimator $\widetilde{\theta}$ defined for complete variable selection problem is shown to be adaptive to any sparse vector θ. Moreover, it seems that the minimax rates may become faster when the matrix Σ is far from identity. We investigate this phenomenon in Section 4.2. All these minimax rates of estimation are, to our knowledge, new in the Gaussian random design regression. Tsybakov [35] has derived minimax rates of estimation in a general random design regression setup, but his results do not apply in our setting as explained in Section 4.2.

1.5 Organization of the paper and some notations

In Section 2, we precise our estimation procedure and explain the heuristics underlying the penalization method. The main results are stated in Section 3. In Section 4, we derive the different minimax rates of estimation and assess the adaptivity of the penalized estimator $\widehat{\theta}_{\widehat{m}}$. We perform a simulation study and compare the behaviour of our estimator with Lasso and adaptive Lasso in Section 5. Section 6 contains a final discussion and some extensions, whereas the proofs are postponed to Section 7.

Throughout the paper, $\|.\|_n^2$ stands for the square of the canonical norm in $\mathbb{R}^n$ reweighted by n. For any vector Z of size n, we recall that $\Pi_m Z$ denotes the orthogonal projection of Z onto the space generated by $(\mathbf{X}_i)_{i\in m}$. The notation X_m stands for $(X_i)_{i\in m}$ and $\mathbf{X}_m$ represents the $n \times d_m$ matrix of the n observations of X_m. For the sake of simplicity, we write $\widetilde{\theta}$ for the penalized estimator $\widehat{\theta}_{\widehat{m}}$. For any $x > 0$, $\lfloor x \rfloor$ is the largest integer smaller than x and $\lceil x \rceil$ is the smallest integer larger than x. Finally, L, L_1, L_2,... denote universal constants that may vary from line to line. The notation $L(.)$ specifies the dependency on some quantities.

2 Estimation procedure

Given a collection of models $\mathcal{M}$ and a penalty $pen : \mathcal{M} \to \mathbb{R}^+$, the estimator $\widetilde{\theta}$ is computed as follows:

Model selection procedure

1. Compute $\widehat{\theta}_m = \arg\min_{\theta' \in S_m} \|Y - X\theta'\|_n^2$ for all models $m \in \mathcal{M}$.
2. Compute $\widehat{m} := \arg\min_{m \in \mathcal{M}} \|\mathbf{Y} - \mathbf{X}\widehat{\theta}_m\|_n^2 \, [1 + pen(m)]$.
3. $\widetilde{\theta} := \widehat{\theta}_{\widehat{m}}$.

The choice of the collection $\mathcal{M}$ and the penalty function $pen(.)$ depends on the problem under study. In what follows, we provide some preliminary results for the parametric estimators $\widehat{\theta}_m$ and we give an heuristic explanation for our penalization method.

For any vector θ' in $\mathbb{R}^p$, we define the mean-squared error $\gamma(.)$ and its empirical counterpart $\gamma_n(.)$ as

$$\gamma(\theta') := \mathbb{E}_\theta \left[(Y - X\theta')^2 \right] \quad \text{and} \quad \gamma_n(\theta') := \|\mathbf{Y} - \mathbf{X}\theta'\|_n^2 \ . \tag{4}$$

The function $\gamma(.)$ is closely connected to the loss function $l(.,.)$ through the relation $l(\beta, \theta) = \gamma(\beta) - \gamma(\theta)$.

Given a model m of size strictly smaller than n, we refer to θ_m as the unique minimizer of $\gamma(.)$ over the subset S_m. It then follows that $\mathbb{E}(Y|X_m) = \sum_{i \in m} \theta_i X_i$ and $\gamma(\theta_m)$ is the conditional variance of Y given X_m. As for it, the least squares estimator $\widehat{\theta}_m$ is the minimizer of $\gamma_n(.)$ over the space S_m.

$$\widehat{\theta}_m := \arg \min_{\theta' \in S_m} \gamma_n(\theta') \ \text{ a.s. } .$$

It is almost surely uniquely defined since Σ is assumed to be non-singular and since $d_m < n$. Besides $\gamma_n(\widehat{\theta}_m)$ equals $\|\mathbf{Y} - \Pi_m \mathbf{Y}\|_n^2$. Let us derive two simple properties of $\widehat{\theta}_m$ that will give us some hints to perform model selection.

Lemma 2.1. *For any model m whose dimension is smaller than $n-1$, the expected mean-squared error of $\widehat{\theta}_m$ and the expected least squares of $\widehat{\theta}_m$ respectively equal*

$$\mathbb{E}\left[\gamma(\widehat{\theta}_m)\right] = \left[l(\theta_m, \theta) + \sigma^2\right] \left(1 + \frac{d_m}{n - d_m - 1}\right) \ , \tag{5}$$

$$\mathbb{E}\left[\gamma_n(\widehat{\theta}_m)\right] = \left[l(\theta_m, \theta) + \sigma^2\right] \left(1 - \frac{d_m}{n}\right) \ . \tag{6}$$

The proof is postponed to the Appendix. From Equation (5), we derive a bias variance decomposition of the risk of the estimator $\widehat{\theta}_m$:

$$\mathbb{E}\left[l(\widehat{\theta}_m, \theta)\right] = l(\theta_m, \theta) + \left[\sigma^2 + l(\theta_m, \theta)\right] \frac{d_m}{n - d_m - 1} \ .$$

Hence, $\widehat{\theta}_m$ converges to θ_m in probability when n converges to infinity. Contrary to the fixed design regression framework, the variance term $\left[\sigma^2 + l(\theta_m, \theta)\right] \frac{d_m}{n - d_m - 1}$ depends on the bias term

$l(\theta_m, \theta)$. Besides, this variance term does not necessarily increase when the dimension of the model increases.

Let us now explain the idea underlying our model selection procedure. We aim at choosing a model $\widehat{m}$ that nearly minimizes the mean-squared error $\gamma(\widehat{\theta}_m)$. Since we do not have access to $\gamma(\widehat{\theta}_m)$ nor to the bias $l(\theta_m, \theta)$, we perform an unbiased estimation of the risk as done by Mallows [24] in the fixed design framework.

$$\begin{aligned}
\gamma\left(\widehat{\theta}_m\right) &\approx \gamma_n\left(\widehat{\theta}_m\right) + \mathbb{E}\left[\gamma\left(\widehat{\theta}_m\right) - \gamma_n\left(\widehat{\theta}_m\right)\right] \\
&\approx \gamma_n\left(\widehat{\theta}_m\right) + \mathbb{E}\left[\gamma_n\left(\widehat{\theta}_m\right)\right] \frac{d_m}{n-d_m}\left[2 + \frac{d_m+1}{n-d_m-1}\right] \\
&\approx \gamma_n\left(\widehat{\theta}_m\right)\left[1 + \frac{d_m}{n-d_m}\left(2 + \frac{d_m+1}{n-d_m-1}\right)\right] . \qquad (7)
\end{aligned}$$

By Lemma 2.1, these approximations are in fact equalities in expectation. Since the last expression only depends on the data, we may compute its minimizer over the collection $\mathcal{M}$. This approximation is effective and minimizing (7) provides a good estimator $\widetilde{\theta}$ when the size of the collection $\mathcal{M}$ is moderate as stated in Theorem 3.1. We recall that $\|\mathbf{Y} - \Pi_m \mathbf{Y}\|_n^2$ equals $\gamma_n(\widehat{\theta}_m)$. Hence, our previous heuristics would lead to a choice of penalty $pen(m) = \frac{d_m}{n-d_m}\left(2 + \frac{d_m+1}{n-d_m-1}\right)$ in our criterion (3), whereas FPE criterion corresponds to $pen(m) = \frac{2d_m}{n-d_m}$. These two penalties are equivalent when the dimension d_m is small in front of n. In Theorem 3.1, we explain why these criteria allow to derive approximate oracle inequalities when there is a small number of models. However, when the size of the collections $\mathcal{M}$ increases, we need to design other penalties that take into account the complexity of the collection $\mathcal{M}$ (see Section 3.2).

3 Oracle inequalities

3.1 A small number of models

In this section, we restrict ourselves to the situation where the collection of models $\mathcal{M}$ only contains a small number of models as defined in [9] Sect 3.1.2.

($\mathbb{H}_{Pol}$): for each $d \geq 1$ the number of models $m \in \mathcal{M}$ such that $d_m = d$ grows at most polynomially with respect to d. In other words, there exists α and β such that for any $d \geq 1$, $\mathrm{Card}\left(\{m \in \mathcal{M},\ d_m = d\}\right) \leq \alpha d^\beta$.

($\mathbb{H}_\eta$): The dimension d_m of every model m in $\mathcal{M}$ is smaller than ηn. Moreover, the number of observations n is larger than $6/(1-\eta)$.

Assumption ($\mathbb{H}_{Pol}$) states that there is at most a polynomial number of models with a given dimension. It includes in particular the problem of ordered variable selection, on which we will focus in this section. Let us introduce the collection of models relevant for this issue. For any positive number i smaller or equal to p, we define the model $m_i := \{1, \ldots, i\}$ and the nested collection $\mathcal{M}_i := \{m_0, m_1, \ldots m_i\}$. Here, m_0 refers to the empty model. Any collection $\mathcal{M}_i$ satisfies ($\mathbb{H}_{Pol}$) with $\beta = 0$ and $\alpha = 1$.

Theorem 3.1. *Let η be any positive number smaller than one. Assume that the collection $\mathcal{M}$ satisfies ($\mathbb{H}_{Pol}$) and ($\mathbb{H}_\eta$). If the penalty $pen(.)$ is lower bounded as follows*

$$pen(m) \geq K\frac{d_m}{n-d_m} \text{ for all } m \in \mathcal{M} \text{ and some } K > 1 \ , \tag{8}$$

then

$$\mathbb{E}\left[l(\widetilde{\theta},\theta)\right] \leq L(K,\eta) \inf_{m\in\mathcal{M}} \left[l(\theta_m,\theta) + \frac{n-d_m}{n} pen(m)\left[\sigma^2 + l(\theta_m,\theta)\right]\right] + \tau_n \ , \tag{9}$$

where the error term τ_n is defined as

$$\tau_n = \tau_n\left[Var(Y), K, \eta, \alpha, \beta\right] := L_1(K,\eta,\alpha,\beta)\left[\frac{\sigma^2}{n} + n^{3+\beta}\, Var(Y)\exp\left[-nL_2(K,\eta)\right]\right] \ ,$$

and $L_2(K,\eta)$ is positive.

The theorem applies for any n, any p and there is no hidden dependency on n or p in the constants. Besides, observe that the theorem does not depend at all on the covariance matrix Σ between the covariates. If we choose the penalty $pen(m) = K\frac{d_m}{n-d_m}$, we obtain an approximate oracle inequality.

$$\mathbb{E}\left[l(\widetilde{\theta},\theta)\right] \leq L(K,\eta) \inf_{m\in\mathcal{M}} \mathbb{E}\left[l(\widehat{\theta}_m,\theta)\right] + \tau_n\left[\mathrm{Var}(Y), K, \eta, \alpha, \beta\right] \ ,$$

thanks to Lemma 2.1. The term in $n^{3+\beta}\mathrm{Var}(Y)\exp[-nL_2(K,\eta)]$ converges exponentially fast to 0 when n goes to infinity and is therefore considered as negligible. One interesting feature of this oracle inequality is that it allows to consider models of dimensions as close to n as we want providing that n is large enough. This will not be possible in the next section when handling more complex collections of models.

If we have stated that $\widetilde{\theta}$ performs almost as well as the oracle model, one may wonder whether it is possible to perform exactly as well as the oracle. In the next proposition, we shall prove that under additional assumption the estimator $\widetilde{\theta}$ with $K = 2$ follows an asymptotic exact oracle inequality. We state the result for the problem of ordered variable selection. Let us assume for a moment that the set of covariates is infinite, i.e. $p = +\infty$. In this setting, we define the subset Θ of sequences $\theta = (\theta_i)_{i\geq 1}$ such that $< X, \theta >$ converges in L^2. In the following proposition, we assume that $\theta \in \Theta$.

Definition 3.1. *Let s and R be two positive numbers. We define the so-called ellipsoid $\mathcal{E}'_s(R)$ as*

$$\mathcal{E}'_s(R) := \left\{(\theta_i)_{i\geq 0}, \quad \sum_{i=1}^{+\infty} \frac{l\left(\theta_{m_{i-1}}, \theta_{m_i}\right)}{i^{-s}} \leq R^2\sigma^2\right\} \ .$$

In Section 4.1, we explain why we call this set $\mathcal{E}'_s(R)$ an ellipsoid.

Proposition 3.2. *Assume there exists s, s', and R such that $\theta \in \mathcal{E}'_s(R)$ and such that for any positive numbers R', $\theta \notin \mathcal{E}'_{s'}(R')$. We consider the collection $\mathcal{M}_{\lfloor n/2\rfloor}$ and the penalty $pen(m) = 2\frac{d_m}{n-d_m}$. Then, there exists a constant $L(s,R)$ and a sequence τ_n converging to zero at infinity such that, with probability, at least $1 - L(s,R)\frac{\log n}{n^2}$,*

$$l(\widetilde{\theta},\theta) \leq \left[1+\tau(n)\right] \inf_{m\in\mathcal{M}_{\lfloor n/2\rfloor}} l(\widehat{\theta}_m,\theta) \ . \tag{10}$$

Admittedly, we make n go to the infinity in this proposition but we are still in a high dimensional setting since $p = +\infty$ and since the size of the collection $\mathcal{M}_{\lfloor n/2 \rfloor}$ goes to infinity with n. Let us briefly discuss the assumption on θ. Roughly speaking, it ensures that the oracle model has a dimension not too close to zero (larger than $\log^2(n)$) and small before n (smaller than $n/\log n$). Notice that it is classical to assume that the bias is non-zero for every model m for proving the asymptotic optimality of Mallows' C_p (*cf.* Shibata [31] and Birgé and Massart [9]). Here, we make a stronger assumption because the bound (10) holds in probability and because the design is Gaussian. Moreover, our stronger assumption has already been made by Stone [32] and Arlot [3]. We refer to Arlot [3] Sect.4.1 for a more complete discussion of this assumption.

The choice of the collection $\mathcal{M}_{\lfloor n/2 \rfloor}$ is arbitrary and one can extend it to many collections that satisfy $(\mathbb{H}_{Pol})$ and $(\mathbb{H}_\eta)$. As mentioned in Section 2, the penalty $pen(m) = 2\frac{d_m}{n-d_m}$ corresponds to the FPE model selection procedure. In conclusion, the choice of the FPE criterion turns out to be asymptotically optimal when the complexity of $\mathcal{M}$ is small.

We now underline that the condition $K > 1$ in Theorem 3.1 is almost necessary. Indeed, choosing K smaller than one yields terrible statistical performances.

Proposition 3.3. *Suppose that p is larger than $n/2$. Let us consider the collection $\mathcal{M}_{\lfloor n/2 \rfloor}$ and assume that for some $\nu > 0$,*

$$pen(m) = (1-\nu)\frac{d_m}{n-d_m} \ , \tag{11}$$

for any model $m \in \mathcal{M}_{\lfloor n/2 \rfloor}$. Then given $\delta \in (0,1)$, there exists some $n_0(\nu,\delta)$ only depending on ν and δ such that for $n \geq n_0(\nu,\delta)$,

$$\mathbb{P}_\theta\left[d_{\widehat{m}} \geq \frac{n}{4}\right] \geq 1-\delta \quad \textit{and} \quad \mathbb{E}\left[l(\widetilde{\theta},\theta)\right] \geq l(\theta_{m_{\lfloor n/2 \rfloor}},\theta) + L(\delta,\nu)\sigma^2 \ .$$

If one chooses a too small penalty, then the dimension $d_{\widehat{m}}$ of the selected model is huge and the penalized estimator $\widetilde{\theta}$ performs poorly. The hypothesis $p \geq n/2$ is needed for defining the collection $\mathcal{M}_{\lfloor n/2 \rfloor}$. Once again, the choice of the collection $\mathcal{M}_{\lfloor n/2 \rfloor}$ is rather arbitrary and the result of Proposition 3.3 still holds for collections $\mathcal{M}$ which satisfy $(\mathbb{H}_{Pol})$ and $(\mathbb{H}_\eta)$ and contain at least one model of large dimension. Theorem 3.1 and Proposition 3.3 tell us that $\frac{d_m}{n-d_m}$ is the minimal penalty.

In practice, we advise to choose K between 2 and 3. Admittedly, $K = 2$ is asymptotically optimal by Proposition 3.2. Nevertheless, we have observed on simulations that $K = 3$ gives slightly better results when n is small. For ordered variable selection, we suggest to take the collection $\mathcal{M}_{\lfloor n/2 \rfloor}$.

3.2 A general model selection theorem

In this section, we study the performance of the penalized estimator $\widetilde{\theta}$ for general collections $\mathcal{M}$. Classically, we need to penalize stronger the models m, incorporating the complexity of the collection. As a special case, we shall consider the problem of complete variable selection. This is why we define the collections $\mathcal{M}_p^d$ that consist of all subsets of $\{1,\dots,p\}$ of size less or equal to d.

Definition 3.2. *Given a collection $\mathcal{M}$, we define the function $H(.)$ by*

$$H(d) := \frac{1}{d}\log\left[Card\left(\{m \in \mathcal{M},\, d_m = d\}\right)\right] \ ,$$

for any integer $d \geq 1$.

This function measures the complexity of the collection $\mathcal{M}$. For the collection $\mathcal{M}_p^d$, $H(k)$ is upper bounded by $\log(ep/k)$ for any $k \leq d$ (see Eq.(4.10) in [25]). Contrary to the situation encountered in ordered variable selection, we are not able to consider models of arbitrary dimensions and we shall do the following assumption.

$(\mathbb{H}_{K,\eta})$: Given $K > 1$ and $\eta > 0$, the collection $\mathcal{M}$ and the number η satisfy

$$\forall m \in \mathcal{M}, \quad \frac{\left[1 + \sqrt{2H(d_m)}\right]^2 d_m}{n - d_m} \leq \eta < \eta(K) \ , \tag{12}$$

where $\eta(K)$ is defined as $\eta(K) := [1 - 2(3/(K+2))^{1/6}]^2 \bigvee [1 - (3/K+2)^{1/6}]^2/4$.

The function $\eta(K)$ is positive and increases when K is larger than one. Besides, $\eta(K)$ converges to one when K converges to infinity. We do not claim that the expression of $\eta(K)$ is optimal. We are more interested in its behavior when K is large.

Theorem 3.4. *Let* $K > 1$ *and let* $\eta < \eta(K)$. *Assume that* n *is larger than some quantity* $n_0(K)$ *only depending on* K *and the collection* $\mathcal{M}$ *satisfies* $(\mathbb{H}_{K,\eta})$. *If the penalty* $pen(.)$ *is lower bounded as follows*

$$pen(m) \geq K \frac{d_m}{n - d_m} \left(1 + \sqrt{2H(d_m)}\right)^2 \quad \textit{for any } m \in \mathcal{M} \ , \tag{13}$$

then

$$\mathbb{E}\left[l(\widetilde{\theta}, \theta)\right] \leq L(K,\eta) \inf_{m \in \mathcal{M}} \left\{ l(\theta_m, \theta) + \frac{n - d_m}{n} pen(m) \left[\sigma^2 + l(\theta_m, \theta)\right] \right\} + \tau_n \ , \tag{14}$$

where τ_n *is defined as*

$$\tau_n = \tau_n \left[Var(Y), K, \eta \right] := \sigma^2 \frac{L_1(K,\eta)}{n} + L_2(K,\eta) n^{5/2} Var(Y) \exp\left[-n L_3(K,\eta)\right] \ ,$$

and $L_3(K,\eta)$ *is positive.*

This theorem provides an oracle type inequality of the same type as the one obtained in the Gaussian sequential framework by Birgé and Massart [8]. The risk of the penalized estimator $\widetilde{\theta}$ almost achieves the infimum of the risks plus a penalty term depending on the function $H(.)$. As in Theorem 3.1, the error term $\tau_n \left[\mathrm{Var}(Y), K, \eta\right]$ depends on θ but this part goes exponentially fast to 0 with n.

Comments:

- As for Theorem 3.1, the result holds for arbitrary large p as long as n is larger than the quantity $n_0(K)$ (independent of p). There is no hidden dependency on p except in the complexity function $H(.)$ and Assumption $\mathbb{H}_{K,\eta}$ that we shall discuss for the particular case of complete variable selection. Moreover, one may easily check Assumption $\mathbb{H}_{K,\eta}$ since it only depends on the collection $\mathcal{M}$ and not on some unknown quantity.
- This result (as well as of Theorem 3.1) does not depend at all on the covariance matrix Σ between the covariates.

- The penalty introduced in this theorem only depends on the collection $\mathcal{M}$ and a number $K > 1$. Hence, performing the procedure does not require any knowledge on σ^2, Σ, or θ. We give hints at the end of the section for choosing the constant K.

- Observe that Theorem 3.1 is not just corollary of Theorem 3.4. If we apply Theorem 3.4 to the problem of ordered selection, then the maximal size of the model has to be smaller than $n\frac{\eta(K)}{1+\eta(K)}$, which depends on K and is always smaller than $n/2$. In contrast, Theorem 3.1 handles models of size up to $n-7$.

3.3 Application to complete variable selection

Let us now restate Theorem 3.4 for the particular issue of complete variable selection. Consider $K > 1$, $\eta < \eta(K)$ and $d > 1$ such that $\mathcal{M}_p^d$ satisfies Assumption $(\mathbb{H}_{K,\eta})$. If we take for any model $m \in \mathcal{M}_p^d$ the penalty term

$$pen(m) = K\frac{d_m}{n-d_m}\left[1+\sqrt{2\log\left(\frac{ep}{d_m}\right)}\right]^2 , \tag{15}$$

then we get

$$\mathbb{E}\left[l(\widetilde{\theta},\theta)\right] \leq L(K,\eta)\inf_{m\in\mathcal{M}_p^d}\left\{l(\theta_m,\theta)+\frac{d_m}{n}\log\left(\frac{ep}{d_m}\right)\sigma^2\right\}+\tau_n\left[\mathrm{Var}(Y),K,\eta\right] .$$

We shall prove in Section 4.2, that the term $\log(p/d_m)$ is unavoidable and that the obtained estimator is optimal from a minimax point of view. If the true parameter θ belongs to some unknown model m, then the rates of estimation of $\tilde{\theta}$ is of the order $\frac{d_m}{n}\log(p/d_m)\sigma^2$. Let us compare our result with other procedures.

- The oracle type inequalities look similar to the ones obtained by Birgé and Massart [8], Bunea *et al.* [10] and Baraud *et al.* [4]. However, Birgé and Massart and Bunea *et al.* assume that the variance σ^2 is known. Moreover, Birgé and Massart and Baraud *et al.* only consider a fixed design setting. Yet, Bunea *et al.* allow the design to be random, but they assume that the regression functions are bounded (Assumption A.2 in their paper) which is not the case here. Moreover, they only get risk bounds with respect to the empirical norm $\|.\|_n$ and not the integrated loss $l(.,.)$.

- As mentioned previously, our oracle inequality holds for any covariance matrix Σ. In contrast, Lasso and Dantzig selector estimators have been shown to satisfy oracle inequalities under assumptions on the empirical design $\mathbf{X}$. In [13], Candès and Tao indeed assume that the singular values of $\mathbf{X}$ restricted to any subset of size proportional to the sparsity of θ are bounded away from zero. Bickel *et al.* [5] introduce an extension of this condition prove both for the Lasso and the Dantzig selector. In a recent work [12], Candès and Plan state that if the empirical correlation between the covariates is smaller than $L(\log p)^{-1}$,then the Lasso follows an oracle inequality in a majority of cases. Their condition is in fact almost necessary. On the one hand, they give examples of some low correlated situations, where the Lasso performs poorly. On the other hand, they prove that the Lasso fails to work well if the correlation between the covariates if larger than $L(\log p)^{-1}$. Yet, Candès and Plan consider the loss function $\|\mathbf{X}\widehat{\theta}-\mathbf{X}\theta\|_n^2$, whereas we use the *integrated* loss $l(\widehat{\theta},\theta)$, but this does not really change the impact of their result. We refer to their paper for further details. The main point is that for some correlation structures, our procedure still works well,

whereas the Lasso and the Dantzig selector procedures perform poorly. In many problems such as GGM estimation, the correlation between the covariates may be high and even the relaxed assumptions of Candès and Plan may not be fulfilled. In Section 5, we illustrate this phenomenon by comparing our procedure with the Lasso on numerical examples for independent and highly correlated covariates.

- Suppose that the covariates are independent and that θ belongs to some model m, the rates of convergence of the Lasso is then of the order $\frac{d_m}{n}\log(p)\sigma^2$, whereas ours is $\frac{d_m}{n}\log(p/d_m)\sigma^2$. Consider the case where p, and d_m are of the same order whereas n is large. Our model selection procedure therefore outperforms the Lasso by a $\log(p)$ factor even if the covariates are independent.

- Let us restate Assumption $(\mathbb{H}_{K,\eta})$ for the particular collection $\mathcal{M}_p^d$. Given some $K > 1$ and some $\eta < \eta(K)$, the collection $\mathcal{M}_p^d$ satisfies $(\mathbb{H}_{K,\eta})$ if

$$d \leq \eta \frac{n}{1 + \left[1 + \sqrt{2(1 + \log(p/d))}\right]^2} \ . \tag{16}$$

  If p is much larger than n, the dimension d of the largest model has to be be smaller than the order $\eta\frac{n}{2\log(p)}$. Candès and Plan state a similar condition for the lasso. We believe that this condition is unimprovable. Indeed, Wainwright states in Th.2 of [38] a result going in this sense: it is impossible to estimate reliably the support of a k-sparse vector θ if n is smaller than the order $k\log(p/k)$. If $\log(p)$ is larger than n, then we cannot apply Theorem 3.4. This ultra-high dimensional setting is also not handled by the theory for the Lasso and the Dantzig selector. Finally, if p is of the same order as n, then Condition (16) is satisfied for dimensions d of the same order as n. Hence, our method works well even when the sparsity is of the same order as n, which is not the case for the Lasso or the Dantzig selector.

Let us discuss the practical choice of d and K for complete variable selection. From numerical studies, we advise to take $d \leq \frac{n}{2.5\left[2+\log\left(\frac{p}{n}\vee 1\right)\right]} \wedge p$ even if this quantity is slightly larger than what is ensured by the theory. The practical choice of K depends on the aim of the study. If one aims at minimizing the risk, $K = 1.1$ gives rather good result. A larger K like 1.5 or 2 allows to obtain a more conservative procedure and consequently a lower FDR. We compare these values of K on simulated examples in Section 5.

3.4 Penalties based on a prior distribution

The penalty defined in Theorem 3.4 only depends on the models through their cardinality. However, the methodology developed in the proof may easily extend to the case where the user has some *prior* knowledge of the relevant models. Let $\pi_{\mathcal{M}}$ be a prior probability measure on the collection $\mathcal{M}$. For any non-empty model $m \in \mathcal{M}$, we define l_m by

$$l_m := -\frac{\log\left(\pi_{\mathcal{M}}(m)\right)}{d_m} \ .$$

By convention, we set $l_\emptyset$ to 1. We define in the next proposition penalty functions based on the quantity l_m that allow to get non-asymptotic oracle inequalities.

Assumption $(\mathbb{H}^l_{K,\eta})$: Given $K > 1$ and $\eta > 0$, the collection $\mathcal{M}$, the numbers l_m and the number η satisfy

$$\forall m \in \mathcal{M}, \quad \frac{\left[1+\sqrt{2l_m}\right]^2 d_m}{n-d_m} \leq \eta < \eta(K) \ , \tag{17}$$

where $\eta(K)$ is defined as in $(\mathbb{H}_{K,\eta})$.

Proposition 3.5. *Let $K > 1$ and let $\eta < \eta(K)$. Assume that $n \geq n_O(K)$ and that Assumption $(\mathbb{H}^l_{K,\eta})$ is fulfilled. If the penalty $pen(.)$ is lower bounded as follows*

$$pen(m) \geq K\frac{d_m}{n-d_m}\left(1+\sqrt{2l_m}\right)^2 \quad \textit{for any } m \in \mathcal{M} \setminus \{\emptyset\} \ , \tag{18}$$

then

$$\mathbb{E}\left[l(\widetilde{\theta},\theta)\right] \leq L(K,\eta) \inf_{m\in\mathcal{M}} \left\{ l(\theta_m,\theta) + \frac{n-d_m}{n} pen(m)\left[\sigma^2 + l(\theta_m,\theta)\right] \right\} + \tau_n \ , \tag{19}$$

where $L(K,\eta)$ and τ_n are the same as in Theorem 3.4.

Comments:

- In this proposition, the penalty (18) as well as the risk bound (19) depend on the prior distribution $\pi_{\mathcal{M}}$. In fact, the bound (19) means that $\widetilde{\theta}$ achieves the trade-off between the bias and some prior weight, which is of the order

$$-\log[\pi_{\mathcal{M}}(m)][\sigma^2 + l(\theta_m, \theta)])/n \ .$$

  This emphasizes that $\widetilde{\theta}$ favours models with a high prior probability. Similar risk bounds are obtained in the fixed design regression framework in Birgé and Massart [7].

- If the proofs of Proposition 3.5 and Theorem 3.4 are very similar, Proposition 3.5 does not imply the theorem.

- Roughly speaking, Assumption $(\mathbb{H}^l_{K,\eta})$ requires that the prior probability $\pi_{\mathcal{M}}(m)$ is not exponentially small with respect to n.

4 Minimax lower bounds and Adaptivity

Throughout this section, we emphasize the dependency of the expectations $\mathbb{E}(.)$ and the probabilities $\mathbb{P}(.)$ on θ by writing $\mathbb{E}_\theta$ and $\mathbb{P}_\theta$. We have stated in Section 3 that the penalized estimator $\widetilde{\theta}$ performs almost as well as the best of the estimators $\widehat{\theta}_m$. We now want to compare the risk of $\widetilde{\theta}$ with the risk of any other possible estimator estimator $\widehat{\theta}$. There is no hope to make a pointwise comparison with an arbitrary estimator. Therefore, we classically consider the maximal risk over some suitable subsets Θ of $\mathbb{R}^p$. The *minimax risk* over the set Θ is given by $\inf_{\widehat{\theta}} \sup_{\theta\in\Theta} \mathbb{E}_\theta[l(\widehat{\theta}, \theta)]$, where the infimum is taken over all possible estimators $\widehat{\theta}$ of θ. Then, the estimator $\tilde{\theta}$ is said to be *approximately minimax* with respect to the set Θ if the ratio

$$\frac{\sup_{\theta\in\Theta} \mathbb{E}_\theta \left[l\left(\tilde{\theta}, \theta\right)\right]}{\inf_{\widehat{\theta}} \sup_{\theta\in\Theta} \mathbb{E}_\theta \left[l\left(\widehat{\theta}, \theta\right)\right]}$$

is smaller than a constant that does not depend on σ^2, n, or p. The minimax rates of estimation were extensively studied in the fixed design Gaussian regression framework and we refer for instance to [8] for a detailed discussion. In this section, we apply a classical methodology known as Fano's Lemma in order to derive minimax rates of estimation for ordered and complete variable selection. Then, we deduce adaptive properties of the penalized estimator $\widetilde{\theta}$.

4.1 Adaptivity with respect to ellipsoids

In this section, we prove that the estimator $\widetilde{\theta}$ introduced in Section 3.1 to perform ordered variable selection is adaptive to a large class of ellipsoids.

Definition 4.1. *For any non increasing sequence $(a_i)_{1\le i\le p+1}$ such that $a_1 = 1$ and $a_{p+1} = 0$ and any $R > 0$, we define the ellipsoid $\mathcal{E}_a(R)$ by*

$$\mathcal{E}_a(R) := \left\{ \theta \in \mathbb{R}^p, \sum_{i=1}^{p} \frac{l\left(\theta_{m_{i-1}}, \theta_{m_i}\right)}{a_i^2} \le R^2 \right\} \ .$$

This definition is very similar to the notion of ellipsoids introduced in [36]. Let us explain why we call this set an ellipsoid. Assume for one moment that the $(X_i)_{1\leq i\leq p}$ are independent identically distributed with variance one. In this case, the term $l\left(\theta_{m_{i-1}},\theta_{m_i}\right)$ equals θ_i^2 and the definition of $\mathcal{E}_a(R)$ translates in

$$\mathcal{E}_a(R)=\left\{\theta\in\mathbb{R}^p,\sum_{i=1}^{p}\frac{\theta_i^2}{a_i^2}\leq R^2\right\}\ ,$$

which precisely corresponds to a *classical* definition of an ellipsoid. If the $(X_i)_{1\leq i\leq p}$ are not i.i.d. with unit variance, it is always possible to create a sequence X_i' of i.i.d. standard Gaussian variables by orthonormalizing the X_i using Gram-Schmidt process. If we call θ' the vector in $\mathbb{R}^p$ such that $X\theta=X'\theta'$, then it holds that $l\left(\theta_{m_{i-1}},\theta_{m_i}\right)=\theta_i'^2$. Then, we can express $\mathcal{E}_a(R)$ using the coordinates of θ' as previously:

$$\mathcal{E}_a(R)=\left\{\theta\in\mathbb{R}^p,\sum_{i=1}^{p}\frac{\theta_i'^2}{a_i^2}\leq R^2\right\}\ .$$

The main advantage of this definition is that it does not directly depend on the covariance of $(X_i)_{1\leq i\leq p}$.

Proposition 4.1. *For any sequence $(a_i)_{1\leq i\leq p}$ and any positive number R, the minimax rate of estimation over the ellipsoid $\mathcal{E}_a(R)$ is lower bounded by*

$$\inf_{\widehat{\theta}}\sup_{\theta\in\mathcal{E}_a(R)}\mathbb{E}_\theta\left[l(\widehat{\theta},\theta)\right]\geq L\sup_{1\leq i\leq p}\left[a_i^2R^2\wedge\frac{\sigma^2 i}{n}\right]\ . \tag{20}$$

This result is analogous to the lower bounds obtained in the fixed design regression framework (see e.g. [25] Th. 4.9). Hence, the estimator $\widetilde{\theta}$ built in Section 3.1 is adaptive to a large class of ellipsoids.

Corollary 4.2. *Assume that n is larger than 12. We consider the penalized estimator $\widetilde{\theta}$ with the collection $\mathcal{M}_{\lfloor n/2\rfloor}$ and the penalty $pen(m)=K\frac{d_m}{n-d_m}$. Let $\mathcal{E}_a(R)$ be an ellipsoid whose radius R satisfies $\frac{\sigma^2}{n}\leq R^2\leq\sigma^2n^\beta$ for some $\beta>0$. Then, $\widetilde{\theta}$ is approximately minimax on $\mathcal{E}_a(R)$*

$$\sup_{\theta\in\mathcal{E}_a(R)}l(\widetilde{\theta},\theta)\leq L(K,\beta)\inf_{\widehat{\theta}}\sup_{\theta\in\mathcal{E}_a(R)}\mathbb{E}_\theta\left[l(\widehat{\theta},\theta)\right]\ ,$$

if either $n\geq 2p$ or $a_{\lfloor n/2\rfloor+1}^2R^2\leq\sigma^2/2$.

In the fixed design framework, one may build adaptive estimators to any ellipsoid satisfying $R^2\geq\sigma^2/n$ so that the ellipsoid is not degenerate (see e.g. [25] Sect. 4.3.3). In our setting, when p is small the estimator $\widetilde{\theta}$ is adaptive to all the ellipsoids that have a moderate radius $\sigma^2/n\leq R^2\leq n^\beta$. The technical condition $R^2\leq n^\beta$ is not really restrictive. It comes from the term $n^3l(0_p,\theta)\exp(-nL(K))$ in Theorem 3.1 which goes exponentially fast to 0 with n. When p is larger, $\widetilde{\theta}$ is adaptive to the ellipsoids that also satisfies $a_{\lfloor n/2\rfloor+1}^2R^2\leq\sigma^2/2$. In other words, we require that the ellipsoid is well approximated by the space $S_{m_{\lfloor n/2\rfloor}}$ of vectors θ whose support is included in $\{1,\dots,\lfloor n/2\rfloor\}$. If this condition is not fulfilled, the estimator $\widetilde{\theta}$ is not proved to be minimax on $\mathcal{E}_a(R)$. For such situations, we believe on the one hand that the estimator $\widetilde{\theta}$ should be refined and on the other hand that our lower bounds are not sharp. Finally, the collection $\mathcal{M}_{\lfloor n/2\rfloor}$ may be replaced by any $\mathcal{M}_{\lfloor n\eta\rfloor}$ in Corollary 4.2.

Since the methods used for minimax lower bounds and the oracle inequalities are analogous to the ones in the Gaussian sequence framework, one may also adapt in our setting the arguments developed in [25] Sect. 4.3.5 to derive minimax rates of estimation over other sets such Besov bodies. However, this is not really relevant for the regression model (1).

4.2 Adaptivity with respect to sparsity

Our aim is now to analyze the minimax risk for the complete variable selection problem. Let us fix an integer k between 1 and p. We are interested in estimating the vector θ within the class of vectors with a most k non-zero components. This typically corresponds to the situation encountered in graphical modeling when estimating the neighborhoods of large sparse graphs. As the graph is assumed to be sparse, only a small number of components of θ are non-zero.

In the sequel, the set $\Theta[k,p]$ stands for the subset of vectors $\theta \in \mathbb{R}^p$, such that at most k coordinates of θ are non-zero. For any $r > 0$, we denote $\Theta[k,p](r)$ the subset of $\Theta[k,p]$ such that any component of θ is smaller than r in absolute value.

First, we derive a lower bound for the minimax rates of estimation when the covariates are independent. Then, we prove the estimator $\widetilde{\theta}$ defined with some collection $\mathcal{M}_p^d$ and the penalty (15) is adaptive to any sparse vector θ. Finally, we investigate the minimax rates of estimation for correlated covariates.

Proposition 4.3. *Assume that the covariates X_i are independent and have a unit variance. For any $k \leq p$ and any radius $r > 0$,*

$$\inf_{\widehat{\theta}} \sup_{\theta \in \Theta[k,p](r)} \mathbb{E}_\theta \left[l(\widehat{\theta}, \theta) \right] \geq Lk \left[r^2 \wedge \sigma^2 \frac{1 + \log\left(\frac{p}{k}\right)}{n} \right] . \tag{21}$$

Thanks to Theorem 3.4, we derive the minimax rate of estimation over $\Theta[k,p]$.

Corollary 4.4. *Consider $K > 0$, $\beta > 0$, and $\eta < \eta(K)$. Assume that $n \geq n_0(K)$ and that the covariates X_i are independent and have a unit variance. Let d be a positive integer such that $\mathcal{M}_p^d$ satisfies $(\mathbb{H}_{K,\eta})$. The penalized estimator $\widetilde{\theta}$ defined with the collection $\mathcal{M}_p^d$ and the penalty (15) is adaptive minimax over the sets $\Theta[k,p](n^\beta)$*

$$\sup_{\theta \in \Theta[k,p]} \mathbb{E}_\theta \left[l(\widetilde{\theta}, \theta) \right] \leq L(K, \beta, \eta) \inf_{\widehat{\theta}} \sup_{\theta \in \Theta[k,p](n^\beta)} \mathbb{E}_\theta \left[l(\widehat{\theta}, \theta) \right] ,$$

for any k smaller than d.

Hence, the minimax rates of estimation over $\Theta[k,p](n^\beta)$ is of order $k\frac{\log\left(\frac{ep}{k}\right)}{n}$, which is similar to the rates obtained in the fixed design regression framework. As in previous Section, we restrict ourselves to a radius r in $\Theta[k,p](r)$ smaller than n^β because of the term $\tau_n(\mathrm{Var}(Y), K, \eta)$ which depends on $l(0_p, \theta)$ but goes exponentially fast to 0 when n goes to infinity. Let us interpret Corollary 4.4 with regard to Condition (16). If p is of the same order as n, the estimator $\widetilde{\theta}$ is simultaneously minimax over all sets $\Theta[k,p](n^\beta)$ when k is smaller than a constant times n. If p is much larger than n, the estimator $\widetilde{\theta}$ is simultaneously minimax over all sets $\Theta[k,p](n^\beta)$ with k smaller than $Ln/\log(p)$. We conjecture that the minimax rate of estimation is larger than $k\log(p/k)/n$ when k becomes larger than $n/\log p$. Let us mention that Tsybakov [35] has proved general minimax lower bounds for aggregation in Gaussian random design regression. However, his result does not apply in our Gaussian design setting setting since he assumes that the density

of the covariates X_i is lower bounded by a constant μ_0.

We have proved that the estimator $\widetilde{\theta}$ is adaptive to an unknown sparsity when the covariates are independent. The performance of $\widetilde{\theta}$ exhibited in Theorem 3.4 do not depend on the covariance matrix Σ. Hence, the minimax rates of estimation on $\Theta[k,p]$ is smaller or equal to the order $k\log(p/k)/n$ for any dependence between the covariance. One may then wonder whether the minimax rate of estimation over $\Theta[k,p]$ is not faster when the covariates are correlated. We are unable to derive the minimax rates for a general covariance matrix Σ. This is why we restrict ourselves to particular examples of correlation structures. Let us first consider a pathological situation: Assume that $X_1,\dots,X_k$ are independent and that $X_{k+1},\dots,X_p$ are all equal to X_1. Admittedly, the covariance matrix Σ is henceforth non invertible. In the discussion, we mention that Theorems 3.1 and 3.4 easily extend when Σ is non-invertible if we take into account that the estimators $\widehat{\theta}_m$ and $\widehat{m}$ are non-necessarily uniquely defined. We may derive from Lemma 2.1 that the estimator $\widehat{\theta}_{\{1,\dots,k\}}$ achieves the rate k/n over $\theta[k,p](n^\beta)$. Conversely, the parametric rate k/n is optimal. However, the estimator $\widetilde{\theta}$ defined with the collection $\mathcal{M}_p^k$ and penalty (15) only achieves the rate $k\log(p/k)/n$. Hence, $\widetilde{\theta}$ is not minimax over $\Theta[k,p]$ for this particular covariance matrix and the minimax rate is degenerate. This emergence of faster rates for correlation covariates also occurs for testing problems in the model (1) as stated in [36] Sect. 4.3. This is why we provide sufficient conditions on Σ so that the minimax rate of estimation is still of the same order as in the independent case. In the following proposition, $\|.\|$ refers to the canonical norm in $\mathbb{R}^p$.

Proposition 4.5. *Let Ψ denote the correlation matrix of the covariates $(X_i)_{1\le i\le p}$. Let k be a positive number smaller $p/2$ and let $\delta>0$. Assume that*

$$(1-\delta)^2\|\theta\|^2 \ \le \theta^*\Psi\theta \le (1+\delta)^2\|\theta\|^2\ , \tag{22}$$

for all $\theta\in\mathbb{R}^p$ with at most $2k$ non-zero components. Then, the minimax rate of estimation over $\Theta[k,p](r)$ is lower bounded as follows

$$\inf_{\widehat{\theta}}\sup_{\theta\in\Theta[k,p](r)}\mathbb{E}_\theta\left[l(\widehat{\theta},\theta)\right]\ge L(1-\delta)^2k\left[r^2\wedge\sigma^2\frac{1+\log\left(\frac{p}{k}\right)}{(1+\delta)^2n}\right]\ .$$

Assumption (22) corresponds to the δ-Restricted Isometry Property of order $2k$ introduced by Candès and Tao [14]. Under such a condition, the minimax rates of estimation is the same as the one in the independent case up to a constant depending on δ and the estimator $\widetilde{\theta}$ defined in Corollary 4.4 is still approximately minimax over such sets $\Theta[k,p]$.

However, the δ-Restricted Isometry Property is quite restrictive and seems not to be necessary so that the minimax rate of estimation stays of the order $k\log(p/k)/n$. Besides, in many situations this condition is not fulfilled. Assume for instance that the random vector X is a Gaussian Graphical model with respect to a given sparse graph. We expect that the correlation between two covariates is large if they are neighbors in the graph and small if they are far-off (w.r.t. the graph distance). This is why we derive lower bounds on the rate of estimation for correlation matrices often used to model stationary processes.

Proposition 4.6. *Let $X_1,\dots,X_p$ form a stationary process on the one dimensional torus. More precisely, the correlation between X_i and X_j is a function of $|i-j|_p$ where $|.|_p$ refers to the toroidal distance defined by:*

$$|i-j|_p := (|i-j|)\wedge(p-|i-j|)\ .$$

$\Psi_1(\omega)$ and $\Psi_2(t)$ respectively refer to the correlation matrix of X such that

$$\begin{aligned} \mathrm{corr}(X_i, X_j) &:= \exp(-\omega|i-j|_p) \text{ where } \omega > 0, \\ \mathrm{corr}(X_i, X_j) &:= (1+|i-j|_p)^{-t} \text{ where } t > 0. \end{aligned}$$

Then, the minimax rates of estimation are lower bounded as follows

$$\inf_{\widehat{\theta}} \sup_{\theta \in \Theta[k,p]} \mathbb{E}_{\theta,\Psi_1(\omega)} \left[l(\widehat{\theta}, \theta) \right] \geq L \frac{k\sigma^2}{n} \left[1 + \log \left(\frac{\lfloor p \lceil \log(4k)/\omega \rceil^{-1} \rfloor}{k} \right) \right] ,$$

if k is smaller than $p/\lceil \log(4k)/\omega \rceil$ and

$$\inf_{\widehat{\theta}} \sup_{\theta \in \Theta[k,p]} \mathbb{E}_{\theta,\Psi_2(t)} \left[l(\widehat{\theta}, \theta) \right] \geq L \frac{k\sigma^2}{n} \left[1 + \log \left(\frac{\lfloor p \lceil (4k)^{\frac{1}{t}} - 1 \rceil^{-1} \rfloor}{k} \right) \right] ;$$

if k is smaller than $p/\lceil (4k)^{\frac{1}{t}} - 1 \rceil$.

In the proof of the proposition, we justify that the correlations considered are well-defined at least when p is odd. Let us mention that these correlation models are quite classical when modelling the correlation of time series (see e.g. [20])

If the range ω is larger than $1/p^\gamma$ or if the range t is larger than γ for some $\gamma < 1$, the lower bounds are of order $\sigma^2 \frac{k}{n}(1 + \log p/k)$. As a consequence, for any of these correlation models the minimax rate of estimation is of the same order as the minimax rate of estimation for independent covariates. This means that the estimator $\widetilde{\theta}$ defined in Proposition 4.4 is rate-optimal for these correlations matrices.

In conclusion, the estimator $\widetilde{\theta}$ defined in Corollary 4.4 may not be adaptive to the covariance matrix Σ but rather achieves the minimax rate over all covariance matrices Σ:

$$\sup_{\Sigma \geq 0} \sup_{\theta \in \Theta[k,p](n^\beta)} \mathbb{E}_\theta \left[l(\widetilde{\theta}, \theta) \right] \leq L(K, \beta, \eta) \inf_{\widehat{\theta}} \sup_{\Sigma \geq 0} \sup_{\theta \in \Theta[k,p](n^\beta)} \mathbb{E}_\theta \left[l(\widehat{\theta}, \theta) \right] .$$

Nevertheless, the result makes sense if one considers GGMs since the resulting covariance matrices are typically far from being independent.

5 Numerical study

In this section, we carry out a small simulation study to evaluate the performance of our estimator $\widetilde{\theta}$. As pointed out earlier, an interesting feature of our criterion lies in its flexibility. However, we restrict ourselves here to the variable selection problem. Indeed, it allows to assess the efficiency of our procedure with having regard to the Lasso [34] and adaptive Lasso proposed by Zou [41]. Even if these two procedures assume that the conditional variance σ^2 is known, they give good results in practice and the comparison with our method is of interest. The calculations are made with *R* `www.r-project.org/`.

5.1 Simulation scheme

We consider the regression model (1) with $p = 20$, and $\sigma^2 = 1$. The number of observations n equal 15, 20, and 30. We perform two simulation experiments.

1. First simulation experiment: The covariance matrix Σ_1 is the identity matrix. This corresponds to the situation where the covariates are all independent. The vector θ_1 has all its components to zero except the three first ones, which respectively equal 2, 1, and 0.5.

2. Second simulation experiment: Let A be the $p \times p$ matrix whose lines $(a_1, \dots, a_p)$ are respectively defined by

$$\begin{aligned} a_1 &:= (1, -1, 0, \dots, 0)/\sqrt{2} \\ a_2 &:= (-1, 1.2, 0, \dots, 0)/\sqrt{1 + 1.2^2} \\ a_3 &:= (1/\sqrt{2}, 1/\sqrt{2}, 1/p, \dots, 1/p)/\sqrt{1/2 + (p-2)/p^2}\ , \end{aligned}$$

   and for $4 \leq j \leq p$, a_j corresponds to the j^{th} canonical vector of $\mathbb{R}^p$. Then, we take the covariance matrix $\Sigma_2 = A^*A$ and the vector $\theta_2^* = (40, 40, 0, \dots, 0)$. This choice of parameters derives from the simulation experiments of [4]. Observe that the two first covariates are highly correlated.

For each sample we estimate θ with our procedure, the Lasso and the adaptive Lasso. For our procedure we use the collection $\mathcal{M}_p^3$ for $n = 15$, $\mathcal{M}_p^4$ for $n = 20$ and, $\mathcal{M}_p^5$ for $n = 30$. The choice of smaller collections for $n = 15$ and 20 is due to Condition (16). We take the penalty (15) with $K = 1.1$ 1.5, and 2. For the Lasso and adaptive Lasso procedures, we first normalize the covariates $(\mathbf{X}_i)$. Here, $2\sqrt{\log p}\sigma$ would be a good choice for the parameter λ of the Lasso. However, we do not have access to σ. Hence, we use an estimation of the variance $\widehat{\text{Var}}(Y)$ which is a (possibly inaccurate) upper bound of σ^2. This is why we choose the parameter λ of the Lasso between $0.3 \times 2\sqrt{\log p \widehat{\text{Var}}(Y)}$ and $2\sqrt{\log p \widehat{\text{Var}}(Y)}$ by leave-one-out cross-validation. The number 0.3 is rather arbitrary. In practice, the performances of the Lasso do not really depend on this number as soon it is neither too small nor close to one. For the adaptive Lasso procedure, the parameters γ and λ are also estimated thanks to leave-one-out cross-validation: γ can take three values $(0.5, 1, 2)$ and the values of λ vary between $0.3 \times 2\sqrt{\log p \widehat{\text{Var}}(Y)}$ and $2\sqrt{\log(p) \widehat{\text{Var}}(Y)}$.

We evaluate the risk ratio

$$\text{ratio.Risk} := \frac{\mathbb{E}\left[l(\widehat{\theta}, \theta)\right]}{\inf_{m \in \mathcal{M}_p^5} \mathbb{E}\left[l(\widehat{\theta}_m, \theta)\right]}$$

as well as the power and the FDR on the basis of 1000 simulations. Here, the power corresponds to the fraction of non-zero components θ estimated as non-zero by the estimator $\widehat{\theta}$, while the FDR is the ratio of the false discoveries over the true discoveries.

$$\text{Power} := \mathbb{E}\left[\frac{\text{Card}(\{i,\ \theta_i \neq 0 \text{ and } \widehat{\theta}_i \neq 0\})}{\text{Card}\left(\{i,\ \theta_i \neq 0\}\right)}\right] \quad \text{and} \quad \text{FDR} := \mathbb{E}\left[\frac{\text{Card}(\{i,\ \theta_i = 0 \text{ and } \widehat{\theta}_i \neq 0\})}{\text{Card}(\{i,\ \widehat{\theta}_i \neq 0\})}\right] .$$

5.2 Results

The results of the first simulation experiment are given in Table 1. We observe that the five estimators perform more or less similarly as expected by the theory. The results of the second simulation study are reported in Table 2. Clearly, the Lasso and adaptive Lasso procedures are not consistent in this situation since the power is close to 0 and the FDR is close to one. Consequently, the risk ratio is quite large and the adaptive Lasso even seems unstable. In contrast,

	$n = 15$			$n = 20$		
Estimator	ratio.Risk	Power	FDR	ratio.Risk	Power	FDR
$K = 1.1$	4.8 ± 0.4	0.67 ± 0.02	0.23 ± 0.02	4.8 ± 0.3	0.77 ± 0.01	0.28 ± 0.02
$K = 1.5$	5.7 ± 0.4	0.62 ± 0.02	0.20 ± 0.01	5.3 ± 0.4	0.74 ± 0.02	0.25 ± 0.01
$K = 2$	7.3 ± 0.5	0.54 ± 0.02	0.17 ± 0.01	6.6 ± 0.5	0.68 ± 0.02	0.21 ± 0.01
Lasso	5.8 ± 0.2	0.64 ± 0.01	0.29 ± 0.02	6.0 ± 0.2	0.74 ± 0.01	0.23 ± 0.01
A. Lasso	4.8 ± 0.3	0.64 ± 0.02	0.30 ± 0.02	4.7 ± 0.4	0.75 ± 0.02	0.30 ± 0.01

	$n = 30$		
Estimator	ratio.Risk	Power	FDR
$K = 1.1$	4.2 ± 0.3	0.87 ± 0.01	0.23 ± 0.02
$K = 1.5$	4.1 ± 0.2	0.84 ± 0.01	0.19 ± 0.01
$K = 2$	4.3 ± 0.2	0.81 ± 0.01	0.14 ± 0.01
Lasso	6.6 ± 0.2	0.83 ± 0.01	0.18 ± 0.01
A. Lasso	4.3 ± 0.5	0.86 ± 0.02	0.26 ± 0.01

Table 1: Our procedure with $K = 1.1$, 1.5, and 2 and Lasso and adaptive Lasso procedures: Estimation and 95% confidence interval of Risk ratio (ratio.Risk), Power and FDR when $p = 20$, $\Sigma = \Sigma_2$, $\theta = \theta_2$, and $n = 15$, 20, and 30.

	$n = 15$			$n = 20$		
Estimator	ratio.Risk	Power	FDR	ratio.Risk	Power	FDR
$K = 1.1$	5.3 ± 0.4	0.77 ± 0.03	0.41 ± 0.02	6.4 ± 0.5	0.87 ± 0.02	0.39 ± 0.02
$K = 1.5$	5.3 ± 0.4	0.76 ± 0.03	0.41 ± 0.02	5.9 ± 0.5	0.87 ± 0.02	0.36 ± 0.02
$K = 2$	5.5 ± 0.5	0.75 ± 0.03	0.40 ± 0.02	5.5 ± 0.5	0.86 ± 0.02	0.33 ± 0.02
Lasso	13.5 ± 0.3	0.02 ± 0.01	0.99 ± 0.01	16.7 ± 0.3	0.02 ± 0.01	0.98 ± 0.01
A. Lasso	15.0 ± 1.2	0.02 ± 0.01	0.90 ± 0.02	20.5 ± 1.8	0.04 ± 0.01	0.89 ± 0.02

	$n = 30$		
Estimator	ratio.Risk	Power	FDR
$K = 1.1$	4.5 ± 0.3	0.96 ± 0.02	0.24 ± 0.02
$K = 1.5$	3.9 ± 0.3	0.95 ± 0.01	0.19 ± 0.02
$K = 2$	3.5 ± 0.3	0.94 ± 0.01	0.16 ± 0.02
Lasso	22.0 ± 0.3	0.02 ± 0.01	0.99 ± 0.01
A. Lasso	31.8 ± 3.0	0.04 ± 0.01	0.88 ± 0.02

Table 2: Our procedure with $K = 1.1$, 1.5, and 2 and Lasso and adaptive Lasso procedures: Estimation and 95% confidence interval of Risk ratio (ratio.Risk), Power and FDR when $p = 20$, $\Sigma = \Sigma_1$, $\theta = \theta_1$, and $n = 15$, 20, and 30.

our method exhibits a large power and a reasonable FDR.

In the two studies, choosing a larger K reduces the power of the estimator but also decreases the FDR. It seems that the choice $K = 1.1$ yields a good risk ratio, whereas $K = 2$ gives a better control of the FDR. Contrary to the parameter λ for the lasso, we do not need an *ad-hoc* method such as cross-validation to calibrate K. The second example is certainly quite pathological but it illustrates that our estimator $\widetilde{\theta}$ performs well even when the Lasso does not provide an accurate

estimation. The good behavior of our method illustrates the strength of Theorem 3.4 that does not depend on the correlation of the explanatory variables.

6 Discussion and concluding remarks

Until now, we have assumed that the covariance matrix Σ of the covariates is non-singular. If Σ is singular, the estimators $\widehat{\theta}_m$ and the model $\widehat{m}$ are not necessarily uniquely defined. However, upon defining $\widehat{\theta}_m$ as *one* of the minimizers of $\gamma_n(\theta')$ over S_m, one may readily extend the oracle inequalities stated in Theorem 3.1 and 3.4.

Let us recall the main features of our method. We have defined a model selection criterion that satisfies oracle inequalities regardless of the correlation between the covariates and regardless of the collection of models. Hence, the estimator $\widetilde{\theta}$ achieves nice adaptive properties for ordered variable selection or for complete variable selection. Besides, one can easily combine this method with prior knowledge on the model by choosing a proper collection $\mathcal{M}$ or by modulating the penalty $pen(.)$. Moreover, we may easily calibrate the penalty even when σ^2 is unknown, whereas the Lasso-type procedures require a cross-validation strategy to choose the parameter λ. The compensation for these nice properties is a computational cost that depends linearly on the size of $\mathcal{M}$. Hence, the complete variable selection problem is NP-hard. This makes it intractable when p becomes too large (i.e. more than 50). In contrast, our criterion applies for arbitrary p when considering ordered variable selection since the size of $\mathcal{M}$ is linear with n. In situations where one has a good prior knowledge on the true model, the collection $\mathcal{M}$ is then not too large and our criterion is also fastly calculable even for large p.

For complete variable selection, Lasso-type procedures are computationally feasible even when p is large and achieve oracle inequalities under assumptions on the covariance structure. However, there are both theoretical and practical problems with these estimators. On the one hand, they are known to perform poorly for some covariance structures. On the other hand, there is some room for improvement in the practical calibration of the lasso, especially when σ^2 is unknown. In a future work, we would like to combine the strength of our method with these computationally fast algorithms. The problem at hand is to design a fast data-driven method that picks a subcollection $\widehat{\mathcal{M}}$ of reasonable size. Afterwards, one applies our procedure to $\widehat{\mathcal{M}}$ instead of $\mathcal{M}$. A direction that needs further investigation is taking for $\widehat{M}$ all the subsets of the regularization path given by the lasso.

7 Proofs

7.1 Some notations and probabilistic tools

First, let us define the random variable ϵ_m by

$$Y \quad = \quad X\theta_m + \epsilon_m + \epsilon \ a.s. \ . \tag{23}$$

By definition of θ_m, ϵ_m follows a normal distribution and is independent of ϵ and of X_m. Hence, the variance of ϵ_m equals $l(\theta_m, \theta)$. The vectors $\boldsymbol{\epsilon}$ and $\boldsymbol{\epsilon}_m$ refer to the n samples of ϵ and ϵ_m. For any model m and any vector Z of size n, $\Pi_m^{\perp} Z$ stands for $Z - \Pi_m Z$. For any subset m of $\{1, \ldots, p\}$, Σ_m denotes the covariance matrix of the vector X_m^*. Moreover, we define the row vector $Z_m := X_m \sqrt{\Sigma_m^{-1}}$ in order to deal with standard Gaussian vectors. Similarly to the matrix $\mathbf{X}_m$, the $n \times d_m$ matrix $\mathbf{Z}_m$ stands for the n observations of Z_m. The notation $\langle ., . \rangle_n$ refers to

the empirical inner product associated with the norm $\|.\|_n$. Lastly, $\varphi_{\max}(A)$ denotes the largest eigenvalue (in absolute value) of a symmetric square matrix A.

We shall extensively use the explicit expression of $\widehat{\theta}_m$:

$$\mathbf{X}\widehat{\theta}_m = \mathbf{X}_m(\mathbf{X}_m^*\mathbf{X}_m)^{-1}\mathbf{X}_m^*\mathbf{Y} \ . \tag{24}$$

Let us state a first lemma that gives the expressions of $\gamma_n(\widehat{\theta}_m)$, $\gamma(\widehat{\theta}_m)$, and the loss $l(\widehat{\theta}_m, \theta_m)$.

Lemma 7.1. *For any model m of size smaller than n,*

$$\gamma_n\left(\widehat{\theta}_m\right) = \|\Pi_m^{\perp}\left(\boldsymbol{\epsilon} + \boldsymbol{\epsilon}_m\right)\|_n^2 \ , \tag{25}$$

$$\gamma\left(\widehat{\theta}_m\right) = \sigma^2 + l(\theta_m, \theta) + l(\widehat{\theta}_m, \theta_m) \ , \tag{26}$$

$$l(\widehat{\theta}_m, \theta_m) = (\boldsymbol{\epsilon} + \boldsymbol{\epsilon}_m)^*\mathbf{Z}_m(\mathbf{Z}_m^*\mathbf{Z}_m)^{-2}\mathbf{Z}_m^*(\boldsymbol{\epsilon} + \boldsymbol{\epsilon}_m) \ . \tag{27}$$

The proof is postponed to the Appendix.

We now introduce the main probabilistic tools used throughout the proofs. First, we need to bound the deviations of χ^2 random variables.

Lemma 7.2. *For any integer $d > 0$ and any positive number x,*

$$\mathbb{P}\left(\chi^2(d) \leq d - 2\sqrt{dx}\right) \leq \exp(-x) \ ,$$

$$\mathbb{P}\left(\chi^2(d) \geq d + 2\sqrt{dx} + 2x\right) \leq \exp(-x) \ .$$

These bounds are classical and are shown by applying Laplace method. We refer to Lemma 1 in [22] for more details. Moreover, we state a refined bound for the lower deviations of a χ^2 distribution.

Lemma 7.3. *For any integer $d > 0$ and any positive number x,*

$$\mathbb{P}\left[\chi^2(d) \leq d\left[\left(1 - \delta_d - \sqrt{\frac{2x}{d}}\right) \vee 0\right]^2\right] \leq \exp(-x) \ ,$$

$$\textit{where } \delta_d := \sqrt{\frac{\pi}{2d}} + \exp(-d/16) \ . \tag{28}$$

The proof is postponed the Appendix. Finally, we shall bound the largest eigenvalue of standard Wishart matrices and standard inverse Wishart matrices. The following deviation inequality is taken from Theorem 2.13 in [17].

Lemma 7.4. *Let Z^*Z be a standard Wishart matrix of parameters (n, d) with $n > d$. For any positive number x,*

$$\mathbb{P}\left\{\varphi_{\max}\left[(Z^*Z)^{-1}\right] \geq \left[n\left(1 - \sqrt{\frac{d}{n}} - x\right)^2\right]^{-1}\right\} \leq \exp(-nx^2/2) \ ,$$

and

$$\mathbb{P}\left[\varphi_{\max}\left(Z^*Z\right) \leq n\left(1 + \sqrt{\frac{d}{n}} + x\right)^2\right] \leq \exp(-nx^2/2) \ .$$

7.2 Proof of Theorem 3.1

Proof of Theorem 3.1. For the sake of simplicity we divide the main steps of the proof in several lemmas. First, let us fix a model m in the collection $\mathcal{M}$. By definition of $\widehat{m}$, we know that

$$\gamma_n(\widetilde{\theta})\left[1+pen(\widehat{m})\right] \leq \gamma_n(\theta_m)\left[1+pen(m)\right] \ .$$

Subtracting $\gamma(\theta)$ to both sides of this inequality yields

$$l(\widetilde{\theta},\theta) \leq l(\theta_m,\theta) + \gamma_n(\theta_m)pen(m) + \overline{\gamma}_n(\theta_m) - \gamma_n(\widetilde{\theta})pen(\widehat{m}) - \overline{\gamma}_n(\widetilde{\theta}) \ , \tag{29}$$

where $\overline{\gamma}_n(.) := \gamma_n(.) - \gamma(.)$. The proof is based on the concentration of the term $-\overline{\gamma}_n(\widetilde{\theta})$. More precisely, we shall prove that with overwhelming probability this quantity is of the same order as the penalty term $\gamma_n(\widetilde{\theta})pen(\widehat{m})$.

Let κ_1 and κ_2 be two positive numbers smaller than one that we shall fix later. For any model $m' \in \mathcal{M}$, we introduce the random variables $A_{m'}$ and $B_{m'}$ as

$$\begin{aligned} A_{m'} &:= \kappa_1 + 1 - \frac{\|\Pi^{\perp}_{m'}\boldsymbol{\epsilon}_{m'}\|_n^2}{l(\theta_{m'},\theta)} + \kappa_2 n\varphi_{\max}\left[(\mathbf{Z}^*_{m'}\mathbf{Z}_{m'})^{-1}\right]\frac{\|\Pi_m(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_{m'})\|_n^2}{l(\theta_{m'},\theta)+\sigma^2} \\ &- K\frac{d_{m'}}{n-d_{m'}}\frac{\|\Pi_{m'^{\perp}}(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_{m'})\|_n^2}{l(\theta_{m'},\theta)+\sigma^2} \ , \end{aligned} \tag{30}$$

$$\begin{aligned} B_{m'} &:= \kappa_1^{-1}\frac{\langle \Pi^{\perp}_{m'}\boldsymbol{\epsilon}, \Pi^{\perp}_{m'}\boldsymbol{\epsilon}_{m'}\rangle_n^2}{\sigma^2 l(\theta_{m'},\theta)} + \frac{\|\Pi_{m'}\boldsymbol{\epsilon}\|_n^2}{\sigma^2} + \kappa_2 n\varphi_{\max}\left[(\mathbf{Z}^*_{m'}\mathbf{Z}_{m'})^{-1}\right]\frac{\|\Pi_{m'}(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_{m'})\|_n^2}{l(\theta_{m'},\theta)+\sigma^2} \\ &- K\frac{d_{m'}}{n-d_{m'}}\frac{\|\Pi^{\perp}_{m'}(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_{m'})\|_n^2}{l(\theta_{m'},\theta)+\sigma^2} \ . \end{aligned} \tag{31}$$

We recall that the notations ϵ_m, Z_m, $\langle .,. \rangle_n$, and $\varphi_{\max}(.)$ are defined in Section 7.1. We may upper bound the expression $-\overline{\gamma}_n(\widetilde{\theta}) - \gamma_n(\widetilde{\theta})pen(\widehat{m})$ with respect to $A_{\widehat{m}}$ and $B_{\widehat{m}}$ as follows.

Lemma 7.5. *Almost surely, it holds that*

$$-\overline{\gamma}_n(\widetilde{\theta}) - \gamma_n(\widetilde{\theta})pen(\widehat{m}) - \sigma^2 + \|\boldsymbol{\epsilon}\|_n^2 \leq l(\widetilde{\theta},\theta)\left[A_{\widehat{m}} \vee (1-\kappa_2)\right] + \sigma^2 B_{\widehat{m}} \ . \tag{32}$$

Let us set the constants

$$\kappa_1 := \frac{1}{4} \quad \text{and} \quad \kappa_2 := \frac{(K-1)(1-\sqrt{\eta})^2}{16} \wedge 1 \ . \tag{33}$$

We do not claim that this choice is optimal, but we are not really concerned about the constants for this result. The core of this proof consists in showing that with overwhelming probability the variable $A_{\widehat{m}}$ is smaller than 1 and $B_{\widehat{m}}$ is smaller than a constant over n.

Lemma 7.6. *The event Ω_1 defined as*

$$\Omega_1 := \left\{A_{\widehat{m}} \leq \frac{7}{8}\right\} \bigcap \left\{\kappa_2 n\varphi_{max}\left[(\mathbf{Z}^*_{\widehat{m}}\mathbf{Z}_{\widehat{m}})^{-1}\right] \leq \frac{K-1}{4}\right\}$$

satisfies $\mathbb{P}(\Omega_1^c) \leq L\mathrm{Card}(\mathcal{M})\exp\left[-nL'(K,\eta)\right]$, where $L'(K,\eta)$ is positive.

Lemma 7.7. *There exists an event Ω_2 of probability larger than $1-\exp(-nL)$ with $L>0$ such that*

$$\mathbb{E}\left[B_{\widehat{m}}\mathbf{1}_{\Omega_1\cap\Omega_2}\right] \leq \frac{L(K,\eta,\alpha,\beta)}{n} \ .$$

Gathering the upper bound (29) and Lemma 7.5, 7.6, and 7.7, we conclude that

$$\begin{aligned}\mathbb{E}\left[l(\widetilde{\theta},\theta)\mathbf{1}_{\Omega_1\cap\Omega_2}\left(\kappa_2\wedge\frac{1}{8}\right)\right] &\leq l(\theta_m,\theta)+\mathbb{E}\left[\gamma_n(\theta_m)pen(m)\right]\\ &+ \sigma^2\frac{L(K,\eta,\alpha,\beta)}{n}+\mathbb{E}\left[\mathbf{1}_{\Omega_1\cap\Omega_2}\left(\overline{\gamma}_n\left(\theta_m\right)+\sigma^2-\|\boldsymbol{\epsilon}\|_n^2\right)\right]\ .\end{aligned}$$

As the expectation of the random variable $\overline{\gamma}_n\left(\theta_m\right)+\sigma^2-\|\boldsymbol{\epsilon}\|_n^2$ is zero, it holds that

$$\begin{aligned}\mathbb{E}\left[\mathbf{1}_{\Omega_1\cap\Omega_2}\left(\overline{\gamma}_n\left(\theta_m\right)+\sigma^2-\|\boldsymbol{\epsilon}\|_n^2\right)\right] &= \mathbb{E}\left[\mathbf{1}_{\Omega_1^c\cup\Omega_2^c}\left(\overline{\gamma}_n\left(\theta_m\right)+\sigma^2-\|\boldsymbol{\epsilon}\|_n^2\right)\right]\\ &\leq \sqrt{\mathbb{P}(\Omega_1^c)+\mathbb{P}(\Omega_2^c)}\left[\sqrt{\mathbb{E}\left[\|\boldsymbol{\epsilon}_m\|_n^2-l(\theta_m,\theta)\right]^2}+2\sqrt{\mathbb{E}\left[\langle\boldsymbol{\epsilon},\boldsymbol{\epsilon}_m\rangle_n^2\right]}\right]\\ &\leq \sqrt{\mathbb{P}(\Omega_1^c)+\mathbb{P}(\Omega_2^c)}\sqrt{\frac{2}{n}}\left[l(\theta_m,\theta)+\sigma\sqrt{2l(\theta_m,\theta)}\right]\ .\end{aligned}$$

The probabilities $\mathbb{P}(\Omega_1^c)$ and $\mathbb{P}(\Omega_2^c)$ converge to 0 at an exponential rate with respect to n. Hence, by taking the infimum over all the models $m\in\mathcal{M}$, we obtain

$$\begin{aligned}\mathbb{E}\left[l(\widetilde{\theta},\theta)\mathbf{1}_{\Omega_1\cap\Omega_2}\right] &\leq L(K,\eta)\inf_{m\in\mathcal{M}}\left[l(\theta_m,\theta)+\left(\sigma^2+l(\theta_m,\theta)\right)pen(m)\right]+L_2(K,\eta,\alpha,\beta)\frac{\sigma^2}{n}+\\ &+ L_3(K,\eta)\sqrt{\frac{\mathrm{Card}(\mathcal{M})}{n}}\left[\sigma^2+l(0_p,\theta)\right]\exp\left[-nL_4(K,\eta)\right]\ , \qquad (34)\end{aligned}$$

with $L_4(K,\eta)>0$. In order to conclude, we need to control the loss of the estimator $\widetilde{\theta}$ on the event of small probability $\Omega_1^c\cup\Omega_2^c$. Thanks to the following lemma, we may upper bound the r-th risk of the estimators $\widehat{\theta}_m$.

Proposition 7.8. *For any model m and any integer $r\geq 2$ such that $n-d_m-2r+1>0$,*

$$\mathbb{E}\left[l(\widehat{\theta}_m,\theta_m)^r\right]^{\frac{1}{r}}\leq Lrd_mn\left[\sigma^2+l(\theta_m,\theta)\right]\ .$$

The proof is postponed to Section 7.4. We derive from this bound a strong control on $\mathbb{E}\left[l(\widetilde{\theta},\theta)\mathbf{1}_{\Omega_1^c\cup\Omega_2^c}\right]$.

Lemma 7.9.

$$\mathbb{E}\left[l(\widetilde{\theta},\theta)\mathbf{1}_{\Omega_1^c\cup\Omega_2^c}\right]\leq L(K,\eta)n^2\,Card(\mathcal{M})\,Var(Y)\exp\left[-nL'(K,\eta)\right]\ , \qquad (35)$$

where $L'(K,\eta)$ is positive.

By Assumptions $(\mathbb{H}_{Pol})$ and $(\mathbb{H}_\eta)$, the cardinality of the collection of $\mathcal{M}$ is smaller than $\alpha n^{1+\beta}$. We gather the upper bounds (34) and (35) and so we conclude. □

Proof of Lemma 7.5. Thanks to Lemma 7.1, we decompose $\overline{\gamma}_n(\widetilde{\theta})$ as

$$\overline{\gamma}_n(\widetilde{\theta})=\|\Pi_{\widehat{m}}^{\perp}(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_{\widehat{m}})\|_n^2-\sigma^2-l(\theta_{\widehat{m}},\theta)-(1-\kappa_2)l(\widetilde{\theta},\theta_{\widehat{m}})-\kappa_2(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_{\widehat{m}})^*\mathbf{Z}_{\widehat{m}}(\mathbf{Z}_{\widehat{m}}^*\mathbf{Z}_{\widehat{m}})^{-2}\mathbf{Z}_{\widehat{m}}^*(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_{\widehat{m}})\ .$$

Since $2ab \leq \kappa_1 a^2 + \kappa_1^{-1} b^2$ for any $\kappa_1 > 0$, it holds that

$$\begin{aligned} -\|\Pi_{\widehat{m}}^{\perp}(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_{\widehat{m}})\|_n^2 + \|\boldsymbol{\epsilon}\|_n^2 &= \|\Pi_{\widehat{m}}\boldsymbol{\epsilon}\|_n^2 - \|\Pi_{\widehat{m}}^{\perp}\boldsymbol{\epsilon}_{\widehat{m}}\|_n^2 - 2\langle \Pi_{\widehat{m}}^{\perp}\boldsymbol{\epsilon}, \Pi_{\widehat{m}}^{\perp}\boldsymbol{\epsilon}_{\widehat{m}}\rangle_n \\ &\leq \sigma^2 \left[\kappa_1^{-1} \frac{\langle \Pi_{\widehat{m}}^{\perp}\boldsymbol{\epsilon}, \Pi_{\widehat{m}}^{\perp}\boldsymbol{\epsilon}_{\widehat{m}}\rangle_n^2}{\sigma^2 l(\theta_{\widehat{m}},\theta)} + \frac{\|\Pi_{\widehat{m}}\boldsymbol{\epsilon}\|_n^2}{\sigma^2}\right] + l(\theta_{\widehat{m}},\theta)\left[-\frac{\|\Pi_{\widehat{m}}^{\perp}\boldsymbol{\epsilon}_{\widehat{m}}\|_n^2}{l(\theta_{\widehat{m}},\theta)} + \kappa_1\right] . \end{aligned}$$

Besides, we upper bound Expression (27) of $l(\widetilde{\theta},\theta_{\widehat{m}})$ using the largest eigenvalue of $(\mathbf{Z}_{\widehat{m}}^* \mathbf{Z}_{\widehat{m}})^{-1}$.

$$\begin{aligned} (\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_{\widehat{m}})^* \mathbf{Z}_{\widehat{m}} (\mathbf{Z}_{\widehat{m}}^* \mathbf{Z}_{\widehat{m}})^{-2} \mathbf{Z}_{\widehat{m}}^* (\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_{\widehat{m}}) &\leq \varphi_{\max}\left[(\mathbf{Z}_{\widehat{m}}^* \mathbf{Z}_{\widehat{m}})^{-1}\right] (\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_{\widehat{m}})^* \mathbf{Z}_{\widehat{m}} (\mathbf{Z}_{\widehat{m}}^* \mathbf{Z}_{\widehat{m}})^{-1} \mathbf{Z}_{\widehat{m}}^* (\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_{\widehat{m}}) \\ &\leq \left[\sigma^2 + l(\theta_{\widehat{m}},\theta)\right] n \varphi_{\max}\left[(\mathbf{Z}_{\widehat{m}}^* \mathbf{Z}_{\widehat{m}})^{-1}\right] \frac{\|\Pi_{\widehat{m}}(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_{\widehat{m}})\|_n^2}{\sigma^2 + l(\theta_{\widehat{m}},\theta)} \end{aligned} \quad (36)$$

Thanks to Assumption (8), we upper bound the penalty terms as follows:

$$-\gamma_n(\widetilde{\theta}) pen(\widehat{m}) \leq -\left[\sigma^2 + l(\theta_{\widehat{m}},\theta)\right] \frac{\|\Pi_{\widehat{m}}^{\perp}(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_{\widehat{m}})\|_n^2}{\sigma^2 + l(\theta_{\widehat{m}},\theta)} K \frac{d_{\widehat{m}}}{n - d_{\widehat{m}}} .$$

By gathering the four last identities, we get

$$-\overline{\gamma}_n(\widetilde{\theta}) - \gamma_n(\widetilde{\theta}) pen(\widehat{m}) - \sigma^2 + \|\boldsymbol{\epsilon}\|_n^2 \leq l(\widetilde{\theta},\theta)\left[A_{\widehat{m}} \vee (1-\kappa_2)\right] + \sigma^2 B_{\widehat{m}} ,$$

since $l(\widetilde{\theta},\theta)$ decomposes into the sum $l(\widetilde{\theta},\theta_{\widehat{m}}) + l(\theta_{\widehat{m}},\theta)$. □

Proof of Lemma 7.6. We recall that for any model $m \in \mathcal{M}$,

$$\begin{aligned} A_m &:= \frac{5}{4} - \frac{\|\Pi_m^{\perp}\boldsymbol{\epsilon}_m\|_n^2}{l(\theta_m,\theta)} + \kappa_2 n \varphi_{\max}\left[(\mathbf{Z}_m^* \mathbf{Z}_m)^{-1}\right] \frac{\|\Pi_m(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_m)\|_n^2}{l(\theta_m,\theta)+\sigma^2} \\ &- K \frac{d_m}{n-d_m} \frac{\|\Pi_m^{\perp}(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_m)\|_n^2}{l(\theta_m,\theta)+\sigma^2} . \end{aligned}$$

In order to control the variable $A_{\widehat{m}}$, we shall simultaneously bound the deviations of the four random variables involved in any variable A_m.

Since $\mathbf{X}_m$ is independent of $\boldsymbol{\epsilon}_m/\sqrt{l(\theta_m,\theta)}$ and since $\boldsymbol{\epsilon}_m/\sqrt{l(\theta_m,\theta)}$ is a standard Gaussian vector of size n, the random variable $n\|\Pi_m^{\perp}\boldsymbol{\epsilon}_m\|_n^2/l(\theta_m,\theta)$ follows a χ^2 distribution with $n - d_m$ degrees of freedom conditionally on $\mathbf{X}_m$. As this distribution does not depend on $\mathbf{X}_m$, $n\|\Pi_m^{\perp}\boldsymbol{\epsilon}_m\|_n^2/l(\theta_m,\theta)$ follows a χ^2 distribution with $n - d_m$ degrees of freedom. Similarly, the random variables $n\|\Pi_m(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_m)\|_n^2/[l(\theta_m,\theta)+\sigma^2]$ and $n\|\Pi_m^{\perp}(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_m)\|_n^2/[l(\theta_m,\theta)+\sigma^2]$ follow χ^2 distributions with respectively d_m and $n - d_m$ degrees of freedom. Besides, the matrix $(\mathbf{Z}_m^* \mathbf{Z}_m)$ follows a standard Wishart distribution with parameters (n, d_m).

Let x be a positive number we shall fix later. By Lemma 7.2 and 7.4, there exists an event Ω_1' of large probability

$$P(\Omega_1^{'c}) \leq 4\exp(-nx)\mathrm{Card}(\mathcal{M}) ,$$

such that for conditionally on Ω_1',

$$\frac{\|\Pi_m^\perp \boldsymbol{\epsilon}_m\|_n^2}{l(\theta_m,\theta)} \geq \frac{n-d_m}{n} - 2\sqrt{\frac{(n-d_m)x}{n}} , \tag{37}$$

$$\frac{\|\Pi_m(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_m)\|_n^2}{\sigma^2+l(\theta_m,\theta)} \leq \frac{d_m}{n} + 2\sqrt{\frac{d_m x}{n}} + 2x , \tag{38}$$

$$\frac{\|\Pi_m^\perp(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_m)\|_n^2}{\sigma^2+l(\theta_m,\theta)} \geq \frac{n-d_m}{n} - 2\sqrt{\frac{(n-d_m)x}{n}} , \tag{39}$$

$$\varphi_{\max}\left[(\mathbf{Z}_m^*\mathbf{Z}_m)^{-1}\right] \leq \left\{ n\left[\left(1-\sqrt{\frac{d_m}{n}}-\sqrt{2x}\right)\vee 0\right]^2\right\}^{-1} , \tag{40}$$

for every model $m \in \mathcal{M}$. Let us prove that for a suitable choice of the number x, $A_{\widehat{m}}\mathbf{1}_{\Omega_1'}$ is smaller than $7/8$. First, we constrain $n\kappa_2\varphi_{\max}\left[(\mathbf{Z}_{\widehat{m}}^*\mathbf{Z}_{\widehat{m}})^{-1}\right]$ to be smaller than $\frac{K-1}{4}$ on the event Ω_1'. By (40), it holds that

$$n\varphi_{\max}\left[(\mathbf{Z}_{\widehat{m}}^*\mathbf{Z}_{\widehat{m}})^{-1}\right] \leq \left[\left(1-\sqrt{\eta}-\sqrt{2x}\right)\vee 0\right]^{-2} .$$

Constraining x to be smaller than $\frac{\left(1-\sqrt{\eta}\right)^2}{8}$ ensures that the largest eigenvalue of $(\mathbf{Z}_{\widehat{m}}^*\mathbf{Z}_{\widehat{m}})^{-1}$ satisfies

$$n\varphi_{\max}\left[(\mathbf{Z}_{\widehat{m}}^*\mathbf{Z}_{\widehat{m}})^{-1}\right] \leq \frac{4}{\left(1-\sqrt{\eta}\right)^2} .$$

By definition (33) of κ_2, it follows that $n\kappa_2\varphi_{\max}\left[(\mathbf{Z}_{\widehat{m}}^*\mathbf{Z}_{\widehat{m}})^{-1}\right] \leq (K-1)/4$. Applying inequality $2ab \leq \delta a^2 + \delta^{-1}b^2$ to the bounds (37), (38), and (39) yields

$$\begin{aligned}
-\frac{\|\Pi_{\widehat{m}}^\perp \boldsymbol{\epsilon}_{\widehat{m}}\|_n^2}{l(\theta_{\widehat{m}},\theta)} &\leq -\frac{1}{2} + \frac{d_{\widehat{m}}}{2n} + 2x \\
\kappa_2 n\varphi_{\max}\left[(\mathbf{Z}_{\widehat{m}}^*\mathbf{Z}_{\widehat{m}})^{-1}\right]\frac{\|\Pi_{\widehat{m}}(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_{\widehat{m}})\|_n^2}{\sigma^2+l(\theta_{\widehat{m}},\theta)} &\leq \frac{K-1}{2}\left[\frac{d_{\widehat{m}}}{n}+\frac{3x}{2}\right] \\
-K\frac{d_{\widehat{m}}}{n-d\widehat{m}}\frac{\|\Pi_{\widehat{m}}^\perp(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_{\widehat{m}})\|_n^2}{\sigma^2+l(\theta_{\widehat{m}},\theta)} &\leq -K\frac{d_{\widehat{m}}}{2n} + x\frac{2K\eta}{1-\eta} .
\end{aligned}$$

Gathering these three inequalities, we get

$$A_{\widehat{m}}\mathbf{1}_{\Omega_1'} \leq \frac{3}{4} + x\left[2+\frac{3(K-1)}{4}+2K\frac{\eta}{1-\eta}\right] .$$

If we set x to

$$x := \left[8\left(2+\frac{3(K-1)}{4}+2K\frac{\eta}{1-\eta}\right)\right]^{-1} \wedge \frac{\left(1-\sqrt{\eta}\right)^2}{8} ,$$

then $A_{\widehat{m}}\mathbf{1}_{\Omega_1'}$ is smaller than $\frac{7}{8}$ and the result follows. □

Proof of Lemma 7.7. We shall simultaneously bound the deviations of the random variables involved in the definition of B_m for all models $m \in \mathcal{M}$. Let us first define the random variable E_m as

$$E_m := \kappa_1^{-1}\frac{\langle \Pi_m^\perp\boldsymbol{\epsilon}, \Pi_m^\perp\boldsymbol{\epsilon}_m\rangle_n^2}{\sigma^2 l(\theta_m,\theta)} + \frac{\|\Pi_m\boldsymbol{\epsilon}\|_n^2}{\sigma^2} .$$

Factorizing by the norm of $\boldsymbol{\epsilon}$, we get

$$E_m \leq \kappa_1^{-1} \frac{\|\boldsymbol{\epsilon}\|_n^2}{\sigma^2} \frac{\langle \frac{\Pi_m^\perp \boldsymbol{\epsilon}}{\|\Pi_m^\perp \boldsymbol{\epsilon}\|_n}, \Pi_m^\perp \boldsymbol{\epsilon}_m \rangle_n^2}{l(\theta_m, \theta)} + \frac{\|\Pi_m \boldsymbol{\epsilon}\|_n^2}{\sigma^2} \ . \tag{41}$$

The variable $n\frac{\|\boldsymbol{\epsilon}\|_n^2}{\sigma^2}$ follows a χ^2 distribution with n degrees of freedom. By Lemma 7.2 there exists an event Ω_2 of probability larger than $1 - \exp(n/8)$ such that $\frac{\|\boldsymbol{\epsilon}\|_n^2}{\sigma^2}$ is smaller than 2. As $\kappa_1^{-1} = 4$, we obtain

$$E_m \mathbf{1}_{\Omega_2} \leq 8 \frac{\langle \frac{\Pi_m^\perp \boldsymbol{\epsilon}}{\|\Pi_m^\perp \boldsymbol{\epsilon}\|_n}, \Pi_m^\perp \boldsymbol{\epsilon}_m \rangle_n^2}{l(\theta_m, \theta)} + \frac{\|\Pi_m \boldsymbol{\epsilon}\|_n^2}{\sigma^2} \ .$$

Since $\boldsymbol{\epsilon}$, $\boldsymbol{\epsilon}_m$, and $\mathbf{X}_m$ are independent, it holds that conditionally on $\mathbf{X}_m$ and $\boldsymbol{\epsilon}$,

$$n \frac{\langle \frac{\Pi_m^\perp \boldsymbol{\epsilon}}{\|\Pi_m^\perp \boldsymbol{\epsilon}\|_n}, \Pi_m^\perp \boldsymbol{\epsilon}_m \rangle_n^2}{l(\theta_m, \theta)} \sim \chi^2(1) \ .$$

Since the distribution depends neither on $\mathbf{X}_m$ nor on $\boldsymbol{\epsilon}$, this random variable follows a χ^2 distribution with 1 degree of freedom. Besides, it is independent of the variable $\frac{\|\Pi_m . \boldsymbol{\epsilon}\|_n^2}{\sigma^2}$. Arguing as previously, we work out the distribution

$$\frac{n\|\Pi_m \boldsymbol{\epsilon}\|_n^2}{\sigma^2} \sim \chi^2(d_m) \ .$$

Consequently, the variable $E_m \mathbf{1}_{\Omega_2}$ is upper bounded by a random variable that follows the distribution of

$$\frac{8}{n} T_1 + \frac{1}{n} T_2 \ ,$$

where T_1 and T_2 are two independent χ^2 distribution with respectively 1 and d_m degrees of freedom. Moreover, the random variables $n\frac{\|\Pi_m(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_m)\|_n^2}{l(\theta_m,\theta)+\sigma^2}$ and $n\frac{\|\Pi_m^\perp(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_m)\|_n^2}{l(\theta_m,\theta)+\sigma^2}$ respectively follow a χ^2 distribution with d_m and $n - d_m$ degrees of freedom.

Let us bound the deviations of the random variables $E_m \mathbf{1}_{\Omega_2}$, $\frac{\|\Pi_m(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_m)\|_n^2}{l(\theta_m,\theta)+\sigma^2}$, and $\frac{\|\Pi_m^\perp(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_m)\|_n^2}{l(\theta_m,\theta)+\sigma^2}$ for any model $m \in \mathcal{M}$. We apply Lemma 1 in [22] for $E_m \mathbf{1}_{\Omega_2}$ and Lemma 7.2 for the two remaining random variables. Hence, for any $x > 0$, there exists an event $\mathbb{F}(x)$ of large probability

$$\begin{aligned} \mathbb{P}\left[\mathbb{F}(x)^c\right] &\leq e^{-x} \left(\sum_{m \in \mathcal{M}} e^{-\xi_1 d_m} + e^{-\xi_2 d_m} + e^{-\xi_3 d_m} \right) \\ &\leq e^{-x} \left[3 + \alpha \sum_{d=1}^{+\infty} d^\beta (e^{-\xi_1 d} + e^{-\xi_2 d} + e^{-\xi_3 d}) \right] \ , \end{aligned}$$

such that conditionally on $\mathbb{F}(x)$,

$$\left\{ \begin{array}{ccc} E_m \mathbf{1}_{\Omega_2} & \leq & \frac{d_m+8}{n} + \frac{2}{n}\sqrt{\left[d_m + 8^2\right](\xi_1 d_m + x)} + 16\frac{\xi_1 d_m + x}{n} \\ \frac{\|\Pi_m(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_m)\|_n^2}{l(\theta_m,\theta)+\sigma^2} & \leq & \frac{1}{n}\left(d_m + 2\sqrt{d_m\left[d_m \xi_2 + x\right]} + 2\left(d_m \xi_2 + x\right) \right) \\ -\frac{K d_m}{n-d_m} \frac{\|(\Pi_m^\perp \boldsymbol{\epsilon}+\boldsymbol{\epsilon}_m)\|_n^2}{\sigma^2 + l(\theta_m,\theta)} & \leq & -\frac{K d_m}{n(n-d_m)}\left(n - d_m - 2\sqrt{(n-d_m)(\xi_3 d_m + x)} \right) \end{array} \right. \ ,$$

for all models $m \in \mathcal{M}$. We shall fix later the positive constants ξ_1, ξ_2, and ξ_3. Let us apply extensively the inequality $2ab \leq \tau a^2 + \tau^{-1} b^2$. Hence, conditionally on $\mathbb{F}(x)$, the model $\widehat{m}$ satisfies

$$\left\{ \begin{array}{rcl} E_{\widehat{m}} \mathbf{1}_{\Omega_2} & \leq & \frac{d_{\widehat{m}}}{n} \left[1 + 2\sqrt{\xi_1} + 17\xi_1 + \tau_1 \right] + \frac{x}{n} \left[17 + \tau_1^{-1} \right] + \frac{72}{n} \\ \frac{\|\Pi_{\widehat{m}}(\boldsymbol{\epsilon} + \boldsymbol{\epsilon}_{\widehat{m}})\|_n^2}{l(\theta_{\widehat{m}}, \theta) + \sigma^2} & \leq & \frac{d_{\widehat{m}}}{n} \left[1 + 2\sqrt{\xi_2} + 2\xi_2 + \tau_2 \right] + \frac{x}{n} \left[2 + \tau_2^{-1} \right] \\ -\frac{K d_{\widehat{m}}}{n - d_{\widehat{m}}} \frac{\|\Pi_{\widehat{m}}^{\perp}(\boldsymbol{\epsilon} + \boldsymbol{\epsilon}_{\widehat{m}})\|_n^2}{\sigma^2 + l(\theta_{\widehat{m}}, \theta)} & \leq & -K \frac{d_{\widehat{m}}}{n} \left[1 - 2\sqrt{\xi_3 \frac{d_{\widehat{m}}}{n - d_{\widehat{m}}}} - \tau_3 \right] + K \frac{x}{n} \tau_3^{-1} \frac{d_{\widehat{m}}}{n - d_{\widehat{m}}} \ . \end{array} \right.$$

By Lemma 7.6, we know that conditionally on Ω_1, $\kappa_2 n \varphi_{\max} \left[(\mathbf{Z}_{\widehat{m}}^* \mathbf{Z}_{\widehat{m}})^{-1} \right]$ is smaller than $\frac{K-1}{4}$. By assumption $(\mathbb{H}_\eta)$, the ratio $\frac{d_{\widehat{m}}}{n - d_{\widehat{m}}}$ is smaller than $\frac{\eta}{1-\eta}$. Gathering these inequalities we upper bound $B_{\widehat{m}}$ on the event $\Omega_1 \cap \Omega_2 \cap \mathbb{F}(x)$,

$$B_{\widehat{m}} \leq \frac{d_{\widehat{m}}}{n} U + \frac{x}{n} V + \frac{72}{n} \ ,$$

where U and V are defined as

$$\begin{aligned} U &:= 1 + 2\sqrt{\xi_1} + 17\xi_1 + \tau_1 + \frac{K-1}{4} \left[1 + 2\sqrt{\xi_2} + 2\xi_2 + \tau_2 \right] - K \left[1 - 2\sqrt{\xi_3} \sqrt{\frac{\eta}{1-\eta}} - \tau_3 \right] \\ V &:= 17 + \tau_1^{-1} + \frac{K-1}{4} \left[2 + \tau_2^{-1} \right] + K \tau_3^{-1} \frac{\eta}{1-\eta} \ . \end{aligned}$$

Looking closely at U, one observes that it is the sum of the quantity $-\frac{3(K-1)}{4}$ and an expression that we can make arbitrary small by choosing the positive constants ξ_1, ξ_2, ξ_3, τ_1, τ_2, and τ_3 small enough. Consequently, there exists a suitable choice of these constants only depending on K and η that constrains the quantity U to be non positive. It follows that for any $x > 0$, with probability larger than $1 - e^{-x} L(K, \eta, \alpha, \beta)$,

$$B_{\widehat{m}} \mathbf{1}_{\Omega_1 \cap \Omega_2} \leq \frac{x}{n} L(K, \eta) + \frac{L'(K, \eta)}{n} \ .$$

Integrating this upper bound for any $x > 0$, we conclude

$$\mathbb{E} \left[B_{\widehat{m}} \mathbf{1}_{\Omega_1 \cap \Omega_2} \right] \leq \frac{L(K, \eta, \alpha, \beta)}{n} \ .$$

□

Proof of Lemma 7.9. We perform a very crude upper bound by controlling the sum of the risk of every estimator $\widehat{\theta}_m$.

$$\mathbb{E} \left[l(\widetilde{\theta}, \theta) 1_{\Omega_1^c \cup \Omega_2^c} \right] \leq \sqrt{\mathbb{P}(\Omega_1^c) + \mathbb{P}(\Omega_2^c)} \sqrt{\sum_{m \in \mathcal{M}} \mathbb{E} \left[l(\widehat{\theta}_m, \theta)^2 \right]} \ .$$

As for any model $m \in \mathcal{M}$, $l(\widehat{\theta}_m, \theta) = l(\theta_m, \theta) + l(\widehat{\theta}_m, \theta_m)$, it follows that

$$\mathbb{E} \left[l(\widehat{\theta}_m, \theta)^2 \right] \leq 2 \left\{ l(\theta_m, \theta)^2 + \mathbb{E} \left[l(\widehat{\theta}_m, \theta_m)^2 \right] \right\} \ .$$

For any model $m \in \mathcal{M}$, it holds that $n - d_m - 3 \geq (1 - \eta) n - 3$, which is positive by assumption $(\mathbb{H}_\eta)$. Hence, we may apply Proposition 7.8 with $r = 2$ to all models $m \in \mathcal{M}$:

$$\begin{aligned} \mathbb{E} \left[l(\widehat{\theta}_m, \theta_m)^2 \right] &\leq L \left[d_m n (\sigma^2 + l(\theta_m, \theta)) \right]^2 \\ &\leq L n^4 \mathrm{Var}(Y)^2 \ , \end{aligned}$$

since for any model m, $\sigma^2 + l(\theta_m, \theta) \leq \mathrm{Var}(Y)$. By summing this bound for all models $m \in \mathcal{M}$ and applying Lemma 7.6 and 7.7, we get

$$\mathbb{E}\left[l(\widetilde{\theta}, \theta)\mathbf{1}_{\Omega_1^c \cup \Omega_2^c}\right] \leq n^2 \mathrm{Card}(\mathcal{M}) L(K,\eta) \mathrm{Var}(Y) \exp\left[-nL'(K,\eta)\right] \ ,$$

where $L'(K,\eta)$ is positive.

□

7.3 Proof of Theorem 3.4 and Proposition 3.5

Proof of Theorem 3.4. This proof follows the same approach as the one of Theorem 3.1. We shall only emphasize the differences with this previous proof. The bound (29) still holds. Let us respectively define the three constants κ_1, κ_2 and $\nu(K)$ as

$$\begin{aligned}
\kappa_1 &:= \frac{\sqrt{\frac{3}{K+2}}}{1-\sqrt{\eta}-\nu(K)} \ , \qquad \kappa_2 := \frac{(K-1)\left[1-\sqrt{\eta}\right]^2\left[1-\sqrt{\eta}-\nu(K)\right]^2}{16} \wedge 1 \ , \\
\nu(K) &:= \left(\frac{3}{K+2}\right)^{1/6} \wedge \frac{1-\left(\frac{3}{K+2}\right)^{1/6}}{2} \ .
\end{aligned}$$

We also introduce the random variables $A_{m'}$ and $B_{m'}$ for any model $m' \in \mathcal{M}$.

$$\begin{aligned}
A_{m'} &:= \kappa_1 + 1 - \frac{\|\Pi^{\perp}_{m'}\boldsymbol{\epsilon}_{m'}\|_n^2}{l(\theta_{m'},\theta)} + \kappa_2 n \varphi_{\max}\left[(\mathbf{Z}^*_{m'}\mathbf{Z}_{m'})^{-1}\right]\frac{\|\Pi_{m'}(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_{m'})\|_n^2}{l(\theta_{m'},\theta)+\sigma^2} \\
&- K\left[1+\sqrt{2H(d'_m)}\right]^2 \frac{d_{m'}}{n-d_{m'}}\frac{\|\Pi_{m'^{\perp}}(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_{m'})\|_n^2}{l(\theta_{m'},\theta)+\sigma^2} \ , \\
B_{m'} &:= \kappa_1^{-1}\frac{\langle \Pi^{\perp}_{m'}\boldsymbol{\epsilon}, \Pi^{\perp}_{m'}\boldsymbol{\epsilon}_{m'}\rangle_n^2}{\sigma^2 l(\theta_{m'},\theta)} + \frac{\|\Pi_{m'}\boldsymbol{\epsilon}\|_n^2}{\sigma^2} + \kappa_2 n\varphi_{\max}\left[(\mathbf{Z}^*_{m'}\mathbf{Z}_{m'})^{-1}\right]\frac{\|\Pi_{m'}(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_{m'})\|_n^2}{l(\theta_{m'},\theta)+\sigma^2} \\
&- K\frac{d_{m'}}{n-d_{m'}}\left[1+\sqrt{2H(d'_m)}\right]^2\frac{\|\Pi^{\perp}_{m'}(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_{m'})\|_n^2}{l(\theta_{m'},\theta)+\sigma^2} \ .
\end{aligned}$$

The bound given in Lemma 7.5 clearly extends to

$$-\overline{\gamma}_n(\widetilde{\theta}) - \gamma_n(\widetilde{\theta})pen(\widehat{m}) - \sigma^2 + \|\boldsymbol{\epsilon}\|_n^2 \leq l(\widetilde{\theta},\theta)\left[A_{\widehat{m}} \vee (1-\kappa_2)\right] + \sigma^2 B_{\widehat{m}} \ .$$

As previously, we control the variable $A_{\widehat{m}}$ on an event of large probability Ω_1 and take the expectation of $B_{\widehat{m}}$ on an event of large probability $\Omega_1 \cap \Omega_2$.

Lemma 7.10. *Let Ω_1 be the event*

$$\Omega_1 := \{A_{\widehat{m}} \leq s(K,\eta)\} \bigcap \left\{\kappa_2 n \varphi_{max}\left[(\mathbf{Z}^*_{\widehat{m}}\mathbf{Z}_{\widehat{m}})^{-1}\right] \leq \frac{(K-1)\left(1-\sqrt{\eta}-\nu(K)\right)^2}{4}\right\} \ ,$$

where $s(K,\eta)$ is a function smaller than one. Then, $\mathbb{P}\left(\Omega_1^c\right) \leq L(K) n \exp\left[-nL'(K,\eta)\right]$ with $L'(K,\eta) > 0$.

The function $s(K,\eta)$ is given explicitly in the proof of Lemma 7.10

Lemma 7.11. *Let us assume that n is larger than some quantities $n_0(K)$. Then, there exists an event Ω_2 of probability larger than $1 - \exp\left[-nL(K,\eta)\right]$ where $L(K,\eta) > 0$ such that*

$$\mathbb{E}\left[B_{\widehat{m}}\mathbf{1}_{\Omega_1\cap\Omega_2}\right] \leq \frac{L(K,\eta)}{n} \ .$$

Gathering inequalities (29), (32), Lemma 7.10 and 7.11, we obtain as on the previous proof that

$$\begin{aligned}\mathbb{E}\left[l(\widetilde{\theta},\theta)\mathbf{1}_{\Omega_1\cap\Omega_2}\right] &\leq L(K,\eta)\inf_{m\in\mathcal{M}}\left[l(\theta_m,\theta)+\left(\sigma^2+l(\theta_m,\theta)\right)pen(m)\right]+\\ &+ L'(K,\eta)\left[\frac{\sigma^2}{n}+\left(\sigma^2+l(0_p,\theta)\right)n\exp\left[-nL''(K,\eta)\right]\right] . \qquad (42)\end{aligned}$$

Afterwards, we control the loss of the estimator $\widetilde{\theta}$ on the event of small probability $\Omega_1^c\cup\Omega_2^c$.

Lemma 7.12. *If n is larger than some quantity $n_0(K)$,*

$$\mathbb{E}\left[l(\widetilde{\theta},\theta)1_{\Omega_1^c\cup\Omega_2^c}\right] \leq n^{5/2}\left(\sigma^2+l(0_p,\theta)\right)L(K,\eta)\exp\left[-nL'(K,\eta)\right] ,$$

where $L(K,\eta)$ is positive.

Gathering this last bound with (42) enables to conclude. □

Proof of Lemma 7.10. This proof is analogous to the proof of Lemma 7.6, except that we shall change the weights in the concentration inequalities in order to take into account the complexity of the collection of models. Let x be a positive number we shall fix later. Applying Lemma 7.2, Lemma 7.3, and Lemma 7.4 ensures that there exists an event Ω_1' such that

$$P(\Omega_1^{'c})\leq 4\exp(-nx)\sum_{m\in\mathcal{M}}\exp\left[-d_mH(d_m)\right] ,$$

and for all models $m\in\mathcal{M}$,

$$\begin{aligned}\frac{\|\Pi_m^{\perp}\boldsymbol{\epsilon}_m\|_n^2}{l(\theta_m,\theta)} &\geq \frac{n-d_m}{n}\left[\left(1-\delta_{n-d_m}-\sqrt{\frac{2d_mH(d_m)}{n-d_m}}-\sqrt{\frac{2xn}{n-d_m}}\right)\vee 0\right]^2 , (43)\\ \frac{\|\Pi_m(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_m)\|_n^2}{\sigma^2+l(\theta_m,\theta)} &\leq \frac{2d_m}{n}\left[1+\sqrt{H(d_m)}+H(d_m)\right]+3x , \qquad (44)\\ \frac{\|\Pi_m^{\perp}(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_m)\|_n^2}{\sigma^2+l(\theta_m,\theta)} &\geq \frac{n-d_m}{n}\left[\left(1-\delta_{n-d_m}-\sqrt{\frac{2d_mH(d_m)}{n-d_m}}-\sqrt{\frac{2xn}{n-d_m}}\right)\vee 0\right]^2 , (45)\\ n\varphi_{\max}\left[\left(\mathbf{Z}_m^*\mathbf{Z}_m\right)^{-1}\right] &\leq \left[\left(1-\left(1+\sqrt{2H(d_m)}\right)\sqrt{\frac{d_m}{n}}-\sqrt{2x}\right)\vee 0\right]^{-2} .\end{aligned}$$

We recall that δ_d is defined in (28). Besides, it holds that

$$\mathbb{P}(\Omega_1^{'c})\leq 4\exp[-nx]\sum_{d=0}^{n}\mathrm{Card}\left[\{m\in\mathcal{M},\ d_m=d\}\right]\exp[-dH(d)]\leq 4n\exp[-nx] .$$

By Assumption $(\mathbb{H}_{K,\eta})$, the expression $\left(1+\sqrt{2H(d_m)}\right)\sqrt{\frac{d_m}{n}}$ is bounded by $\sqrt{\eta}$. Hence, conditionally on Ω_1',

$$n\varphi_{\max}\left[\left(\mathbf{Z}_{\widehat{m}}^*\mathbf{Z}_{\widehat{m}}\right)^{-1}\right] \leq \left[\left(1-\sqrt{\eta}-\sqrt{2x}\right)\vee 0\right]^{-2} ,$$

Constraining x to be smaller than $\frac{\left(1-\sqrt{\eta}\right)^2}{8}$ ensures that

$$n\kappa_2\varphi_{\max}\left[\left(\mathbf{Z}_{\widehat{m}}^*\mathbf{Z}_{\widehat{m}}\right)^{-1}\right]\mathbf{1}_{\Omega_1'} \leq \frac{(K-1)(1-\sqrt{\eta}-\nu(K))^2}{4} \ .$$

By assumption $(\mathbb{H}_{K,\eta})$, the dimension of any model $m \in \mathcal{M}$ is smaller than $n/2$. If n is larger than some quantities only depending on K, then $\delta_{n/2}$ is smaller than $\nu(K)$. Let us assume first that this is the case. We recall that $\nu(K)$ is defined at the beginning of the proof of Theorem 3.4. Since $\nu(K) \leq 1-\sqrt{\eta}$, inequality (43) becomes

$$\frac{\|\Pi_m^\perp \boldsymbol{\epsilon}_{\widehat{m}}\|_n^2}{l(\theta_{\widehat{m}},\theta)} \quad \geq \quad \left(1-\frac{d_{\widehat{m}}}{n}\right)\left[1-\nu(K)-\sqrt{\eta}\right]^2 - 2\sqrt{2x} \ .$$

Bounding analogously the remaining terms of $A_{\widehat{m}}$, we get

$$A_{\widehat{m}} \quad \leq \quad \kappa_1 + 1 - \left[1-\sqrt{\eta}-\delta_{n/2}\right]^2 + \frac{d_{\widehat{m}}}{n}(1-\sqrt{\eta}-\delta_{n/2})^2 U_1 + \sqrt{x}U_2 + xU_3 \ ,$$

where U_1, U_2, and U_3 are respectively defined as

$$\left\{\begin{array}{rcl} U_1 & := & -K\left[1+\sqrt{2H(d_{\widehat{m}})}\right]^2 + 1 + (K-1)/2\left[1+\sqrt{H(d_{\widehat{m}})}\right]^2 \leq 0 \\ U_2 & := & 2\sqrt{2}\left[1+K\eta\right] \\ U_3 & := & \frac{3}{4}(K-1)\left[1-\sqrt{\eta}-\nu(K)\right]^2 \ . \end{array}\right.$$

Since U_1 is non-positive, we obtain an upper bound of $A_{\widehat{m}}$ that does not depend anymore on $\widehat{m}$. By assumption $(\mathbb{H}_{K,\eta})$, we know that $\eta < (1-\nu(K)-(\frac{3}{K+2})^{1/6})^2$. Hence, coming back to the definition of κ_1 allows to prove that κ_1 is strictly smaller than $[1-\sqrt{\eta}-\nu(K)]^2$. Setting

$$x := \left[\frac{\left[1-\sqrt{\eta}-\nu(K)\right]^2-\kappa_1}{4U_2}\right]^2 \wedge \frac{\left[1-\sqrt{\eta}-\nu(K)\right]^2-\kappa_1}{4U_3} \wedge \frac{\left(1-\sqrt{\eta}\right)^2}{8} \ ,$$

we get

$$A_{\widehat{m}} \leq 1 - \frac{1}{2}\left[(1-\sqrt{\eta}-\nu(K))^2-\kappa_1\right] < 1 \ ,$$

on the event Ω_1'.

In order to take into account the case $\delta_{n/2} \geq \nu(K)$, we only have to choose a large constant $L(K)$ in the upper bound of $\mathbb{P}(\Omega_1^c)$. □

Proof of Lemma 7.11. Once again, the sketch of the proof closely follows the proof of Lemma 7.11. Let us consider the random variables E_m defined as

$$E_m := \kappa_1^{-1}\frac{\langle \Pi_{m'}^\perp \boldsymbol{\epsilon}, \Pi_{m'}^\perp \boldsymbol{\epsilon}_{m'}\rangle_n^2}{\sigma^2 l(\theta_{m'},\theta)} + \frac{\|\Pi_{m'}\boldsymbol{\epsilon}\|_n^2}{\sigma^2} \ .$$

Since $n\|\boldsymbol{\epsilon}\|_n^2/\sigma^2$ follows a χ^2 distribution with n degrees of freedom, there exists an event Ω_2 of probability larger than $1-\exp\left[-nL(K)\right]$such that $\|\boldsymbol{\epsilon}\|_n^2/\sigma^2$ is smaller than $\kappa_1^{-1} = \sqrt{(K+2)/3}[1-\sqrt{\eta}-\nu(K)]$ on Ω_2. The constant $L(K)$in the exponential is positive. We shall simultaneously upper bound the deviations of the random variables E_m, $\frac{\|\Pi_m(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_m)\|_n^2}{l(\theta_m,\theta)+\sigma^2}$, and $\frac{\|\Pi_m^\perp(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_m)\|_n^2}{\sigma^2+l(\theta_m,\theta)}$. Let ξ

be some positive constant that we shall fix later. For any $x > 0$, we define an event $\mathbb{F}(x)$ such that conditionally on $\mathbb{F}(x) \cap \Omega_2$,

$$\left\{ \begin{array}{rcl} E_m & \leq & \frac{d_m+\kappa_1^{-2}}{n} + \frac{2}{n}\sqrt{\left[d_m+\kappa_1^{-4}\right]\left[d_m(\xi+H(d_m))+x\right]} \\ & + & 2\kappa_1^{-2}\frac{\xi(d_m+H(d_m))+x}{n} \\ \frac{\|\Pi_m(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_m)\|_n^2}{l(\theta_m,\theta)+\sigma^2} & \leq & \frac{1}{n}\left[d_m + 2\sqrt{d_m\left[d_m(\frac{1}{16}+H(d_m))+x\right]} + 2\left[d_m(\frac{1}{16}+H(d_m))+x\right]\right] \\ \frac{\|\Pi_m^\perp \epsilon_m+\epsilon\|_n^2}{\sigma^2+l(\theta_m,\theta)} & \geq & \frac{n-d_m}{n}\left[\left(1-\delta_{n-d_m}-\sqrt{\frac{d_m(1+2H(d_m))}{n-d_m}}-\sqrt{\frac{2x}{n-d_m}}\right)\vee 0\right]^2 , \end{array} \right.$$

for any model $m \in \mathcal{M}$. Then, the probability of $\mathbb{F}(x)$ satisfies

$$\begin{aligned} \mathbb{P}\left[\mathbb{F}(x)^c\right] &\leq e^{-x}\left[\sum_{m\in\mathcal{M}} \exp\left[-d_m H(d_m)\right]\left(e^{-\xi d_m}+e^{-\frac{d_m}{16}}+e^{-\frac{d_m}{2}}\right)\right] \\ &\leq e^{-x}\left(\frac{1}{1-e^{-\xi}}+\frac{1}{1-e^{-1/16}}+\frac{1}{1-e^{-1/2}}\right) . \end{aligned}$$

Let us expand the three deviation bounds thanks to the inequality $2ab \leq \tau a^2 + \tau^{-1}b^2$:

$$\begin{aligned} E_m &\leq \frac{d_m}{n}\left[1+2\sqrt{\xi}+2\kappa_1^{-2}\xi+\tau_1\xi+\tau_2\right]+\frac{x}{n}\left[2\kappa_1^{-2}+\tau_2^{-1}+\tau_1\right] \\ &+ \frac{\kappa_1^{-2}}{n}\left[1+\tau_1^{-1}\kappa_1^{-2}\right]+\frac{d_m H(d_m)}{n}\left[2\kappa_1^{-2}+\tau_1\right]+2\frac{d_m\sqrt{H(d_m)}}{n} \\ &\leq \frac{d_m}{n}\left(1+\sqrt{2H(d_m)}\right)^2\left[\kappa_1^{-2}+2\sqrt{\xi}+2\kappa_1^{-2}\xi+\tau_1\xi+\tau_2\right] \\ &+ \frac{x}{n}\left[2\kappa_1^{-2}+\tau_2^{-1}+\tau_1\right]+\frac{\kappa_1^{-2}}{n}\left[1+\tau_1^{-1}\kappa_1^{-2}\right] . \end{aligned}$$

Similarly, we get

$$\frac{\|\Pi_m(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_m)\|_n^2}{l(\theta_m,\theta)+\sigma^2} \leq 2\frac{d_m}{n}\left[1+\sqrt{2H(d_m)}\right]^2+5\frac{x}{n} .$$

If n is larger than some quantity $n_0(K)$, then $\delta_{n/2}$ is smaller than $\nu(K)$. Applying Assumption $(\mathbb{H}_{K,\eta})$, we get

$$\begin{aligned} -K\frac{d_m}{n-d_m}&\left(1+\sqrt{2H(d_m)}\right)^2\frac{\|\Pi_m^\perp(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_m)\|_n^2}{l(\theta_m,\theta)+\sigma^2} \\ &\leq -K\frac{d_m}{n}\left(1+\sqrt{2H(d_m)}\right)^2\left[\left(1-\sqrt{\eta}-\nu(K)-\sqrt{\frac{2x}{n-d_m}}\right)\vee 0\right]^2 \\ &\leq -K\frac{d_m}{n}\left(1+\sqrt{2H(d_m)}\right)^2\left[\left(1-\sqrt{\eta}-\nu(K)\right)^2-\tau_3\right]+2K\eta\tau_3^{-1}\frac{x}{n} . \end{aligned}$$

Let us combine these three bounds with the definitions of B_m, κ_1, and κ_2. Hence, Conditionally to the event $\Omega_1 \cap \Omega_2 \cap \mathbb{F}(x)$,

$$B_{\widehat{m}} \leq \frac{d_{\widehat{m}}}{n}\left[1+\sqrt{2H(\widehat{m})}\right]^2 U_1+\frac{x}{n}U_2+\frac{L(K,\eta)}{n}U_3 , \tag{46}$$

where

$$\left\{ \begin{array}{rcl} U_1 & := & -\frac{K-1}{6}\left(1-\sqrt{\eta}-\nu(K)\right)^2+K\tau_3+2\sqrt{\xi}+2\kappa_1^{-2}\xi+\tau_1\xi+\tau_2 , \\ U_2 & := & \tau_2^{-1}+\tau_1+L(K,\eta)(1+\tau_3^{-1}) , \\ U_3 & := & 1+\tau_1^{-1} . \end{array} \right.$$

Since $K > 1$, there exists a suitable choice of the constants ξ, τ_1, and τ_2, only depending on K and η that constrains U_1 to be non positive. Hence, conditionally on the event $\Omega_1 \cap \Omega_2 \cap \mathbb{F}(x)$,

$$B_{\widehat{m}} \leq \frac{L(K,\eta)}{n} + L'(K,\eta)\frac{x}{n} \ .$$

Since $\mathbb{P}\left[\mathbb{F}(x)^c\right] \leq e^{-x}L(K,\eta)$, we conclude by integrating the last expression with respect to x. □

Proof of Lemma 7.12. As in the ordered selection case, we apply Cauchy-Schwarz inequality

$$\mathbb{E}\left[l(\widetilde{\theta},\theta)1_{\Omega_1^c\cup\Omega_2^c}\right] \quad \leq \quad \sqrt{\mathbb{P}(\Omega_1^c)+\mathbb{P}(\Omega_2^c)}\sqrt{\mathbb{E}\left[l(\widetilde{\theta},\theta)^2\right]} \ .$$

However, there are too many models to bound efficiently the risk of $\widetilde{\theta}$ by the sum of the risks of the estimators $\widehat{\theta}_m$. This is why we use here Hölder's inequality

$$\begin{aligned}\mathbb{E}\left[l(\widetilde{\theta},\theta)1_{\Omega_1^c\cup\Omega_2^c}\right] &\leq L(K)\sqrt{n}\exp\left[-nL(K,\eta)\right]\sqrt{\mathbb{E}\left[\sum_{m\in\mathcal{M}}\mathbf{1}_{m=\widehat{m}}l(\widehat{\theta}_m,\theta)^2\right]} \\ &\leq L(K)\sqrt{n}\exp\left[-nL(K,\eta)\right]\sqrt{\sum_{m\in\mathcal{M}}\mathbb{P}\left(m=\widehat{m}\right)^{1/u}\mathbb{E}\left[l(\widehat{\theta}_m,\theta)^{2v}\right]^{1/v}}, \quad (47)\end{aligned}$$

where $v := \left\lfloor \frac{n}{8} \right\rfloor$, and $u =: \frac{v}{v-1}$. We assume here that n is larger than 8. For any model $m \in \mathcal{M}$, the loss $l(\widehat{\theta}_m,\theta)$ decomposes into the sum $l(\theta_m,\theta) + l(\widehat{\theta}_m,\theta_m)$. Hence,we obtain the following upper bound by applying Minkowski's inequality

$$\mathbb{E}\left[l(\widehat{\theta}_m,\theta)^{2v}\right]^{1/2v} \leq l(\theta_m,\theta) + \mathbb{E}\left[l(\widehat{\theta}_m,\theta_m)^{2v}\right]^{1/2v} \leq \mathrm{Var}(Y) + \mathbb{E}\left[l(\widehat{\theta}_m,\theta_m)^{2v}\right]^{1/2v} \ . \quad (48)$$

We shall upper bound this last term thanks to Proposition 7.8. Since v is smaller than $n/8$ and since d_m is smaller than $n/2$, it follows that for any model $m \in \mathcal{M}$, $n - d_m - 4v + 1$ is positive and

$$\mathbb{E}\left[l(\widehat{\theta}_m,\theta_m)^{2v}\right]^{1/2v} \leq 2vLnd_m\left(\sigma^2 + l(\theta_m,\theta)\right) \ ,$$

for any model $m \in \mathcal{M}$. Since $d_m \leq n$ and since $\sigma^2 + l(\theta_m,\theta) \leq \mathrm{Var}(Y)$, we obtain

$$\mathbb{E}\left[l(\widehat{\theta}_m,\theta_m)^{2v}\right]^{1/2v} \leq 2vLn^2\mathrm{Var}(Y) \ . \quad (49)$$

Gathering upper bounds (47), (48), and (49) we get

$$\begin{aligned}\mathbb{E}\left[l(\widetilde{\theta},\theta)1_{\Omega_1^c\cup\Omega_2^c}\right] &\leq L(K)\sqrt{n}\exp\left[-nL'(K,\eta)\right] \\ &\times \left[\mathrm{Var}(Y) + 2vLn^2\mathrm{Var}(Y)\right]\sqrt{\sum_{m\in\mathcal{M}}\mathbb{P}\left(m=\widehat{m}\right)^{1/u}} \ .\end{aligned}$$

Since the sum over $m \in \mathcal{M}$ of $\mathbb{P}\left(m=\widehat{m}\right)$ is one, the last term of the previous expression is maximized when every $\mathbb{P}\left(m=\widehat{m}\right)$ equals $\frac{1}{\mathrm{Card}(\mathcal{M})}$. Hence,

$$\mathbb{E}\left[l(\widetilde{\theta},\theta)1_{\Omega_1^c\cup\Omega_2^c}\right] \quad \leq \quad n^{5/2}\mathrm{Var}(Y)L(K,\eta)\mathrm{Card}(\mathcal{M})^{1/(2v)}\exp\left[-nL'(K,\eta)\right] \ ,$$

where $L'(K,\eta)$ is positive. Let us first bound the cardinality of the collection $\mathcal{M}$. We recall that the dimension of any model $m \in \mathcal{M}$ is assumed to be smaller than $n/2$ by $(\mathbb{H}_{K,\eta})$. Besides, for any $d \in \{1,\dots,n/2\}$, there are less than $\exp(dH(d))$ models of dimension d. Hence,

$$\log\left(\mathrm{Card}(\mathcal{M})\right) \leq \log(n) + \sup_{d=1,\dots,n/2} dH(d) \ .$$

By assumption $(\mathbb{H}_{K,\eta})$, $dH(d)$ is smaller than $n/2$. Thus, $\log(\mathrm{Card}(\mathcal{M})) \leq \log(n) + n/2$ and it follows that $\mathrm{Card}(\mathcal{M})^{1/(2v)}$ is smaller than an universal constant providing that n is larger than 8. All in all, we get

$$\mathbb{E}\left[l(\widetilde{\theta},\theta)1_{\Omega_1^c \cup \Omega_2^c}\right] \quad \leq \quad n^{5/2}\mathrm{Var}(Y)L(K,\eta)\exp\left[-nL'(K,\eta)\right] \ ,$$

where $L'(K,\eta)$ is positive. □

Proof of Proposition 3.5. We apply the same arguments as in the proof of Theorem 3.4, except that we replace $H(d_m)$ by l_m.

$$\begin{aligned}
A_{m'} &:= \kappa_1 + 1 - \frac{\|\Pi_{m'}^{\perp}\boldsymbol{\epsilon}_{m'}\|_n^2}{l(\theta_{m'},\theta)} + \kappa_2 n\varphi_{\max}\left[(\mathbf{Z}_{m'}^*\mathbf{Z}_{m'})^{-1}\right]\frac{\|\Pi_{m'}(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_{m'})\|_n^2}{l(\theta_{m'},\theta)+\sigma^2} \\
&- K\left[1+\sqrt{2l_{m'}}\right]^2\frac{d_{m'}}{n-d_{m'}}\frac{\|\Pi_{m'^{\perp}}(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_{m'})\|_n^2}{l(\theta_{m'},\theta)+\sigma^2} \ , \\
B_{m'} &:= \kappa_1^{-1}\frac{\langle\Pi_{m'}^{\perp}\boldsymbol{\epsilon},\Pi_{m'}^{\perp}\boldsymbol{\epsilon}_{m'}\rangle_n^2}{\sigma^2 l(\theta_{m'},\theta)} + \frac{\|\Pi_{m'}\boldsymbol{\epsilon}\|_n^2}{\sigma^2} + \kappa_2 n\varphi_{\max}\left[(\mathbf{Z}_{m'}^*\mathbf{Z}_{m'})^{-1}\right]\frac{\|\Pi_{m'}(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_{m'})\|_n^2}{l(\theta_{m'},\theta)+\sigma^2} \\
&- K\frac{d_{m'}}{n-d_{m'}}\left[1+\sqrt{2l'_m}\right]^2\frac{\|\Pi_{m'}^{\perp}(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_{m'})\|_n^2}{l(\theta_{m'},\theta)+\sigma^2} \ .
\end{aligned}$$

In fact, Lemma 7.10, 7.11, and 7.12 are still valid for this penalty. The previous proofs of these three lemma depend on the quantity $H(d_m)$ through the properties:

$$H(d_m) \text{ satisfies assumption } (\mathbb{H}_{K,\eta}) \text{ and } \textstyle\sum_{m\in\mathcal{M},\ d_m=d}\exp(-dH(d_m)) \leq 1.$$

Under the assumptions of Proposition 3.5, l_m satisfies the corresponding Assumption $(\mathbb{H}^l_{K,\eta})$ and is such that $\sum_{m\in\mathcal{M},\ d_m=d}\exp(-dl_m)) \leq 1$. Hence, the proofs of these lemma remain valid in this setting if we replace $H(d_m)$ by l_m.

There is only one small difference at the end of the proof of Lemma 7.12 when bounding $\log\left(\mathrm{Card}(\mathcal{M})\right)$. By definition of l_m,

$$\mathrm{Card}(\mathcal{M}) - 1 \leq \sup_{m\in\mathcal{M}\setminus\{\emptyset\}} \exp(d_m l_m) \ .$$

Hence, $\log(\mathrm{Card}(\mathcal{M}) \leq 1 + \sup_{m\in\mathcal{M}\setminus\{\emptyset\}} d_m l_m$, which is smaller than $1 + n/2$ by Assumption $(\mathbb{H}^l_{K,\eta})$. Hence, the upper bound shown in the proof of Lemma 7.12 is still valid.

□

7.4 Proof of Proposition 7.8

Proof of Proposition 7.8. Let m be a subset of $\{1,\dots,p\}$. Thanks to (27), we know that

$$l(\widehat{\theta}_m,\theta_m) = (\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_m)^* \mathbf{Z}_m \left(\mathbf{Z}_m^*\mathbf{Z}_m\right)^{-2}\mathbf{Z}_m^*\left(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_m\right) \ .$$

Applying Cauchy-Schwarz inequality, we decompose the r-th loss of $\widehat{\theta}_m$ in two terms

$$\begin{aligned}\mathbb{E}\left[l(\widehat{\theta}_m,\theta_m)^r\right]^{\frac{1}{r}} &\leq \mathbb{E}\left[\left\|(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_m)(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_m)^*\right\|_F^r \left\|\mathbf{Z}_m\left(\mathbf{Z}_m^*\mathbf{Z}_m\right)^{-2}\mathbf{Z}_m^*\right\|_F^r\right]^{\frac{1}{r}} \\ &\leq \mathbb{E}\left[\left\|(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_m)(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_m)^*\right\|_F^r\right]^{\frac{1}{r}} \mathbb{E}\left\{tr\left[\left(\mathbf{Z}_m^*\mathbf{Z}_m\right)^{-2}\right]^{\frac{r}{2}}\right\}^{\frac{1}{r}}, \qquad (50)\end{aligned}$$

by independence of $\boldsymbol{\epsilon}$, $\boldsymbol{\epsilon}_m$, and $\mathbf{Z}_m$. Here, $\|.\|_F$ stands for the Frobenius norm in the space of square matrices. We shall successively upper bound the two terms involved in (50).

$$\left\|(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_m)(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_m)^*\right\|_F^r = \left[\sum_{1\leq i,j\leq n}(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_m)[i]^2(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_m)[j]^2\right]^{r/2} .$$

This last expression corresponds to the $L_{r/2}$ norm of a Gaussian chaos of order 4. By Theorem 3.2.10 in [18], such chaos satisfy a Khintchine-Kahane type inequality:

Lemma 7.13. *For all $d\in\mathbb{N}$ there exists a constant $L_d\in(0,\infty)$ such that, if X is a Gaussian chaos of order d with values in any normed space F with norm $\|.\|$ and if $1<s<q<\infty$, then*

$$\left(\mathbb{E}\|X\|^q\right)^{\frac{1}{q}} \leq L_d\left(\frac{q-1}{s-1}\right)^{d/2}\mathbb{E}\left[\|X\|^s\right]^{\frac{1}{s}} .$$

Let us assume that r is larger than four. Applying the last lemma with $d=4$, $q=r/2$, and $s=2$ yields

$$\mathbb{E}\left[\left\|(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_m)(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_m)^*\right\|_F^r\right]^{\frac{2}{r}} \leq L_4(r/2-1)^2\mathbb{E}\left[\left\|(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_m)(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_m)^*\right\|_F^4\right]^{\frac{1}{2}} .$$

By standard Gaussian properties, we compute the fourth moment of this chaos and obtain

$$\mathbb{E}\left[\left\|(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_m)(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_m)^*\right\|_F^4\right]^{\frac{1}{2}} \leq Ln^2\left[\sigma^2+l(\theta_m,\theta)\right]^2 .$$

Hence, we get the upper bound

$$\mathbb{E}\left[\left\|(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_m)(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_m)^*\right\|_F^r\right]^{\frac{1}{r}} \leq L(r-1)n\left[\sigma^2+l(\theta_m,\theta)\right] . \qquad (51)$$

Straightforward computations allow to extend this bound to $r=2$ and $r=3$.

Let us turn to bounding the second term of (50). Since the eigenvalues of the matrix $\left(\mathbf{Z}_m^*\mathbf{Z}_m\right)^{-1}$ are almost surely non-negative, it follows that

$$tr\left[\left(\mathbf{Z}_m^*\mathbf{Z}_m\right)^{-2}\right] \leq tr\left[\left(\mathbf{Z}_m^*\mathbf{Z}_m\right)^{-1}\right]^2 .$$

Consequently, we shall upper bound the r-th moment of the trace of an inverse standard Wishart matrix. For any couple of matrices A and B respectively of size $p_1\times q_1$ and $p_2\times q_2$, we define the Kronecker product matrix $A\otimes B$ as the matrix of size $p_1p_2\times q_1q_2$ that satisfies:

$$A\otimes B[i_2+p_2(i_1-1);j_2+q_2(j_1-1)] := A[i_1;j_1]B[i_2;j_2] , \quad \text{for any } \begin{cases} 1\leq i_1\leq p_1 \\ 1\leq i_2\leq p_2 \\ 1\leq j_1\leq q_1 \\ 1\leq j_2\leq q_2 \end{cases} .$$

For any matrix A, $\otimes^k A$ refers to the k-th power of A with respect to the Kronecker product. Since $tr(A)^k = tr\left(\otimes^k A\right)$ for any square matrix A, we obtain

$$\begin{aligned}\mathbb{E}\left[tr(\mathbf{Z}_m^*\mathbf{Z}_m)^{-1}\right]^k &= \mathbb{E}\left[tr\left(\otimes^k(\mathbf{Z}_m^*\mathbf{Z}_m)^{-1}\right)\right] \\ &= tr\left[\mathbb{E}\left(\otimes^k(\mathbf{Z}_m^*\mathbf{Z}_m)^{-1}\right)\right] \\ &\leq \sqrt{d_m^k}\left\|\mathbb{E}\left[\otimes^k(\mathbf{Z}_m^*\mathbf{Z}_m)^{-1}\right]\right\|_F ,\end{aligned}$$

thanks to Cauchy-Schwarz inequality. In Equation (4.2) of [37], Von Rosen has characterized recursively the expectation of $\otimes^k(\mathbf{Z}_m^*\mathbf{Z}_m)^{-1}$ as long as $n - d_m - 2k - 1$ is positive:

$$\mathrm{vec}\left(\mathbb{E}\left[\otimes^{k+1}(\mathbf{Z}_m^*\mathbf{Z}_m)^{-1}\right]\right) = A(n,d_m,k)^{-1}\mathrm{vec}\left(\mathbb{E}\left[\otimes^k(\mathbf{Z}_m'\mathbf{Z}_m)^{-1}\right]\otimes I\right) , \tag{52}$$

where 'vec' refers to the vectorized version of the matrix. See Section 2 of [37] for more details about this definition. $A(n,d_m,k)$ is a symmetric matrix of size $d_m^{k+1}\times d_m^{k+1}$ which only depends on n, d_m, and k and is known to be diagonally dominant. More precisely, any diagonal element of $A(n,d_m,k)$ is greater or equal to one plus the corresponding row sums of the absolute values of the off-diagonal elements. Hence, the matrix A is invertible and its smallest eigenvalue is larger or equal to one. Consequently, $\varphi_{\max}\left(A^{-1}\right)$ is smaller or equal to one. It then follows from (52) that

$$\begin{aligned}\left\|\mathbb{E}\left[\otimes^{k+1}(\mathbf{Z}_m^*\mathbf{Z}_m)^{-1}\right]\right\|_F &= \left\|\mathrm{vec}\left(\mathbb{E}\left[\otimes^{k+1}(\mathbf{Z}_m^*\mathbf{Z}_m)^{-1}\right]\right)\right\|_F \\ &\leq \varphi_{\max}(A^{-1})\left\|\mathrm{vec}\left(\mathbb{E}\left[\otimes^k(\mathbf{Z}_m^*\mathbf{Z}_m)^{-1}\right]\otimes I\right)\right\|_F \\ &\leq \sqrt{d_m}\left\|\mathbb{E}\left[\otimes^k(\mathbf{Z}_m^*\mathbf{Z}_m)^{-1}\right]\right\|_F .\end{aligned}$$

By induction, we obtain

$$\mathbb{E}\left[tr(\mathbf{Z}_m^*\mathbf{Z}_m)^{-1}\right]^r \leq d_m^r , \tag{53}$$

if $n - d_m - 2r + 1 > 0$. Combining upper bounds (51) and (53) enables to conclude

$$\mathbb{E}\left[l(\widehat{\theta}_m,\theta_m)^r\right]^{\frac{1}{r}} \leq Lrd_m n(\sigma^2 + l(\theta_m,\theta)) .$$

□

7.5 Proof of Proposition 3.2

Proof of Proposition 3.2. Let m_* be the model that minimizes the loss function $l(\widehat{\theta}_m,\theta)$:

$$m_* = \arg\inf_{m\in\mathcal{M}_{\lfloor n/2\rfloor}} l(\widehat{\theta}_m,\theta) .$$

It is almost surely uniquely defined. Contrary to the oracle m^*, the model m_* is random. By definition of $\widehat{m}$, we derive that

$$l(\widetilde{\theta},\theta) \leq l(\widehat{\theta}_{m_*},\theta) + \gamma_n(\widehat{\theta}_{m^*})pen(m_*) + \overline{\gamma}_n(\widehat{\theta}_{m_*}) - \gamma_n(\widetilde{\theta})pen(\widehat{m}) - \overline{\gamma}_n(\widetilde{\theta}) , \tag{54}$$

where $\overline{\gamma}_n$ is defined in the proof of Theorem 3.1. The proof divides in two parts. First, we state that on an event Ω_1 of large probability, the dimensions of $\widehat{m}$ and of m^* are moderate. Afterwards, we prove that on another event of large probability $\Omega_1\cap\Omega_2\cap\Omega_3$, the ratio $l(\widetilde{\theta},\theta)/l(\widehat{\theta}_{m^*},\theta)$ is close to one.

Lemma 7.14. *Let us define the event Ω_1 as:*

$$\Omega_1 := \left\{ \log^2(n) < d_{m_*} < \frac{n}{\log n} \quad \textit{and} \quad \log^2(n) < d_{\widehat{m}} < \frac{n}{\log n} \right\} .$$

The event Ω_1 is achieved with large probability: $\mathbb{P}(\Omega_1) \geq 1 - \frac{L(R,s)}{n^2}$.

Lemma 7.15. *There exists an event Ω_2 of probability larger than $1 - L\frac{\log n}{n}$ such that*

$$\left[-\overline{\gamma}_n(\widetilde{\theta}) - \gamma_n(\widetilde{\theta}) pen(\widehat{m}) - \sigma^2 + \|\boldsymbol{\epsilon}\|_n^2 \right] \mathbf{1}_{\Omega_1 \cap \Omega_2} \leq l(\widetilde{\theta}, \theta) \tau_1(n),$$

where $\tau_1(n)$ is a positive sequence converging to zero when n goes to infinity.

Lemma 7.16. *There exists an event Ω_3 of probability larger than $1 - L\frac{\log n}{n}$ such that*

$$\left[\overline{\gamma}_n(\widehat{\theta}_{m^*}) + \gamma_n(\widehat{\theta}_{m^*}) pen(m^*) + \sigma^2 - \|\boldsymbol{\epsilon}\|_n^2 \right] \mathbf{1}_{\Omega_1 \cap \Omega_3} \leq l\left(\widehat{\theta}_{m_*}, \theta\right) \tau_2(n),$$

where $\tau_2(n)$ is a positive sequence converging to zero when n goes to infinity.

Gathering these three lemma, we derive from the upper bound (54) the inequality

$$\frac{l(\widetilde{\theta}, \theta)}{l(\widehat{\theta}_{m_*}, \theta)} \mathbf{1}_{\Omega_1 \cap \Omega_2 \cap \Omega_3} \leq \frac{1 + \tau_2(n)}{1 - \tau_1(n)} ,$$

which allows to conclude.

□

Proof of Lemma 7.14. Let us consider the model $m_{R,s}$ defined by $d_{m_{R,s}} := \lfloor (nR^2)^{\frac{1}{1+s}} \rfloor$. If n is larger than some quantity $L(R,s)$, then $d_{m_{R,s}}$ is smaller than $n/2$ and $m_{R,s}$ therefore belongs to the collection $\mathcal{M}_{\lfloor n/2 \rfloor}$. We shall prove that outside an event of small probability, the loss $l(\widehat{\theta}_{m_{R,s}}, \theta)$ is smaller than the loss $l(\widehat{\theta}_m, \theta)$ of all models $m \in \mathcal{M}_{\lfloor n/2 \rfloor}$ whose dimension is smaller than $\log^2(n)$ or larger than $\frac{n}{\log n}$. Hence, the model m_* satisfies $\log^2(n) < d_{m_*} < \frac{n}{\log n}$ with large probability.

First, we need to upper bound the loss $l(\widehat{\theta}_{m_{R,s}}, \theta)$. Since $l(\widehat{\theta}_{m_{R,s}}, \theta) = l(\theta_{m_{R,s}}, \theta) + l(\widehat{\theta}_{m_{R,s}}, \theta_{m_{R,s}})$, it comes to upper bounding both the bias term and the variance term. Since θ belongs to $\mathcal{E}'_s(R)$,

$$\begin{aligned} l\left(\theta_{m_{R,s}}, \theta\right) &= \sum_{i > d_{m_{R,s}}}^{+\infty} l(\theta_{m_{i-1}}, \theta_{m_i}) \\ &\leq (d_{m_i} + 1)^{-s} \sum_{i > d_{m_{R,s}}}^{+\infty} \frac{l(\theta_{m_{i-1}}, \theta_{m_i})}{i^{-s}} \leq \sigma^2 \left(\frac{R^2}{n^s} \right)^{\frac{1}{1+s}} . \end{aligned} \tag{55}$$

Then, we bound the variance term $l(\widehat{\theta}_{m_{R,s}}, \theta_{m_{R,s}})$ thanks to (36) as in the proof of Lemma 7.5.

$$l\left(\widehat{\theta}_{m_{R,s}}, \theta_{m_{R,s}}\right) \leq \left[\sigma^2 + l\left(\theta_{m_{R,s}}, \theta\right)\right] \varphi_{\max}\left[n(\mathbf{Z}^*_{m_{R,s}} \mathbf{Z}_{m_{R,s}})^{-1} \right] \frac{\left\| \Pi_{m_{R,s}}(\boldsymbol{\epsilon} + \boldsymbol{\epsilon}_{m_{R,s}}) \right\|_n^2}{\sigma^2 + l(\theta_{m_{R,s}}, \theta)} .$$

The two random variables involved in this last expression respectively follow (up to a factor n) the distribution of an inverse Wishart matrix with parameters $(n, d_{m_{R,s}})$ and a χ^2 distribution

with $d_{m_{R,s}}$ degrees of freedom. Thanks to Lemma 7.2 and 7.4, we prove that outside an event of probability smaller than $L(R,s)\exp[-L'(R,s)n^{\frac{1}{1+s}}]$ with $L'(R,s)>0$,

$$l\left(\widehat{\theta}_{m_{R,s}},\theta_{m_{R,s}}\right) \leq 4\left[\sigma^2+l\left(\theta_{m_{R,s}},\theta\right)\right]\frac{d_{m_{R,s}}}{n}\ ,$$

if n is large enough. Gathering this last upper bound with (55) yields

$$l\left(\widehat{\theta}_{m_{R,s}},\theta\right)\leq\sigma^2\left[5\frac{R^{\frac{2}{1+s}}}{n^{\frac{s}{1+s}}}+4\left(\frac{R^{\frac{2}{1+s}}}{n^{\frac{s}{1+s}}}\right)^2\right]\leq\sigma^2\frac{C(R,s)}{n^{\frac{s}{1+s}}} \tag{56}$$

where $C(R,s)$ is a constant that only depends on R and s.

Let us prove that the bias term of any model of dimension smaller than $\log^2(n)$ is larger than (56) if n is large enough. Obviously, we only have to consider the model of dimension $\lfloor\log^2(n)\rfloor$. Assume that there exists an infinite increasing sequence of integers u_n satisfying:

$$\sum_{i>\log^2(u_n)} l\left(\theta_{m_{i-1}},\theta_{m_i}\right)\leq\frac{C(R,s)}{(u_{n+1})^{\frac{s}{1+s}}}\ . \tag{57}$$

Then, the sequence (v_n) defined by $v_n:=\log^2(u_n)$ satisfies

$$\sum_{i>v_n} l\left(\theta_{m_{i-1}},\theta_{m_i}\right)\leq C(R,s)\exp\left[-\sqrt{v_{n+1}}\frac{s}{1+s}\right]\ .$$

Let us consider a subsequence of (v_n) such that $\lfloor v_n\rfloor$ is strictly increasing. For the sake of simplicity we still call it v_n. It follows that

$$\begin{aligned}\sum_{i=\lfloor v_0\rfloor+1}^{+\infty}\frac{l\left(\theta_{m_{i-1}},\theta_{m_i}\right)}{i^{-s'}} &= \sum_{n=0}^{+\infty}\sum_{i=\lfloor v_n\rfloor+1}^{\lfloor v_{n+1}\rfloor}\frac{l\left(\theta_{m_{i-1}},\theta_{m_i}\right)}{i^{-s'}}\\ &\leq C(R,s)\sum_{n=0}^{+\infty}\lfloor v_{n+1}\rfloor^{s'}\exp\left[-\sqrt{\lfloor v_{n+1}\rfloor}\frac{s}{1+s}\right]<\infty\ ,\end{aligned}$$

and θ therefore belongs to some ellipsoid $\mathcal{E}_{s'}(R')$. This contradicts the assumption θ does not belong to any ellipsoid $\mathcal{E}_{s'}(R')$. As a consequence, there only exists a finite sequence of integers u_n that satisfy Condition (57). For n large enough, the bias term of any model of dimension less than $\log^2(n)$ is therefore larger than the loss $l(\widehat{\theta}_{m_{R,s}},\theta)$ with overwhelming probability.

Let us turn to the models of dimension larger than $n/\log n$. We shall prove that with large probability, for any model m of dimension larger than $n/\log n$, the variance term $l(\widehat{\theta}_m,\theta_m)$ is larger than the order $\sigma^2/\log n$. For any model $m\in\mathcal{M}_{\lfloor n/2\rfloor}$,

$$l\left(\widehat{\theta}_m,\theta_m\right)\geq\frac{n\sigma^2}{\varphi_{\max}\left(\mathbf{Z}_m^*\mathbf{Z}_m\right)}\frac{\|\Pi_m(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_m)\|_n^2}{\sigma^2+l(\theta_m,\theta)}\ .$$

The two random variables involved in this expression respectively follow (up to a factor n) a Wishart distribution with parameters (n,d_m) and a χ^2 distribution with d_m. Again, we apply Lemma 7.2 and 7.4 to control the deviations of these random variables. Hence, outside an event of probability smaller than $L(\xi)\exp[-n\xi/\log n]$,

$$l\left(\widehat{\theta}_m,\theta_m\right)\geq\sigma^2\left(1+\sqrt{\frac{d_m}{n}}+\sqrt{2\xi\frac{d_m}{n}}\right)^{-2}\frac{d_m}{n}\left(1-2\sqrt{\xi}\right)\ ,$$

for any model m of dimension larger than $n/\log n$. For any model $m \in \mathcal{M}_{\lfloor n/2 \rfloor}$, the ratio d_m/n is smaller than $1/2$. As a consequence, we get

$$l\left(\widehat{\theta}_m, \theta_m\right) \geq \frac{\sigma^2}{\log n}\left(1 - 2\sqrt{\xi}\right)\left(1 + \sqrt{1/2} + \sqrt{\xi}\right)^{-2} .$$

Choosing for instance $\xi = 1/16$ ensures that for n large enough the loss $l(\widehat{\theta}_m, \theta_m)$ is larger than $l(\widehat{\theta}_{m_{R,s}}, \theta)$ for every model m of dimension larger than $n/\log n$ outside an event of probability smaller than $L_1 \exp[-L_2 n/\log n] + L_3(R,s)\exp[-L_4(R,s)n^{1/(1+s)}]$ with $L_4(R,s) > 0$.

Let us now turn to the selected model $\widehat{m}$. We shall prove that outside an event of small probability,

$$\gamma_n\left(\widehat{\theta}_{m_{R,s}}\right)[1 + pen(m_{R,s})] \leq \gamma_n\left(\widehat{\theta}_m\right)[1 + pen(m)] , \tag{58}$$

for all models m of dimension smaller than $\log^2 n$ or larger than $n/\log n$. We first consider the models of dimension smaller than $\log^2(n)$. For any model $m \in \mathcal{M}_{\lfloor n/2 \rfloor}$, $\gamma_n(\widehat{\theta}_m) * n/[\sigma^2 + l(\theta_m, \theta)]$ follows a χ^2 distribution with $n - d_m$ degrees of freedom. Again, we apply Lemma 7.2. Hence, with probability larger than $1 - e/[n^2(e-1)]$, the following upper bound holds for any model m of dimension smaller than $\log^2(n)$.

$$\begin{aligned}
\gamma_n\left(\widehat{\theta}_m\right)[1 + pen(m)] &\geq \sigma^2\left[1 + \frac{l(\theta_m,\theta)}{\sigma^2}\right]\left(1 + 2\frac{d_m}{n - d_m}\right)\left[\frac{n - d_m}{n} - 2\frac{\sqrt{(n-d_m)(d_m + 2\log(n))}}{n}\right] \\
&\geq \sigma^2\left[1 + \frac{l(\theta_m,\theta)}{\sigma^2}\right]\left(1 + \frac{d_m}{n}\right)\left[1 - 2\sqrt{\frac{d_m + 2\log(n)}{n - d_m}}\right] \\
&\geq \sigma^2\left[1 + \frac{l(\theta_m,\theta)}{\sigma^2}\right]\left[1 - 4\frac{\log n}{\sqrt{n}}\right] ,
\end{aligned}$$

for n large enough. Besides, outside an event of probability smaller than $\frac{1}{n^2}$,

$$\begin{aligned}
\gamma_n\left(\widehat{\theta}_{m_{R,s}}\right)[1 + pen(m_{R,s})] &\leq \sigma^2\left[1 + \frac{l(\theta_{m_{R,s}},\theta)}{\sigma^2}\right]\left(1 + 2\frac{d_{m_{R,s}}}{n - d_{m_{R,s}}}\right) \times \\
&\qquad \left[\frac{n - d_{m_{R,s}}}{n} + 2\frac{\sqrt{(n - d_{m_{R,s}})2\log n}}{n} + 4\frac{\log n}{n}\right] \\
&\leq \sigma^2\left[1 + \frac{l(\theta_{m_{R,s}},\theta)}{\sigma^2}\right]\left(1 + \frac{d_{m_{R,s}}}{n}\right)\left[1 + 2\frac{\sqrt{2\log n}}{\sqrt{n - d_{m_{R,s}}}} + 4\frac{\log n}{n - d_{m_{R,s}}}\right] .
\end{aligned}$$

For n large enough, $d_{m_{R,s}}$ is smaller than $\frac{n}{2}$, and the last upper bound becomes:

$$\gamma_n\left(\widehat{\theta}_{m_{R,s}}\right)[1 + pen(m_{R,s})] \leq \sigma^2\left[1 + \frac{C(R,s)}{n^{\frac{s}{1+s}}}\right]^2\left(1 + 10\frac{\log(n)}{\sqrt{n}}\right) .$$

Hence, $\gamma_n\left(\widehat{\theta}_{m_{R,s}}\right)[1 + pen(m_{R,s})] \leq \gamma_n\left(\widehat{\theta}_m\right)[1 + pen(m)]$ if

$$\frac{l(\theta_{m_{\lfloor \log^2 n \rfloor}},\theta)}{\sigma^2} \geq 3\frac{C(R,s)}{n^{\frac{s}{1+s}}} \times \frac{1 + 10\log(n)/\sqrt{n}}{1 - 4\log(n)/\sqrt{n}} + 14\frac{\log(n)}{\sqrt{n}} .$$

As previously, this inequality always holds except for a finite number of n, since θ does not belong to any ellipsoid $\mathcal{E}_{s'}(R')$. Thus, outside an event of probability smaller than $\frac{L}{n^2}$, $d_{\widehat{m}}$ is larger than $\log^2 n$.

Let us now turn to the models of large dimension. Inequality (58) holds if the quantity

$$\|\boldsymbol{\epsilon}\|_n^2 \left(\frac{2d_{m_{R,s}}}{n-d_{m_{R,s}}} - \frac{2d_m}{n-d_m}\right) + \|\Pi_m \boldsymbol{\epsilon}\|_n^2 \left(1+\frac{2d_m}{n-d_m}\right) + \langle \Pi^\perp_{m_{R,s}} \boldsymbol{\epsilon}_{m_{R,s}}, \Pi^\perp_{m_{R,s}} \boldsymbol{\epsilon} + 2\boldsymbol{\epsilon}_{m_{R,s}}\rangle_n \left(1+\frac{2d_{m_{R,s}}}{n-d_{m_{R,s}}}\right) \quad (59)$$

is non-positive. The three following bounds hold outside an event of probability smaller than $\frac{L(\xi)}{n^2}$:

$$\begin{aligned}
\|\boldsymbol{\epsilon}\|_n^2 &\geq 1 - 4\frac{\sqrt{\log n}}{\sqrt{n}}\ , \\
\|\Pi_m \boldsymbol{\epsilon}\|_n^2 &\leq (1+\xi)\frac{d_m}{n}, \text{ for all models } m \text{ of dimension } d_m > \frac{n}{\log n}\ , \\
\langle \Pi^\perp_{m_{R,s}} \boldsymbol{\epsilon}_{m_{R,s}}, \Pi^\perp_{m_{R,s}} \boldsymbol{\epsilon} + 2\boldsymbol{\epsilon}_{m_{R,s}}\rangle_n &\leq l(\theta_{m_{R,s}},\theta)\left[\frac{n-d_{m_{R,s}}}{n} + 4\frac{\sqrt{(n-d_{m_{R,s}})\log n}}{n} + \frac{4\log n}{n}\right] \\
&+ 4\sqrt{l(\theta_{m_{R,s}},\theta)}\sigma\frac{\sqrt{(n-d_{m_{R,s}})\log n}}{n}\ .
\end{aligned}$$

Gathering these three inequalities we upper bound (59) by

$$\sigma^2\frac{d_m}{n-d_m}\left[-2+8\sqrt{\frac{\log n}{n}} + (1+\xi)\left(\frac{n+d_m}{n}\right)\right] + 2\sigma^2\frac{d_{m_{R,s}}}{n-d_{m_{R,s}}} + \\ +\sigma^2 L\left(1+\frac{d_{m_{R,s}}}{n}\right)\left(\frac{l(\theta_{m_{R,s}},\theta)}{\sigma^2} + \frac{\sqrt{l(\theta_{m_{R,s}},\theta)}}{\sigma}\right)\left(1+\sqrt{\frac{\log n}{n-d_{m_{R,s}}}}\right)\ .$$

The dimension of any model $m \in \mathcal{M}_{\lfloor n/2 \rfloor}$ is assumed to be smaller than $n/2$ and the dimensions of the models m considered are larger than $\frac{n}{\log n}$. For ξ small enough and n large enough, the previous expression is therefore upper bounded by

$$\sigma^2\frac{2}{\log n}\left[\frac{3}{2}(1+\xi) - 2 + 8\sqrt{\frac{\log n}{n}}\right] + L\sigma^2\left[\frac{R^{\frac{2}{1+s}}}{n^{\frac{s}{1+s}}} + \frac{R^{\frac{1}{1+s}}}{n^{\frac{a}{2(1+a)}}}\right]\ . \quad (60)$$

For n large enough, this last quantity is clearly non-positive.

All in all, we have proved that for n large enough outside an event of probability smaller than $\frac{L(R,s)}{n^2}$, it holds that

$$\log^2(n) < d_{m_*} < \frac{n}{\log n} \quad \text{and} \quad \log^2(n) < d_{\widehat{m}} < \frac{n}{\log n}\ .$$

□

Proof of Lemma 7.15. Arguing as in the proof of Theorem 3.1, we upper bound

$$-\overline{\gamma}_n(\widetilde{\theta}) - \gamma_n(\widetilde{\theta})pen(\widehat{m}) + \sigma^2 + \|\boldsymbol{\epsilon}\|_n^2 \leq l(\theta_{\widehat{m}},\theta)A_{\widehat{m}} + \sigma^2 B_{\widehat{m}} + (1-\kappa_2(n))l(\widetilde{\theta},\theta_{\widehat{m}})\ , \quad (61)$$

where $A_{\widehat{m}}$ and $B_{\widehat{m}}$ are respectively defined in (30) and in (31). We will fix the quantities $\kappa_1(n)$ and $\kappa_2(n)$ later. Besides, we define and bound the quantity $E_{\widehat{m}}$ as in (41).

Applying Lemma 7.2 and Lemma 7.4 and arguing as in the proofs of Lemma 7.6 and Lemma 7.7, there exists an event Ω_2 of large probability

$$\mathbb{P}(\Omega_1^c) \leq \exp\left[-n/8\right] + 5 \sum_{d=\log^2(n)}^{\frac{n}{\log n}} \exp\left[-\frac{2d}{\log n}\right] \leq \exp\left[-n/8\right] + \frac{5\log n}{2n^2(1-1/\log n)} \ ,$$

and such that conditionally on $\Omega_1 \cap \Omega_2$,

$$\begin{aligned}
\frac{\|\Pi_{\widehat{m}}^{\perp}\boldsymbol{\epsilon}_{\widehat{m}}\|_n^2}{l(\theta_{\widehat{m}},\theta)} &\geq \frac{n-d_{\widehat{m}}}{n} - 2\frac{\sqrt{2(n-d_{\widehat{m}})d_{\widehat{m}}/\log n}}{n}, \\
\frac{\|\Pi_{\widehat{m}}(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_{\widehat{m}})\|_n^2}{\sigma^2+l(\theta_{\widehat{m}},\theta)} &\leq \frac{d_{\widehat{m}}}{n} + \frac{2\sqrt{2}d_{\widehat{m}}}{n\sqrt{\log n}} + 4\frac{d_{\widehat{m}}}{n\log n}, \\
\frac{\|\Pi_{\widehat{m}}^{\perp}(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_{\widehat{m}})\|_n^2}{\sigma^2+l(\theta_{\widehat{m}},\theta)} &\geq \frac{n-d_{\widehat{m}}}{n} - 2\frac{\sqrt{2(n-d_{\widehat{m}})d_{\widehat{m}}/\log n}}{n} \\
\varphi_{\max}\left[(\mathbf{Z}_{\widehat{m}}^*\mathbf{Z}_{\widehat{m}})^{-1}\right] &\leq n^{-1}\left(1-\left(1+\sqrt{\frac{4}{\log n}}\right)\sqrt{\frac{d_{\widehat{m}}}{n}}\right)^{-2} \\
\|\boldsymbol{\epsilon}\|_n^2 &\leq 2 \\
E_{\widehat{m}} &\leq \frac{d_{\widehat{m}}+2\kappa_1^{-1}(n)}{n} + \frac{2}{n}\sqrt{\left[d_{\widehat{m}}+\left(2\kappa_1^{-1}(n)\right)^2\right]\frac{2d_{\widehat{m}}}{\log n}} + 8\kappa_1^{-1}(n)\frac{d_{\widehat{m}}}{n\log n} \ .
\end{aligned}$$

Gathering these six upper bounds, we are able to upper bound $A_{\widehat{m}}$ and $B_{\widehat{m}}$,

$$\begin{aligned}
A_{\widehat{m}} &\leq \kappa_1(n) + L_1\sqrt{\frac{d_{\widehat{m}}}{n\log n}} + \frac{d_{\widehat{m}}}{n}\left[-1+L_2\sqrt{\frac{d_{\widehat{m}}}{(n-d_{\widehat{m}})\log n}} + \kappa_2(n)\frac{1+L_3/\sqrt{\log(n)}}{\left[1-(1+\sqrt{\frac{4}{\log n}})\sqrt{\frac{d_{\widehat{m}}}{n}})\right]^2}\right] \ , \\
B_{\widehat{m}} &\leq \frac{d_{\widehat{m}}}{n}\left[-1+L_1\sqrt{\frac{d_{\widehat{m}}}{(n-d_{\widehat{m}})\log n}} + \kappa_2(n)\frac{1+L_2/\sqrt{\log(n)}}{\left[1-(1+\sqrt{\frac{4}{\log n}})\sqrt{\frac{d_{\widehat{m}}}{n}})\right]^2}\right] \\
&+ L_3\frac{d_{\widehat{m}}}{n}\left[\frac{\kappa_1^{-1}(n)}{d_{\widehat{m}}} + \frac{\kappa_1^{-1}(n)}{\log n} + \frac{1}{\sqrt{\log(n)}} + \frac{\kappa_1^{-1}(n)}{\sqrt{\log(n)d_{\widehat{m}}}}\right] \ .
\end{aligned}$$

Conditionally to the event Ω_1, the dimension of $\widehat{m}$ is moderate. Setting κ_1 to $\frac{1}{\log n}$, we get

$$\begin{aligned}
A_{\widehat{m}} &\leq \frac{L_1}{\log n} + \frac{d_{\widehat{m}}}{n}\left[-1+\frac{L_2}{\log n} + \kappa_2(n)\frac{1+\frac{L_3}{\sqrt{\log n}}}{\left[1-\frac{L_4}{\sqrt{\log(n)}}\right]^2}\right] \ , \\
B_{\widehat{m}} &\leq \frac{d_{\widehat{m}}}{n}\left[-1+\frac{L_1}{\log n} + \kappa_2(n)\frac{1+\frac{L_2}{\sqrt{\log n}}}{\left[1-\frac{L_3}{\sqrt{\log(n)}}\right]^2} + \frac{L_4}{\sqrt{\log n}}\right] \ .
\end{aligned}$$

Hence, there exists a sequence $\kappa_2(n)$ converging to one such that conditionally on $\Omega_1 \cap \Omega_2$, $B_{\widehat{m}}$ is non-positive and $A_{\widehat{m}}$ is bounded by $\frac{L}{\log n}$ when n is large enough. Coming back to the inequality (61) yields

$$\left[-\overline{\gamma}_n(\widetilde{\theta}) - \gamma_n(\widetilde{\theta})pen(\widehat{m}) - \sigma^2 + \|\boldsymbol{\epsilon}\|_n^2\right] \mathbf{1}_{\Omega_1\cap\Omega_2} \leq l(\widetilde{\theta},\theta)\left[\frac{L}{\log n} \vee (1-\kappa_2(n))\right] ,$$

which concludes the proof. □

Proof of Lemma 7.16. We follow a similar approach to the previous proof.

$$\overline{\gamma}_n(\widehat{\theta}_{m_*}) + \gamma_n(\widehat{\theta}_{m_*})pen(m_*) + \sigma^2 - \|\boldsymbol{\epsilon}\|_n^2 \quad \leq \quad C_{m_*} l(\theta_{m_*},\theta) + D_{m_*}\sigma^2 + \kappa_2(n) l(\widehat{\theta}_{m_*},\theta_{m_*}) \quad (62)$$

where for any model $m' \in \mathcal{M}_{\lfloor n/2 \rfloor}$, $C_{m'}$ and $D_{m'}$ are respectively defined as

$$\begin{aligned}
C_{m'} &= \kappa_1(n) + \frac{\|\Pi^{\perp}_{m'}\boldsymbol{\epsilon}_{m'}\|_n^2}{l(\theta_{m'},\theta)} - 1 + 2\frac{d_{m'}}{n-d_{m'}}\frac{\|\Pi^{\perp}_{m'}(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_{m'})\|_n^2}{l(\theta_{m'},\theta)+\sigma^2} \\
&- (1+\kappa_2(n))\frac{n}{\varphi_{\max}\left(\mathbf{Z}^*_{m'}\mathbf{Z}_{m'}\right)}\frac{\|\Pi_m(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_{m'})\|_n^2}{l(\theta_{m'},\theta)+\sigma^2} \\
D_{m'} &= \kappa_1^{-1}(n)\frac{\langle \Pi^{\perp}_{m'}\boldsymbol{\epsilon}, \Pi^{\perp}_{m'}\boldsymbol{\epsilon}_{m'}\rangle_n^2}{\sigma^2 l(\theta_{m'},\theta)} - \frac{\|\Pi_{m'}\boldsymbol{\epsilon}\|_n^2}{\sigma^2} \\
&- (1+\kappa_2(n))\frac{n}{\varphi_{\max}\left(\mathbf{Z}^*_{m'}\mathbf{Z}_{m'}\right)}\frac{\|\Pi_{m'}(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_{m'})\|_n^2}{l(\theta_{m'},\theta)+\sigma^2} + 2\frac{d_{m'}}{n-d_{m'}}\frac{\|\Pi^{\perp}_{m'}(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_{m'})\|_n^2}{l(\theta_{m'},\theta)+\sigma^2} .
\end{aligned}$$

We fix $\kappa_1(n) = 1/\log n$ whereas $\kappa_2(n)$ will be fixed later. Arguing as in the proof of Lemma 7.15, there exists an event Ω_3 of large probability

$$\mathbb{P}(\Omega_3^c) \leq \exp\left[-n/8\right] + 5\sum_{d=\log^2(n)}^{\frac{n}{\log n}} \exp\left[-\frac{2d}{\log n}\right] \leq \exp\left[-n/8\right] + \frac{5\log n}{2n^2(1-1/\log(n))} ,$$

such that conditionally on $\Omega_1 \cap \Omega_3$, the two following bounds hold:

$$\begin{aligned}
C_{m_*} &\leq \frac{L_1}{\log n} + \frac{d_{m_*}}{n}\left[1 + \frac{L_2}{\log n} - (1+\kappa_2(n))\frac{1+L_3\sqrt{\frac{2}{\log n}}}{\left[1+\frac{L_4}{\sqrt{\log n}}\right]^2}\right] , \\
D_{m_*} &\leq \frac{d_{m_*}}{n}\left[1 + \frac{L_1}{\log n} + \frac{L_2}{\sqrt{\log n}} - (1+\kappa_2(n))\frac{1+L_3\sqrt{\frac{2}{\log n}}}{\left[1+\frac{L_4}{\sqrt{\log n}}\right]^2}\right] ,
\end{aligned}$$

if n is large. The main difference with the proof of Lemma 7.15 lies in the fact that we now control the largest eigenvalue of $\mathbf{Z}^*_m\mathbf{Z}_m$ thanks to the second result of Lemma 7.4. There exists a sequence $\kappa_2(n)$ converging to 0 such that conditionally on $\Omega_1 \cap \Omega_3$, D_{m_*} is non-positive and C_{m_*} is bounded by $\frac{L}{\log n}$ when n is large. Coming back to (61) yields

$$\left[\overline{\gamma}_n(\widehat{\theta}_{m_*}) + pen(m_*) + \sigma^2 - \|\epsilon\|_n^2\right] \mathbf{1}_{\Omega_1\cup\Omega_3} \leq l(\widehat{\theta}_{m_*},\theta)\left[\frac{L}{\log n} \vee \kappa_2(n)\right] ,$$

which concludes the proof. □

7.6 Proof of Proposition 3.3

Proof of Proposition 3.3. The approach is similar to the proof of Proposition 1 in [9]. For any model $m \in \mathcal{M}_{\lfloor n/2 \rfloor}$, let us define

$$\Delta(m, m_{\lfloor n/2 \rfloor}) := \gamma_n\left(\widehat{\theta}_{m_{\lfloor n/2 \rfloor}}\right)\left[1 + pen(m_{\lfloor n/2 \rfloor})\right] - \gamma_n\left(\widehat{\theta}_m\right)\left[1 + pen(m)\right] .$$

We shall prove that with large probability the quantity $\Delta(m, m_{\lfloor n/2 \rfloor})$ is negative for any model m of dimension smaller than $n/4$. Hence, with large probability $d_{\widehat{m}}$ will be larger than $n/4$. Let us fix a model m of dimension smaller than $n/4$.

First, we use Expression (25) to lower bound $\gamma_n(\widehat{\theta}_m)$.

$$\begin{aligned}
\gamma_n\left(\widehat{\theta}_m\right) &= \|\Pi_m^{\perp}\left(\boldsymbol{\epsilon} + \boldsymbol{\epsilon}_{m_{\lfloor n/2 \rfloor}}\right)\|_n^2 + \|\Pi_m^{\perp}\left(\boldsymbol{\epsilon}_m - \boldsymbol{\epsilon}_{m_{\lfloor n/2 \rfloor}}\right)\|_n^2 + 2\langle \Pi_m^{\perp}\left(\boldsymbol{\epsilon} + \boldsymbol{\epsilon}_{m_{\lfloor n/2 \rfloor}}\right), \Pi_m^{\perp}\left(\boldsymbol{\epsilon}_m - \boldsymbol{\epsilon}_{m_{\lfloor n/2 \rfloor}}\right)\rangle_n \\
&\geq \|\Pi_m^{\perp}\left(\boldsymbol{\epsilon} + \boldsymbol{\epsilon}_{m_{\lfloor n/2 \rfloor}}\right)\|_n^2 - \left\langle \Pi_m^{\perp}\left(\boldsymbol{\epsilon} + \boldsymbol{\epsilon}_{m_{\lfloor n/2 \rfloor}}\right), \frac{\Pi_m^{\perp}\left(\boldsymbol{\epsilon}_m - \boldsymbol{\epsilon}_{m_{\lfloor n/2 \rfloor}}\right)}{\|\Pi_m^{\perp}\left(\boldsymbol{\epsilon}_m - \boldsymbol{\epsilon}_{m_{\lfloor n/2 \rfloor}}\right)\|_n} \right\rangle_n^2 ,
\end{aligned}$$

since $2ab \geq -a^2 - b^2$ for any number a and b. Hence, we may upper bound $\Delta(m, m_{\lfloor n/2 \rfloor})$ by

$$\begin{aligned}
\Delta(m, m_{\lfloor n/2 \rfloor}) &\leq \|\Pi_{m_{\lfloor n/2 \rfloor}}^{\perp}\left(\boldsymbol{\epsilon} + \boldsymbol{\epsilon}_{m_{\lfloor n/2 \rfloor}}\right)\|_n^2 \left[pen(m_{\lfloor n/2 \rfloor}) - pen(m)\right] \\
&- \left\|[\Pi_m^{\perp} - \Pi_{m_{\lfloor n/2 \rfloor}}^{\perp}]\left(\boldsymbol{\epsilon} + \boldsymbol{\epsilon}_{m_{\lfloor n/2 \rfloor}}\right)\right\|_n^2 \left[1 + pen(m)\right] \\
&+ \left\langle \Pi_m^{\perp}\left(\boldsymbol{\epsilon} + \boldsymbol{\epsilon}_{m_{\lfloor n/2 \rfloor}}\right), \frac{\Pi_m^{\perp}\left(\boldsymbol{\epsilon}_m - \boldsymbol{\epsilon}_{m_{\lfloor n/2 \rfloor}}\right)}{\|\Pi_m^{\perp}\left(\boldsymbol{\epsilon}_m - \boldsymbol{\epsilon}_{m_{\lfloor n/2 \rfloor}}\right)\|_n} \right\rangle_n^2 \left[1 + pen(m)\right] . \qquad (63)
\end{aligned}$$

Arguing as the proof of Lemma 2.1, we observe that $\|\Pi_{m_{\lfloor n/2 \rfloor}}^{\perp}\left(\boldsymbol{\epsilon} + \boldsymbol{\epsilon}_{m_{\lfloor n/2 \rfloor}}\right)\|_n^2 * n/[\sigma^2 + l(\theta_{m_{\lfloor n/2 \rfloor}})]$ follows a χ^2 distribution with $n - \lfloor n/2 \rfloor$ degrees of freedom. Analogously, the random variable $\|[\Pi_m^{\perp} - \Pi_{m_{\lfloor n/2 \rfloor}}^{\perp}]\left(\boldsymbol{\epsilon} + \boldsymbol{\epsilon}_{m_{\lfloor n/2 \rfloor}}\right)\|_n^2 * n/[\sigma^2 + l(\theta_{m_{\lfloor n/2 \rfloor}})]$ follows a χ^2 distribution with $(d_{m_{\lfloor n/2 \rfloor}} - d_m)$ degrees of freedom. Let us turn to the distribution of the third term. Coming back to the definition of ϵ_m, we observe that

$$\epsilon_m - \epsilon_{m_{\lfloor n/2 \rfloor}} = Y - X\theta_m - (Y - X\theta_{m_{\lfloor n/2 \rfloor}}) = X(\theta_m - \theta_{m_{\lfloor n/2 \rfloor}}) .$$

Hence, $\epsilon_m - \epsilon_{m_{\lfloor n/2 \rfloor}}$ is both independent of X_m and of $\epsilon + \epsilon_{m_{\lfloor n/2 \rfloor}}$. Consequently, by conditioning and unconditioning, we conclude that the random variable defined in (63) follows (up to a $[\sigma^2 + l(\theta_{m_{\lfloor n/2 \rfloor}})]/n$ factor) a χ^2 distribution with 1 degree of freedom.

Once again, we apply Lemma 7.2 and the classical deviation bound $\mathbb{P}\left(|\mathcal{N}(0,1)| \geq \sqrt{2x}\right) \leq 2e^{-x}$. Let x be some positive number smaller than one that we shall fix later. There exists an event Ω_x of probability larger than $1 - \exp(-nx/2) - 3\exp(-(n/4-1)x)\frac{1}{1-e^{-x}}$ such for any model of dimension smaller than $n/4$,

$$\begin{aligned}
\frac{\Delta(m, m_{\lfloor n/2 \rfloor})}{\sigma^2 + l(\theta_{m_{\lfloor n/2 \rfloor}})} &\leq \left(\frac{n - \lfloor n/2 \rfloor}{n}\right)\left(1 + 2\sqrt{x} + 2x\right)\left(pen(m_{\lfloor n/2 \rfloor}) - pen(m)\right) \\
&- \frac{\lfloor n/2 \rfloor - d_m}{n}(1 - 2\sqrt{x} - 2x)(1 + pen(m)) .
\end{aligned}$$

We now replace the penalty terms by their values thanks to Assumption (11). Conditionally to Ω_x, we obtain that

$$\frac{\Delta(m, m_{\lfloor n/2 \rfloor})}{\sigma^2 + l(\theta_{m_{\lfloor n/2 \rfloor}})} \leq \frac{\lfloor n/2 \rfloor - d_m}{n}\left\{4(1-\nu)(\sqrt{x} + x)\left[1 + \frac{d_m}{n - d_m}\right] - \nu(1 - 2\sqrt{x} - 2x)\right\} .$$

Since the dimension of the model m is smaller than $n/4$, $\frac{d_m}{n-d_m}$ is smaller than $1/3$. Hence, the last upper bound becomes

$$\frac{\Delta(m, m_{\lfloor n/2 \rfloor})}{\sigma^2 + l(\theta_{m_{\lfloor n/2 \rfloor}})} \leq \frac{\lfloor n/2 \rfloor - d_m}{n} \left\{ \frac{16}{3}(1-\nu)(\sqrt{x}+x) - \nu(1-2\sqrt{x}-2x) \right\} .$$

There exists some $x(\nu)$ such that conditionally on $\Omega_{x(\nu)}$, $\Delta(m, m_{\lfloor n/2 \rfloor})$ is negative for any model m of dimension smaller than $n/4$. Since $\mathbb{P}(\Omega^c_{x(\nu)})$ goes exponentially fast with ν to 0, there exists some $n_0(\nu, \delta)$ such that for any n larger than $n_0(\nu, \delta)$, $\mathbb{P}(\Omega^c_{x(\nu)})$ is smaller than δ. We have proved that with probability larger than $1-\delta$, the dimension of $\widehat{m}$ is larger than $n/4$.

Let us simultaneously lower bound the loss $l(\widehat{\theta}_m, \theta_m)$ for every model $m \in \mathcal{M}$ of dimension larger than $n/4$. In the sequel, $\succeq$ means "stochastically larger than". Thanks to (27), we stochastically lower bound $l(\widehat{\theta}_m, \theta_m)$

$$\begin{aligned} l(\widehat{\theta}_m, \theta_m) &\geq n\varphi_{\max}\left(\mathbf{Z}_m^* \mathbf{Z}_m\right)^{-1} \|\Pi_m(\boldsymbol{\epsilon} + \boldsymbol{\epsilon}_m)\|_n^2 \\ &\succeq \varphi_{\max}\left(n\mathbf{Z}_m^* \mathbf{Z}_m\right)^{-1} \|\Pi_m \boldsymbol{\epsilon}\|_n^2, \end{aligned}$$

where $\mathbf{Z}_m^* \mathbf{Z}_m$ follows a standard Wishart distribution with parameters (n, d_m). Applying Lemma 7.2 and Lemma 7.4 in order to simultaneously lower bound the loss $l(\widehat{\theta}_m, \theta_m)$, we find an event Ω' of probability larger than $1 - \frac{2\exp(-n/4)}{1-e^{-1/16}}$, such that

$$l(\widehat{\theta}_m, \theta_m)\mathbf{1}_{\Omega'} \geq \left(1 + \sqrt{\frac{d_m}{n}} + \sqrt{\frac{2d_m}{16n}}\right)^{-2} \frac{d_m}{2n}\sigma^2 \geq \frac{d_m}{8n}\sigma^2 ,$$

for any model $m \in \mathcal{M}$ of dimension larger than $n/4$. On the event $\Omega_{x(\nu)}$, the dimension $d_{\widehat{m}}$ is larger than $n/4$. As a consequence, $l(\widetilde{\theta}, \theta_{\widehat{m}})\mathbf{1}_{\Omega' \cap \Omega_{x(\nu)}} \geq \frac{\sigma^2}{32}$. All in all, we obtain

$$\begin{aligned} \mathbb{E}\left[l(\widetilde{\theta}, \theta)\right] &\geq l(\theta_{m_{\lfloor n/2 \rfloor}}, \theta) + \mathbb{E}\left[\mathbf{1}_{\Omega' \cap \Omega_{x(\nu)}} l(\widetilde{\theta}, \theta_{\widehat{m}})\right] \\ &\geq l(\theta_{m_{\lfloor n/2 \rfloor}}, \theta) + \left[1 - \mathbb{P}(\Omega^c_{x(\nu)}) - \mathbb{P}(\Omega'^c)\right] \frac{\sigma^2}{32} \\ &\geq l(\theta_{m_{\lfloor n/2 \rfloor}}, \theta) + L(\delta, \nu)\sigma^2 , \end{aligned}$$

if n is larger than some $n_0(\nu, \delta)$.

□

7.7 Proofs of the minimax lower bounds

All these minimax lower bounds are based on Birgé's version of Fano's Lemma [6].

Lemma 7.17. (Birgé's Lemma) *Let (Θ, d) be some pseudo-metric space and $\{\mathbb{P}_\theta, \theta \in \Theta\}$ be some statistical model. Let κ denote some absolute constant smaller than one. Then for any estimator $\widehat{\theta}$ and any finite subset Θ_1 of Θ, setting $\delta = min_{\theta, \theta' \in \Theta_1, \theta \neq \theta'} d(\theta, \theta')$, provided that $\max_{\theta, \theta' \in \Theta_1} \mathcal{K}(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) \leq \kappa \log|\Theta_1|$, the following lower bound holds for every $p \geq 1$,*

$$\sup_{\theta \in \Theta_1} \mathbb{E}_\theta[d^p(\widehat{\theta}, \theta)] \geq 2^{-p}\delta^p(1-\kappa) .$$

First, we compute the Kullback-Leibler divergence between the distribution $\mathbb{P}_\theta$ and $\mathbb{P}_{\theta'}$.

$$\mathcal{K}\left(\mathbb{P}_\theta;\mathbb{P}_{\theta'}\right)=\mathcal{K}\left(\mathbb{P}_\theta(X);\mathbb{P}_{\theta'}(X)\right)+\mathbb{E}_\theta\left[\mathcal{K}\left(\mathbb{P}_\theta(Y|X);\mathbb{P}_{\theta'}(Y|X)\right)\,|\,X\right]$$

The two marginal distributions $\mathbb{P}_\theta(X)$ and $\mathbb{P}_{\theta'}(X)$ are equal. The conditional distributions $\mathbb{P}_\theta(Y|X)$ and $\mathbb{P}_{\theta'}(Y|X)$ are Gaussian with variance σ^2 and with mean respectively equal to $X\theta$ and $X\theta'$. Hence, the conditional Kullback-Leibler divergence equals

$$\mathcal{K}\left(\mathbb{P}_\theta(Y|X);\mathbb{P}_{\theta'}(Y|X)\right)=\frac{\left[X(\theta-\theta')\right]^2}{2\sigma^2}\ .$$

Reintegrating with respect to X yields

$$\mathcal{K}\left(\mathbb{P}_\theta;\mathbb{P}_{\theta'}\right)=\frac{l(\theta',\theta)}{2\sigma^2}\ \text{and}\ \ \mathcal{K}\left(\mathbb{P}_\theta^{\otimes n};\mathbb{P}_{\theta'}^{\otimes n}\right)=n\frac{l(\theta',\theta)}{2\sigma^2}\ . \tag{64}$$

Proof of Proposition 4.1. First, we need a lower bound of the minimax rate of estimation on a subspace of dimension D.

Lemma 7.18. *Let D be some positive number smaller than p and r be some arbitrary positive number. Let S_D be the set of vectors in $\mathbb{R}^p$ whose support in included in $\{1,\dots,D\}$. Then, for any estimator $\widehat{\theta}$ of θ,*

$$\sup_{\theta\in S_D,\, l(0_p,\theta)\le Dr^2}\mathbb{E}_\theta\left[l(\widehat{\theta},\theta)\right]\ge LD\left[r^2\wedge\frac{\sigma^2}{n}\right]\ . \tag{65}$$

Let us fix some $D\in\{1,\dots,p\}$. Consider the set $\Theta_D:=\left\{\theta\in S_D, l(0_p,\theta)\le a_D^2R^2\right\}$. Since the a_j's are non increasing, it holds that

$$\sum_{i=1}^{p}\frac{l(\theta_{m_{i-1}},\theta_{m_i})}{a_i^2}\le\sum_{i=1}^{D}\frac{l(\theta_{m_{i-1}},\theta_{m_i})}{a_D^2}\le\frac{l(0_p,\theta)}{a_D^2}\le R^2\ ,$$

for any $\theta\in\Theta_D$. Hence Θ_D is included in $\mathcal{E}_a(R)$. Applying Lemma 7.18, we get

$$\begin{aligned}\inf_{\widehat{\theta}}\sup_{\theta\in\mathcal{E}_a(R)}&\ge LD\left[\frac{a_D^2R^2}{D}\wedge\frac{\sigma^2}{n}\right]\\&\ge L\left[a_D^2R^2\wedge\frac{D\sigma^2}{n}\right]\ .\end{aligned}$$

Taking the supremum over D in $\{1,\dots,p\}$ enables to conclude.

□

Proof of Lemma 7.18. Let us assume first that $\Sigma=I_p$. Consider the hypercube $\mathcal{C}_D(r):=\{0,r\}^D\times\{0\}^{p-D}$. Thanks to (64), we upper bound the Kullback-Leibler divergence between the distributions $\mathbb{P}_\theta$ and $\mathbb{P}_{\theta'}$

$$\mathcal{K}\left(\mathbb{P}_\theta^{\otimes n};\mathbb{P}_{\theta'}^{\otimes n}\right)\le\frac{nDr^2}{2\sigma^2}\ ,$$

where θ and θ' belong to $\mathcal{C}_D(r)$. Then, we apply Varshamov-Gilbert's lemma (e.g. Lemma 4.7 in [25]) to the set $\mathcal{C}_D(r)$.

Lemma 7.19 (Varshamov-Gilbert's lemma)**.** *Let $\{0,1\}^D$ be equipped with Hamming distance d_H. There exists some subset Θ of $\{0,1\}^D$ with the following properties*

$$d_H(\theta,\theta') > D/4 \text{ for every } (\theta,\theta') \in \Theta^2 \text{ with } \theta \neq \theta' \text{ and } \log|\Theta| \geq D/8 \ .$$

Combining Lemma 7.17 with the set Θ defined in the last lemma yields

$$\inf_{\widehat{\theta}} \sup_{\theta \in \mathcal{C}_D(r)} \mathbb{E}_\theta \left[d_H(\widehat{\theta},\theta) \right] \geq \frac{D}{16} \ ,$$

provided that $\frac{nDr^2}{2\sigma^2} \leq D/16$. Coming back to the loss function $l(.,.)$ yields

$$\inf_{\widehat{\theta}} \sup_{\theta \in \mathcal{C}_D(r)} \mathbb{E}_\theta \left[l(\widehat{\theta},\theta) \right] \geq LDr^2 \ ,$$

if $r^2 \leq L\frac{\sigma^2}{n}$. Finally, we get

$$\inf_{\widehat{\theta}} \sup_{\theta \in S_D,\ l(0_p,\theta) \leq Dr^2} \mathbb{E}_\theta \left[l(\widehat{\theta},\theta) \right] \geq LD \left[r^2 \wedge \frac{\sigma^2}{n} \right] \ .$$

If we no longer assume that the covariance matrix Σ is the identity, we orthogonalize the sequence X_i thanks to Gram-Schmidt process. Applying the previous argument to this new sequence of covariates allows to conclude. □

Proof of Corollary 4.2. This result follows from the upper bound on the risk of $\widetilde{\theta}$ in Theorem 3.1 and the minimax lower bound of Proposition 4.1. Let $\mathcal{E}_a(R)$ an ellipsoid satisfying $\frac{\sigma^2}{n} \leq R^2 \leq \sigma^2 n^\beta$, then $l(0_p,\theta)$ is smaller than $\sigma^2 n^\beta$. By Theorem 3.1, the estimator $\widetilde{\theta}$ defined with the collection $\mathcal{M}_{\lfloor n/2 \rfloor \wedge p}$ and $pen(m) = K\frac{d_m}{n-d_m}$ satisfies

$$\begin{aligned} \mathbb{E}_\theta \left[l(\widetilde{\theta},\theta) \right] &\leq L(K) \inf_{1 \leq i \leq \lfloor n/2 \rfloor \wedge p} \left\{ l(\theta_{m_i},\theta) + K\frac{i}{n-i}[\sigma^2 + l(\theta_{m_i},\theta)] \right\} + L(K,\beta)\frac{\sigma^2}{n} \\ &\leq L(K,\beta) \inf_{1 \leq i \leq \lfloor n/2 \rfloor \wedge p} \left[l(\theta_{m_i},\theta) + \frac{i}{n}\sigma^2 \right] \ . \end{aligned}$$

If θ belongs to $\mathcal{E}_a(R)$, then

$$l(\theta_{m_i},\theta) \leq a_{i+1}^2 \sum_{j=i+1}^{p} \frac{l(\theta_{m_j},\theta_{m_{j-1}})}{a_j^2} \leq R^2 a_{i+1}^2 \ ,$$

since the (a_i)'s are increasing. It follows that

$$\mathbb{E}_\theta \left[l(\widetilde{\theta},\theta) \right] \leq L(K,\beta) \inf_{1 \leq i \leq \lfloor n/2 \rfloor \wedge p} \left[R^2 a_{i+1}^2 + \frac{i}{n}\sigma^2 \right] \ . \tag{66}$$

Let us define $i^* := \sup\left\{ 1 \leq i \leq p \,,\, R^2 a_i^2 \geq \frac{\sigma^2 i}{n} \right\}$, with the convention $\sup \varnothing = 0$. Since $R^2 \geq \sigma^2/n$, i^* is larger or equal to one. By Proposition 4.1, the minimax rates of estimation is lower bounded as follows

$$\inf_{\widehat{\theta}} \sup_{\theta \in \mathcal{E}_a(R)} \mathbb{E}_\theta \left[l(\widehat{\theta},\theta) \right] \geq L \left[a_{i^*+1}^2 R^2 \vee \frac{\sigma^2 i^*}{n} \right] \geq L \left[a_{i^*+1}^2 R^2 + \frac{\sigma^2 i^*}{n} \right] \ .$$

If either $p \leq 2n$ or $a^2_{\lfloor n/2 \rfloor +1} R^2 \leq \sigma^2/2$, then i^* is smaller or equal to $\lfloor n/2 \rfloor \wedge p$ and we obtain thanks to (66) that

$$\begin{aligned} \mathbb{E}_\theta \left[l(\widetilde{\theta}, \theta) \right] &\leq L(K,\beta) \left[a^2_{i^*+1} R^2 + \frac{\sigma^2 i^*}{n} \right] \\ &\leq L(K,\beta) \inf_{\widehat{\theta}} \sup_{\theta \in \mathcal{E}_a(R)} \mathbb{E} \left[l(\widehat{\theta}, \theta) \right] . \end{aligned}$$

□

Proof of Proposition 4.3. First, we use (64) to upper bound the Kullback-Leibler divergence between the distributions corresponding to parameters θ and θ' in the set $\Theta[k,p](r)$

$$\mathcal{K} \left(\mathbb{P}_\theta^{\otimes n}; \mathbb{P}_{\theta'}^{\otimes n} \right) \leq \frac{nkr^2}{2\sigma^2} ,$$

since the covariates are i.i.d standard Gaussian variables. Let us state a combinatorial argument due to Birgé and Massart [7].

Lemma 7.20. *Let $\{0,1\}^p$ be equipped with Hamming distance d_H and given $1 \leq k \leq p/4$, define $\{0,1\}^p_k := \{x \in \{0,1\}^p : d_H(0,x) = k\}$. There exists some subset Θ of $\{0,1\}^p_k$ with the following properties*

$$d_H(\theta, \theta') > k/8 \text{ for every } (\theta, \theta') \in \Theta^2 \text{ with } \theta \neq \theta' \text{ and } \log |\Theta| \geq k/5 \log \left(\frac{p}{k} \right) .$$

Suppose that k is smaller than $p/4$. Applying Lemma 7.17 with Hamming distance d_H and the set $r\Theta$ introduced in Lemma 7.20 yields

$$\inf_{\widehat{\theta}} \sup_{\theta \in \Theta[k,p](r)} \mathbb{E}_\theta \left[d_H \left(\widehat{\theta}, \theta \right) \right] \geq \frac{k}{16} , \qquad \text{provided that} \qquad \frac{nkr^2}{2\sigma^2} \leq \frac{k}{10} \log \left(\frac{p}{k} \right) . \tag{67}$$

Since the covariates X_i are independent and of variance 1, the lower bound (67) is equivalent to

$$\inf_{\widehat{\theta}} \sup_{\theta \in \Theta[k,p](r)} \mathbb{E}_\theta \left[l \left(\widehat{\theta}, \theta \right) \right] \geq \frac{kr^2}{16}.$$

All in all, we obtain

$$\inf_{\widehat{\theta}} \sup_{\theta \in \Theta[k,p](r)} \mathbb{E}_\theta \left[l \left(\widehat{\theta}, \theta \right) \right] \geq Lk \left(r^2 \wedge \frac{\log \left(\frac{p}{k} \right)}{n} \sigma^2 \right) .$$

Since p/k is larger than 4, we obtain the desired lower bound by changing the constant L:

$$\inf_{\widehat{\theta}} \sup_{\theta \in \Theta[k,p](r)} \mathbb{E}_\theta \left[l \left(\widehat{\theta}, \theta \right) \right] \geq Lk \left(r^2 \wedge \frac{1 + \log \left(\frac{p}{k} \right)}{n} \sigma^2 \right) .$$

If p/k is smaller than 4, we know from the proof of Lemma 7.18, that

$$\inf_{\widehat{\theta}} \sup_{\theta \in \mathcal{C}_k(r)} \mathbb{E}_\theta \left[l \left(\widehat{\theta}, \theta \right) \right] \geq Lk \left(r^2 \wedge \frac{\sigma^2}{n} \right) .$$

We conclude by observing that $\log(p/k)$ is smaller than $\log(4)$ and that $\mathcal{C}_k(r)$ is included in $\Theta[k,p](r)$.

□

Proof of Proposition 4.5. Assume first the covariates (X_i) have a unit variance. If this is not the case, then one only has to rescale them. By Condition (22), the Kullback-Leibler divergence between the distributions corresponding to parameters θ and θ' in the set $\Theta[k,p](r)$ satisfies

$$\mathcal{K}\left(\mathbb{P}_\theta^{\otimes n};\mathbb{P}_{\theta'}^{\otimes n}\right) \leq (1+\delta)^2\frac{nkr^2}{2\sigma^2},$$

We recall that $\|.\|$ refers to the canonical norm in $\mathbb{R}^p$. Arguing as in the proof of Proposition 4.3, we lower bound the risk of any estimator $\widehat{\theta}$ with the loss function $\|.\|$,

$$\inf_{\widehat{\theta}}\sup_{\theta\in\Theta[k,p](r)}\mathbb{E}_\theta\left[\|\widehat{\theta}-\theta\|^2\right] \geq Lk\left(r^2\wedge\frac{1+\log\left(\frac{p}{k}\right)}{(1+\delta)^2n}\sigma^2\right) ,$$

Applying again Assumption (22) allows to obtain the desired lower bound on the risk

$$\inf_{\widehat{\theta}}\sup_{\theta\in\Theta[k,p](r)}\mathbb{E}_\theta\left[l(\widehat{\theta},\theta)\right] \geq Lk(1-\delta)^2\left(r^2\wedge\frac{1+\log\left(\frac{p}{k}\right)}{(1+\delta)^2n}\sigma^2\right) .$$

□

Proof of Proposition 4.6. In short, we find a subset $\Phi\subset\{1,\dots,p\}$ whose correlation matrix follows a 1/2-Restricted Isometry Property of size $2k$. We then apply Proposition 4.5 with the subset Φ of covariates.

We first consider the correlation matrix $\Psi_1(\omega)$. Let us pick a maximal subset $\Phi\subset\{1,\dots p\}$ of points that are $\lceil\log(4k)/\omega\rceil$ spaced with respect to the toroidal distance. Hence, the cardinality of Φ is $\lfloor p\lceil\log(4k)/\omega\rceil^{-1}\rfloor$. Assume that k is smaller than this quantity. We call C the correlation matrix of the points that belong to Φ. Obviously, for any $(i,j)\in\Phi^2$, it holds that $|C(i,j)|\leq 1/(4k)$ if $i\neq j$. Hence, any submatrix of C with size $2k$ is diagonally dominant and the sum of the absolute value of its non-diagonal elements is smaller than 1/2. Hence, the eigenvalues of any submatrix of C with size $2k$ lies between 1/2 and 3/2. The matrix C therefore follows a 1/2-Restricted Isometry Property of size $2k$. Consequently, we may apply Proposition 4.5 with the subset of covariates Φ and the result follows. The second case is handled similarly.

Definition of the correlations

Let us now justify why these correlations are well-defined when p is an odd integer. We shall prove that the matrices $\Psi_1(\omega)$ and $\Psi_2(t)$ are non-negative. Observe that these two matrices are symmetric and circulant. This means that there exists a family of numbers $(a_k)_{1\leq k\leq p}$ such that

$$\Psi_1(\omega)[i,j]=a_{i-j \bmod p} \quad \text{for any } 1\leq i,j\leq p .$$

Such matrices are known to be jointly diagonalizable in the same basis and their eigenvalues correspond to the discrete Fourier transform of (a_k). More precisely, their eigenvalues $(\lambda_l)_{1\leq l\leq p}$ are expressed as

$$\lambda_l := \sum_{k=0}^{p-1}\exp\left(\frac{2i\pi kl}{p}\right)a_k . \tag{68}$$

We refer to [27] Sect. 2.6.2 for more details. In the first example, a_k equals $\exp(-\omega(k\wedge(p-k))$, whereas it equals $[1+(k\wedge(p-k))]^{-t}$ in the second example.

CASE 1: Using the expression (68), one can compute λ_l.

$$\begin{aligned}
\lambda_l &= -1 + 2\sum_{k=0}^{(p-1)/2} \cos\left(\frac{2\pi kl}{p}\right)\exp(-k\omega) \\
&= -1 + 2\mathrm{Re}\left\{\sum_{k=0}^{(p-1)/2} \exp\left[k(i2\pi\frac{l}{p} - \omega\right]\right\} \\
&= -1 + 2\mathrm{Re}\left\{\frac{1 - e^{-\omega\frac{p+1}{2}}(-1)^l e^{i2\pi\frac{l}{p}}}{1 - e^{-\omega + i2\pi\frac{l}{p}}}\right\} \\
&= -1 + 2\frac{1 - e^{-\omega}\cos\left(\frac{2\pi l}{p}\right) + e^{-\omega(p+1)/2}(-1)^l\cos\left(\frac{\pi l}{p}\right)(e^{-\omega} - 1)}{1 + e^{-2\omega} - 2e^{-\omega}\cos\left(\frac{2\pi l}{p}\right)}
\end{aligned}$$

Hence, we obtain that

$$\lambda_l \geq 0 \Leftrightarrow 1 + 2e^{-\omega(p+1)/2}(-1)^l\cos\left(\frac{\pi l}{p}\right)\left(e^{-\omega} - 1\right) - e^{-2\omega} \geq 0 \ .$$

It is sufficient to prove that

$$1 - e^{-2\omega} + 2e^{-\omega(p+3)/2} - 2e^{-\omega(p+1)/2} \geq 0 \ .$$

This last expression is non-negative if ω equals zero and is increasing with respect to ω. We conclude that λ_l is non-negative for any $1 \leq l \leq p$. The matrix $\Psi_1(\omega)$ is therefore non-negative and defines a correlation.

CASE 2: Let us prove that the corresponding eigenvalues λ_l are non-negative.

$$\lambda_l = -1 + 2\sum_{k=0}^{(p-1)/2} \cos\left(\frac{2\pi kl}{p}\right)(k+1)^{-t}$$

Using the following identity

$$(k+1)^{-t} = \frac{1}{\Gamma(t)}\int_0^\infty e^{-r(k+1)}r^{t-1}dr \ ,$$

we decompose λ_l into a sum of integrals.

$$\lambda_l = \frac{1}{\Gamma(t)}\left\{\int_0^\infty r^{t-1}e^{-r}\left[-1 + 2\sum_{k=0}^{(p-1)/2}\cos\left(\frac{2\pi kl}{p}\right)e^{-rk}\right]\right\}dr \ .$$

The term inside the brackets corresponds to the eigenvalue for an exponential correlation with parameter r (CASE 1). This expression is therefore non-negative for any $r \geq 0$. In conclusion, the matrix $\Psi_2(t)$ is non-negative and the correlation is defined. □

Appendix

Proof of Lemma 7.1. We recall that $\gamma_n(\widehat{\theta}_m) = \|\mathbf{Y} - \Pi_m \mathbf{Y}\|_n^2$. Thanks to the definition (23) of ϵ and ϵ_m , we obtain the first result. Let us turn to the mean squared error $\gamma(\widehat{\theta}_m)$. In the following computation $\widehat{\theta}_m$ is considered as fixed and we only use that $\widehat{\theta}_m$ belongs to S_m. By definition,

$$\begin{aligned}\gamma(\widehat{\theta}_m) &= \mathbb{E}_{Y,X}\left[Y - X\widehat{\theta}_m\right]^2 = \sigma^2 + \mathbb{E}_X\left[X(\theta - \widehat{\theta}_m)\right]^2 \\ &= \sigma^2 + l(\theta_m, \theta) + l(\widehat{\theta}_m, \theta_m) \ ,\end{aligned}$$

since θ_m is the orthogonal projection of θ with respect to the inner product associated to the loss $l(.,.)$. We then derive that

$$l(\widehat{\theta}_m, \theta_m) = \mathbb{E}_{X_m}\left[X\left(\theta_m - \widehat{\theta}_m\right)\right]^2 = \left(\theta_m - \widehat{\theta}_m\right)^* \Sigma \left(\theta_m - \widehat{\theta}_m\right) \ .$$

Since $\widehat{\theta}_m$ is the least-squares estimator of θ_m, it follows from (23) that

$$l(\widehat{\theta}_m, \theta_m) = (\boldsymbol{\epsilon} + \boldsymbol{\epsilon}_m)^* \mathbf{X}_m (\mathbf{X}_m^* \mathbf{X}_m)^{-1} \Sigma_m (\mathbf{X}_m^* \mathbf{X}_m)^{-1} \mathbf{X}_m^* (\boldsymbol{\epsilon} + \boldsymbol{\epsilon}_m) \ .$$

We replace $\mathbf{X}_m$ by $\mathbf{Z}_m \sqrt{\Sigma_m}$ and therefore obtain

$$l(\widehat{\theta}_m, \theta_m) = (\boldsymbol{\epsilon} + \boldsymbol{\epsilon}_m)^* \mathbf{Z}_m (\mathbf{Z}_m^* \mathbf{Z}_m)^{-2} \mathbf{Z}_m^* (\boldsymbol{\epsilon} + \boldsymbol{\epsilon}_m) \ .$$

□

Proof of Lemma 2.1. Thanks to Equation (25), we know that $\gamma_n(\widehat{\theta}_m) = \|\Pi_m^{\perp}(\boldsymbol{\epsilon} + \boldsymbol{\epsilon}_m)\|_n^2$. The variance of $\epsilon + \epsilon_m$ is $\sigma^2 + l(\theta_m, \theta)$. Since $\boldsymbol{\epsilon} + \boldsymbol{\epsilon}_m$ is independent of $\mathbf{X}_m$, $\gamma_n(\widehat{\theta}_m) * n / [\sigma^2 + l(\theta_m, \theta)]$ follows a χ^2 distribution with $n - d_m$ degrees of freedom and the result follows.

Let us turn to the expectation of $\gamma(\widehat{\theta}_m)$. By (26), $\gamma(\widehat{\theta}_m)$ equals

$$\gamma\left(\widehat{\theta}_m\right) = \sigma^2 + l(\theta_m, \theta) + (\boldsymbol{\epsilon} + \boldsymbol{\epsilon}_{\widehat{m}})^* \mathbf{Z}_{\widehat{m}} (\mathbf{Z}_{\widehat{m}}^* \mathbf{Z}_{\widehat{m}})^{-2} \mathbf{Z}_{\widehat{m}}^* (\boldsymbol{\epsilon} + \boldsymbol{\epsilon}_{\widehat{m}}) \ ,$$

following the arguments of the proof of Lemma 7.1. Since $\epsilon + \epsilon_m$ and X_m are independent, one may integrate with respect to $\boldsymbol{\epsilon} + \boldsymbol{\epsilon}_m$

$$\mathbb{E}\left[\gamma(\widehat{\theta}_m)\right] = \left[\sigma^2 + l(\theta_m, \theta)\right] \left\{1 + \mathbb{E}\left[tr\left(\mathbf{Z}_m^* \mathbf{Z}_m\right)^{-1}\right)\right]\right\} \ ,$$

where the last term it the expectation of the trace of an inverse standard Wishart matrix of parameters (n, d_m). Thanks to [37], we know that it equals $\frac{d_m}{n - d_m - 1}$. □

Proof of Lemma 7.3. The random variable $\sqrt{\chi^2(d)}$ may be interpreted as a Lipschitz function with constant 1 on $\mathbb{R}^d$ equipped with the standard Gaussian measure. Hence, we may apply the Gaussian concentration theorem (see e.g. [25] Th. 3.4). For any $x > 0$,

$$\mathbb{P}\left(\sqrt{\chi^2(d)} \leq \mathbb{E}\left[\sqrt{\chi^2(d)}\right] - \sqrt{2x}\right) \leq \exp(-x) \ . \tag{69}$$

In order to conclude, we need to lower bound $\mathbb{E}\left[\sqrt{\chi^2(d)}\right]$. Let us introduce the variable $Z := 1 - \sqrt{\frac{\chi^2(d)}{d}}$. By definition, Z is smaller or equal to one. Hence, we upper bound $\mathbb{E}(Z)$ as

$$\mathbb{E}(Z) \leq \int_0^1 \mathbb{P}(Z \geq t) dt \leq \int_0^{\sqrt{\frac{1}{8}}} \mathbb{P}(Z \geq t) dt + \mathbb{P}(Z \geq \sqrt{\frac{1}{8}}) \ .$$

Let us upper bound $\mathbb{P}(Z \geq t)$ for any $0 \leq t \leq \sqrt{\frac{1}{8}}$ by applying Lemma 7.2

$$\begin{aligned}\mathbb{P}\left(Z \geq t\right) &\leq \mathbb{P}\left(\chi^2(d) \leq d\left[1-t\right]^2\right) \\ &\leq \mathbb{P}\left(\chi^2(d) \leq d - 2\sqrt{d}\sqrt{dt^2/2}\right) \leq \exp\left(-\frac{dt^2}{2}\right) ,\end{aligned}$$

since $t \leq 2 - \sqrt{2}$. Gathering this upper bound with the previous inequality yields

$$\begin{aligned}\mathbb{E}(Z) &\leq \exp\left(-\frac{d}{16}\right) + \int_0^{+\infty} \exp\left(-\frac{dt^2}{2}\right) dt \\ &\leq \exp\left(-\frac{d}{16}\right) + \sqrt{\frac{\pi}{2d}} .\end{aligned}$$

Thus, we obtain $\mathbb{E}\left(\sqrt{\chi^2(d)}\right) \geq \sqrt{d} - \sqrt{d}\exp(-d/16) - \sqrt{\pi/2}$. Combining this lower bound with (69) allows to conclude. □

Acknowledgements

I gratefully thank Pascal Massart for many fruitful discussions. I also would like to thank the referee for his suggestions that led to an improvement of the paper.

References

[1] H. Akaike. Statistical predictor identification. *Ann. Inst. Statist. Math.*, 22:203–217, 1970.

[2] H. Akaike. A new look at the statistical model identification. *IEEE Trans. Automatic Control*, AC-19:716–723, 1974. System identification and time-series analysis.

[3] S. Arlot. Model selection by resampling penalization, 2008. oai:hal.archives-ouvertes.fr:hal-00262478_v1.

[4] Y. Baraud, C. Giraud, and S. Huet. Gaussian model selection with an unknown variance. *Ann. Statist.*, 37(2):630–672, 2009.

[5] P. Bickel, Y. Ritov, and A. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics (to appear)*, 2009.

[6] L. Birgé. A new lower bound for multiple hypothesis testing. *IEEE Trans. Inform. Theory*, 51(4):1611–1615, 2005.

[7] L. Birgé and P. Massart. Minimum contrast estimators on sieves: exponential bounds and rates of convergence. *Bernoulli*, 4(3):329–375, 1998.

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

[9] L. Birgé and P. Massart. Minimal penalties for Gaussian model selection. *Probab. Theory Related Fields*, 138(1-2):33–73, 2007.

[10] F. Bunea, A. Tsybakov, and M. Wegkamp. Aggregation for Gaussian regression. *Ann. Statist.*, 35(4):1674–1697, 2007.

[11] F. Bunea, A. Tsybakov, and M. Wegkamp. Sparsity oracle inequalities for the Lasso. *Electron. J. Stat.*, 1:169–194 (electronic), 2007.

[12] E. Candès and Y. Plan. Near-ideal model selection by l_1 minimization. *Ann. Statist. (to appear)*, 2009.

[13] E. Candes and T. Tao. The Dantzig selector: statistical estimation when p is much larger than n. *Ann. Statist.*, 35(6):2313–2351, 2007.

[14] E. J. Candes and T. Tao. Decoding by linear programming. *IEEE Trans. Inform. Theory*, 51(12):4203–4215, 2005.

[15] R. G. Cowell, A. P. Dawid, S. L. Lauritzen, and D. J. Spiegelhalter. *Probabilistic networks and expert systems*. Statistics for Engineering and Information Science. Springer-Verlag, New York, 1999.

[16] N. A. C. Cressie. *Statistics for spatial data*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons Inc., New York, 1993. Revised reprint of the 1991 edition, A Wiley-Interscience Publication.

[17] K. R. Davidson and S. J. Szarek. Local operator theory, random matrices and Banach spaces. In *Handbook of the geometry of Banach spaces, Vol. I*, pages 317–366. North-Holland, Amsterdam, 2001.

[18] V. H. de la Peña and E. Giné. *Decoupling*. Probability and its Applications (New York). Springer-Verlag, New York, 1999. From dependence to independence, Randomly stopped processes. U-statistics and processes. Martingales and beyond.

[19] C. Giraud. Estimation of Gaussian graphs by model selection. *Electron. J. Stat.*, 2:542–563, 2008.

[20] T. Gneiting. Power-law correlations, related models for long-range dependence and their simulation. *J. Appl. Probab.*, 37(4):1104–1109, 2000.

[21] M. Kalisch and P. Bühlmann. Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *J. Mach. Learn. Res.*, 8:613–636, 2007.

[22] B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *Ann. Statist.*, 28(5):1302–1338, 2000.

[23] S. L. Lauritzen. *Graphical models*, volume 17 of *Oxford Statistical Science Series*. The Clarendon Press Oxford University Press, New York, 1996. Oxford Science Publications.

[24] C.L. Mallows. Some comments on C_p. *Technometrics*, 15:661–675, 1973.

[25] P. Massart. *Concentration inequalities and model selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin, 2007. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard.

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

[27] H. Rue and L. Held. *Gaussian Markov Random Fields: Theory and Applications*, volume 104 of *Monographs on Statistics and Applied Probability*. Chapman & Hall/CRC, London, 2005.

[28] K. Sachs, O. Perez, D.Pe'er, D. A. Lauffenburger, and G. P. Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308:523–529, 2005.

[29] J. Schäfer and K. Strimmer. An empirical bayes approach to inferring large-scale gene association network. *Bioinformatics*, **21**:754–764, 2005.

[30] G. Schwarz. Estimating the dimension of a model. *Ann. Statist.*, 6(2):461–464, 1978.

[31] R. Shibata. An optimal selection of regression variables. *Biometrika*, 68(1):45–54, 1981.

[32] C. Stone. An asymptotically optimal histogram selection rule. In *Proceedings of the Berkeley conference in honor of Jerzy Neyman and Jack Kiefer, Vol. II (Berkeley, Calif., 1983)*, Wadsworth Statist./Probab. Ser., pages 513–520, Belmont, CA, 1985. Wadsworth.

[33] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1):267–288, 1996.

[34] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1):267–288, 1996.

[35] A. Tsybakov. Optimal rates of aggregation. In *16th Annual Conference on Learning Theory*, volume 2777, pages 303–313. Springer-Verlag, 2003.

[36] N. Verzelen and F. Villers. Goodness-of-fit tests for high-dimensional gaussian linear models. *Ann. Statist. (to appear)*, 2009.

[37] D. von Rosen. Moments for the inverted Wishart distribution. *Scand. J. Statist.*, 15(2):97–109, 1988.

[38] M. J. Wainwright. Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting. Technical Report 725, Department of Statistics, UC Berkeley, 2007.

[39] A. Wille, P. Zimmermann, E. Vranova, A. Fürholz, O. Laule, S. Bleuler, L. Hennig, A. Prelic, P. von Rohr, L. Thiele, E. Zitzler, W. Gruissem, and P. Bühlmann. Sparse graphical Gaussian modelling of the isoprenoid gene network in *arabidopsis thaliana*. *Genome Biology*, **5**(11), 2004.

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

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

Centre de recherche INRIA Saclay – Île-de-France
Parc Orsay Université - ZAC des Vignes
4, rue Jacques Monod - 91893 Orsay Cedex (France)

Centre de recherche INRIA Bordeaux – Sud Ouest : Domaine Universitaire - 351, cours de la Libération - 33405 Talence Cedex
Centre de recherche INRIA Grenoble – Rhône-Alpes : 655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier
Centre de recherche INRIA Lille – Nord Europe : Parc Scientifique de la Haute Borne - 40, avenue Halley - 59650 Villeneuve d'Ascq
Centre de recherche INRIA Nancy – Grand Est : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex
Centre de recherche INRIA Paris – Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex
Centre de recherche INRIA Rennes – Bretagne Atlantique : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex
Centre de recherche INRIA Sophia Antipolis – Méditerranée : 2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399