Comparative utilizations of Information Criteria for Gaussian regression on a random design.

Guilhem Coq

Electronic Journal of Statistics
Vol. 0 (0000)
ISSN: 1935-7524
DOI: 10.1214/154957804100000000

Comparative utilizations of Information Criteria for Gaussian regression on a random design. *

Guilhem Coq

Laboratoire de Mathématiques et Applications
Université de Poitiers, France
e-mail: coq@math.univ-poitiers.fr

Abstract: We consider the problem of estimating an unknown function $f^\star$ in the setting of Gaussian regression on a random design. To this end, we use general Information Criteria, also called penalized likelihood criteria. We introduce several comparative methods of use of those criteria that present the advantage to have reasonnable computational complexity. We also show that those methods are as efficient as classical ones since they satisfy good asymptotic properties as well as an oracle inequality.

Keywords and phrases: Information Criteria, Model selection, Gaussian regression, Oracle inequality.

Contents

*This is an original survey paper

Introduction

We study the problem of Gaussian regression in the case called random design by Baraud [3] or Birgé [5]. In this setting we consider a set of abscisses $x^n = (x_1, \ldots, x_n)$ that is a sample from a specified density w on an interval I of $\mathbb{R}$. The ordinates $y^n = (y_1, \ldots, y_n)$ are the images of those abscisses by an unknown function $f^\star$ deteriorated by a Gaussian white noise. By opposition, the more common fixed design , studied for instance by Baraud [2], considers the same problem except the abscisses are deterministic.

In all that paper, we address the regression problem using information criteria. Let us cite for classical references on this subject Akaike [1] who gives the AIC criterion in the early 70's, Schwarz [17] who presents the Bayesian Information Criterion. Rissanen introduces the notion of stochastic complexity [15, 16] which, along with the MDL principle [14, 8], allows to derive RIC, Rissanen Information Criterion, which is similar to BIC. In a general setting, Nishii studies the asymptotic properties of models selected by a general information criterion in [11].

More recently, we observe an interest for non-asymptotic study of model selection via information criteria. The conjoint work of Barron, Birgé and Massart [4, 6, 9] gives a lot of tools allowing to derive non-asymptotic inequalities for the risks of a model selection procedure. In this optic, Castellan studies in [7]

such bounds for the problem of the histogram while Baraud [2, 3] is interested in the problem of regression, also the topic of this paper.

One of our main concern is to provide methods of model selection, referred to as comparative methods, which are both efficient and fast. The main idea of those methods was firstly introduced by Nishii [10, 11], studied in the regression setting by Rao and Wu [13], and discussed recently in a more general setting by the same authors in [12]. Here, we will use the classical comparative method, referred to as leave-one approach in [12]. We will also introduce several variations of it. The efficiency of those methods will be studied in terms of the behavior of the selected model as well as in terms of risks. As for the rapidity, we will require the methods to present an algorithm that does not require too many computations in view of an implementation on machines.

Part 1 and 2.2 present respectively the modelization and the information criterion used throughout this paper. Comparative methods of use of this criterion are defined in part 2.3. Using them, we show theorem 3.1 which deals with asymptotic stability, theorem 4.1 which studies the asymptotic risk and theorem 4.2 which gives an oracle inequality. Appendices A to E present the proofs of those theorems.

1. Modelization

1.1. Regression space

Let I be an interval of $\mathbb{R}$ endowed with the Lebesgue measure λ and w be a given nonnegative function with integral 1 assumed to vanish only on a set of I of Lebesgue measure 0. The following defines a scalar product on $L^2 := L^2(I, wd\lambda)$

$$\langle f, g \rangle_w = \int_I fgwd\lambda. \tag{1.1}$$

whose associated norm and squared-distance are denoted by $\|.\|_w$ and d_w^2.

We choose $F = \text{Vect}(f_1, \ldots, f_d)$ a d-dimensional subspace of L^2 where $d \leq n$ is not allowed to depend on n. For any support $S \subset \{1, \ldots, d\}$ we denote by F_S the $|S|$-dimensionnal subspace of F

$$F_S = \text{Vect}\left\{f_j,\ j \in S\right\}. \tag{1.2}$$

We choose an unknown function $f^\star \in F$, write $f^\star = \sum a_j^\star f_j$ and call $S^\star$ its support

$$S^\star = \{j \text{ such that } a_j^\star \neq 0\}. \tag{1.3}$$

Recall that, in all that follows, the function f_j belong to L^2. In the sequel, we will, at some points, make the following assumption:

$$f_j \in L^8(I, wd\lambda),\ j = 1, \ldots, d. \tag{H8}$$

1.2. Observation space

We endow the observations space $\mathbb{R}^n$ with the canonical scalar product $\langle .,. \rangle_n$ and its associated norm and squared-distance $\|.\|_n$, d^2.

For $x \in I^n$ and $f \in F$, we denote by $f(x)$ the vector $(f(x_1, \ldots, f(x_n))^T$. To a support $S \subset \{1, \ldots, d\}$ we also associate the subspace

$$E_S = \text{Vect}\,\{f_j(x),\ j \in S\} \subset \mathbb{R}^n \tag{1.4}$$

and shorten $E_{\{1,\ldots,d\}}$ to E.

1.3. Modelization

Let $X = X_1, \ldots, X_n$ be independent variable with density w on I. Let also G be a n-dimensionnal gaussian white noise $G \sim \mathcal{N}(0, \sigma^2 \text{id}_n)$ independent of X. We modelize the abscisses X and ordinates Y of a set of n points in the plane by a regression model on random design :

$$Y = f^\star(X) + G \tag{1.5}$$

Namely, the abscisses are given by a sample of the law w and the ordinates by the images of those abscisses by the unknown function $f^\star$ deteriorated by the noise. Note that all the spaces E_S (1.4) become random.

2. Information criteria and their use

2.1. Maximum likelihood

We are interested in twice the opposite of the log-likelihood of a realization (x, y) of (X, Y) relatively to a function $f \in F$. In our Gaussian setting, it is classical that this quantity, dropping terms not depending on f, is the quadratic error :

$$\begin{array}{rcl} l : F \sim \mathbb{R}^d & \to & \mathbb{R} \\ f & \mapsto & \displaystyle\sum_{i=1}^{n} (y_i - f(x_i))^2 \end{array}$$

Consequently, for any support S, the maximization of the likelihood of (x, y) within the regression subspace F_S amounts to the minimization of the quadratic error l on F_S. We denote by $\widehat{f}_S \in F_S$ the function realizing that minimization :

$$d^2(y, E_S) = \min_{f \in F_S} l(f) =: l(\widehat{f}_S). \tag{2.1}$$

2.2. The criterion

Let us choose any function $\alpha : \mathbb{N} \to \mathbb{R}^+$ referred to in the sequel as the penalty term. We consider the information criterion

$$\mathrm{IC}(S) = d^2(y, E_S) + |S|\alpha(n),\ S \subset \{1, \dots, d\}. \tag{2.2}$$

In the sequel, the following hypothesis on $\alpha(n)$ will be used :

$$\begin{array}{ll} (i) & \alpha(n) = o(n) \\ (ii) & \ln \ln n = o(\alpha(n)). \end{array} \tag{H_α}$$

Note that, in the case where the error G in (1.5) is not Gaussian, the quantity $\mathrm{IC}(S)$ is not strictly speaking an information criterion since those are defined as penalized maximum likelihood.

An inequality of the type $\mathrm{IC}(S) \leq \mathrm{IC}(T)$ expresses the fact that the support S realizes a better trade-off between the goodness of fit (measured by the likelihood term $l(\widehat{f}_S)$) and the complexity of the model needed to obtain such a fit (measured by the penalty $|S|\alpha(n)$) than the support T does.

If one wants to use such a criterion for the estimation of $f^\star$, one has to choose a penalty function $\alpha(n)$ and a sequence of computations of IC that will lead to the selection of a support $\widehat{S}$ and thus to the estimation of $f^\star$ by $\widehat{f}_{\widehat{S}}$.

In this paper, we are not only interested in properly choosing $\alpha(n)$ but also in efficient and fast methods able to select $\widehat{S}$. We present those methods now.

2.3. Methods

By method, we understand a sequence of computations of $\mathrm{IC}(S)$ that eventually leads to the selection of a support $\widehat{S}$ well suited to estimate $f^\star$ by $\widehat{f}_{\widehat{S}}$. We define five of those in the sequel : formulaes (2.3), (2.4), (2.5), (2.6) and part 2.3.5. The global method (2.3) is the most used when the problem of computation speed is not under consideration ; this is the case for instance in Baraud's work [2, 3]. The comparative method (2.4) solves the computationnal problem. The remaining ones are derived from (2.4), in the specific aim of obtaining an oracle inequality (theorem 4.2) on the risk for the descending comparative method (part 2.3.5).

All the methods we use allow to select a support that may take any value in $\mathcal{P}(\{1, \dots, d\})$. In this sense, our basis $(f_j)_j$ is not required to present a natural order as, for instance, basis of polynomials or wavelets would do. Our methods will always select a set of functions to be used for regression, without favouring any of them and regardless of which kind of functions are mixed in the basis.

Let us also stress that, even though this paper is written in the context of linear regression, those methods may be simply transposed to many other parametric model selection problems.

2.3.1. Global method

This is the most straightforward, but also the most computationnal method. The estimated support is chosen as

$$\widehat{S} = \text{Argmin}\left\{\text{IC}(S), S \in \mathcal{P}(\{1,\dots,d\})\right\}. \tag{2.3}$$

2.3.2. Comparative method

Rather than testing any support as in the global method, Nishii [10, 11] suggests the following method, also called leave-one approach by Rao and Wu in [13]. From now on, the expression $-j$ denotes the support $\{1,\dots,d\} \setminus \{j\}$.

$$\begin{aligned} \text{IC}_{\text{ref}} &= \text{IC}(\{1,\dots,d\}) \\ \widehat{S} &= \left\{j \in \{1,\dots,d\} \text{ such that } \text{IC}_{\text{ref}} \leq \text{IC}(-j)\right\} \end{aligned} \tag{2.4}$$

2.3.3. Reversed comparative method

Looking at method (2.4) one may ask why not to consider the following :

$$\begin{aligned} \text{IC}_{\text{ref}} &= \text{IC}(\emptyset) \\ \widehat{S} &= \left\{j \in \{1,\dots,d\} \text{ such that } \text{IC}(\{j\}) \leq \text{IC}_{\text{ref}}\right\}. \end{aligned} \tag{2.5}$$

The reversed method might be transposed to the case where the space of regression has infinite dimension. Indeed, even though the basis (in the classical or Hilbert sense) of F was infinite, using the reversed method would never require to compute an IC with an infinite number of free parameters. By opposition the reference of the regular comparative method (2.4) are not computable in the infinite dimensionnal case.

As will be discussed in part C.2, when the basis $(f_1,\dots,f_d)$ is orthogonal relatively to the scalar product (1.1), the reversed method is asymptotically equivalent to the regular method (2.4). However, in the non-orthogonal case, this method presents asymptotic flaws that will also be precisely described in part C.2. In order to handle these flaws, we present the next method.

2.3.4. Adapted reversed comparative method

For any $j \in \{1,\dots,d\}$, choose a non-vanishing function f_j^N that is normal, relatively to the scalar product (1.1), to the hyperplane F_{-j} defined in (1.2). The orientation of f_j^N as well as its norm does not matter in the sequel.

For any j, the family $\{f_k, k \neq j\} \cup \{f_j^N\}$ is a basis of F and a function does not live in F_{-j} if and only if it has a component along f_j^N. The idea comes that, instead of determining wether $f^\star$ should have a component along f_j, we

will determine if it should have one along f_j^N by following the so called adapted reversed comparative method :

$$\begin{aligned} \mathrm{IC}_{\mathrm{ref}} &= \mathrm{IC}(\emptyset) \\ \widehat{S} &= \left\{ j \in \{1,\dots,d\} \text{ such that } l\left(\widehat{f}_j^N\right) + \alpha(n) \leq \mathrm{IC}_{\mathrm{ref}} \right\}, \end{aligned} \tag{2.6}$$

where $\widehat{f}_j^N$ denotes the function h colinear to f_j^N realizing the minimum of $d^2(y, h(x))$.

Note that the basis $(f_j^N)_j$ is orthogonal if and only if $(f_j)_j$ is. In this case, f_j^N is colinear to f_j which makes (2.5) and (2.6) equivalent methods of selection.

It will be shown in theorem 3.1 that this method satisfies a convergence theorem even though the basis $(f_1,\dots,f_d)$ is not orthogonal.

2.3.5. Descending comparative method

The descending comparative method is designed specifically in the aim of using results from Baraud [3] in order to show theorem 4.2.

Firstly let us set

$$\begin{aligned} S^{(0)} &= \{1,\dots,d\} \\ \mathrm{IC}_{\mathrm{ref}}^{(0)} &= \mathrm{IC}(S^{(0)}). \end{aligned}$$

The first step of the descending method produces new quantities that have superscript (1) as follows

$$\begin{aligned} C^{(1)} &= \left\{ j \in S^{(0)},\ \mathrm{IC}\left(S^{(0)} \setminus \{j\}\right) \leq \mathrm{IC}_{\mathrm{ref}}^{(0)} \right\} \\ J^{(1)} &= \mathrm{Argmin}\left\{ \mathrm{IC}\left(S^{(0)} \setminus \{j\}\right),\ j \in C^{(1)} \right\}. \end{aligned} \tag{2.7}$$

This way, among the functions of $C^{(1)}$ found useless by the criterion, $J^{(1)}$ is the worst one. This is consequently the function we should remove in priority. This is what we do now by refreshing our reference with superscript (1) :

$$\begin{aligned} S^{(1)} &= S^{(0)} \setminus \{J^{(1)}\} \\ \mathrm{IC}_{\mathrm{ref}}^{(1)} &= \mathrm{IC}(S^{(1)}). \end{aligned} \tag{2.8}$$

From there, we start a second step by computing useless functions and the worst one by

$$\begin{aligned} C^{(2)} &= \left\{ j \in S^{(1)},\ \mathrm{IC}\left(S^{(1)} \setminus \{j\}\right) \leq \mathrm{IC}_{\mathrm{ref}}^{(1)} \right\} \\ J^{(2)} &= \mathrm{Argmin}\left\{ \mathrm{IC}\left(S^{(1)} \setminus \{j\}\right),\ j \in C^{(2)} \right\} \end{aligned}$$

and refresh again our reference by adding 1 to all superscripts in (2.8).

This process is repeated until the random final step k_f+1 where $C^{(k_f+1)} = \emptyset$. This means that the criterion does not reject functions anymore and that the current support $S^{(k_f)}$ should be our estimator $\widehat{S}$.

TABLE 1
Methods and their complexities.

Method	Complexity
Global	2^d
comparative	$d+1$
Reversed comparative	$d+1$
Adapted reversed comparative	$d+1$
Descending comparative	$\leq d(d+1)/2$

2.4. Methods complexities

We define here an integer willing to reflect the complexity of a particular method in terms of computations. This integer is simply the number of ICs that one needs to compute in order to select $\widehat{S}$. Table 1 sums up the complexities of the different methods used here. For the descending method the number of computation is random so we only give an upper bound.

All methods derived from the comparative one have polynomial complexity contrarily to the global method that has exponential complexity. Nevertheless they allow a precise selection of the support in the sense that for each of them $\widehat{S}$ may take any value in $\mathcal{P}(\{1,\dots,d\})$. They are what we have called fast methods. The remainder of the paper is devoted to explain in which ways they are also efficient .

3. Stability of comparative methods

Recall that the unknown function $f^\star$ has support $S^\star$ as in (1.3). The following theorem, shown in appendix C, gives conditions on the penalty $\alpha(n)$ in (2.2) that ensure convergence of $\widehat{S}$ to $S^\star$ when the criterion is used with the comparative methods we introcuced in previous section.

Theorem 3.1 *Assume that (H_α) holds. Then*

(i) the comparative method (2.4) is strongly consistent in the sense that $\widehat{S}$ converges to $S^\star$ almost surely,
(ii) the reversed comparative method (2.5) is strongly consistent provided that the basis $(f_1,\dots,f_d)$ is orthonormal relatively to the scalar product (1.1),
(iii) the adapted reversed comparative method (2.6) is strongly consistent,
(iv) the descending comparative method (part 2.3.5) is strongly consistent.

More precisely, in any of those cases, conditions (i) and (ii) in (H_α) ensure respectively $S^\star \subset \widehat{S}$ and $\widehat{S} \subset S^\star$ a.s. above a certain rank.

4. Study of the risks

The aim of this section is to study the behaviour of the risk following the selection of a support $\widehat{S}$ resulting from the use of a comparative method described in part 2.3.

The fact that, in our random design setting, $\widehat{f}_S$ defined in (A.6) is not square integrable prevents us from computing risks in general. We handle this issue by using, in this entire section 4 the truncated estimator $\widetilde{f}_S$, defined in (A.7).

4.1. Asymptotics of the risks

Our aim is to derive asymptotic results similar to those given in [10], except we handle the random design case and use more comparative methods.

4.1.1. The ideal case

Assume for a moment that the user knows the support $S^\star$. Then he will estimate $f^\star$ by $\widetilde{f}_{S^\star}$ and get an oracle risk denoted by $\mathcal{O}R(n, S^\star)$. As shown in lemma B.2, under (H8), this risk satisfies :

$$\mathbb{E}\left[\left\|\widetilde{f}_{S^\star} - f^\star\right\|_w^2\right] \sim \frac{\sigma^2|S^\star|}{n} \tag{4.1}$$

4.1.2. Risk of comparative methods

Theorem 4.1 *Assume that (H_α) and (H8) hold and that either of the following method has been used to determine $\widehat{S}$:*

- ⋆ *comparative method (2.4),*
- ⋆ *reversed comparative method in the orthonormal case (2.5),*
- ⋆ *adapted reversed comparative method (2.6),*
- ⋆ *descending comparative method (part 2.3.5).*

Then the estimation of $f^\star$ by $\widetilde{f}_{\widehat{S}}$ defined in (A.7) presents a risk $R(n)$ equivalent to the oracle risk (4.1) in the sense that

$$R(n) \sim \frac{\sigma^2|S^\star|}{n}$$

The proof of this theorem is given in appendix D.

4.2. An oracle inequality for the risk of the descending method

Our main purpose here is to establish an oracle inequality on the risk achieved by the estimator of $f^\star$ resulting from the use of an information criterion of the form (2.2) along with the descending comparative method (part 2.3.5). The result is given in theorem 4.2.

4.2.1. Definition of the risk

To $S \subset \{1, \dots, d\}$ a support we associate an unknown risk $R^\star(S)$ by

$$R^\star(S) = d_w^2(f^\star, F_S) + \sigma^2|S|/n. \tag{4.2}$$

This quantity $R^\star(S)$ is the actual risk resulting from the estimation of $f^\star$ by $\widehat{f}_S$, but in the fixed design setting.

In our random design and under certain integrability hypothesis, we have shown that one may obtain an explicit value of this risk involving Gram matrices similar to those used in appendix A. However, those expressions are quite useless and we do not give them here. Let us just say that, in the case where $S^\star \subset S$, the risk remains almost unchanged, while in the other case the estimator is biased and a new variance term appears. Also stress that, as n grows, the law of large numbers ensures that those differences vanish.

Regardless of this remark, and as Baraud in [3], we work in the sequel in random design, with the $\widetilde{f}_S$ estimator defined in (A.7), but still with quantities $R^\star(S)$.

4.2.2. Baraud's result

Let $\mathcal{F}$ be a family of supports. We associate with it the oracle risk

$$\mathcal{O}_{\mathcal{F}}(f^\star) = \min_{S \in \mathcal{F}} R^\star(S). \tag{4.3}$$

This oracle is the minimum risk the user could achieve if he knew by advance the support $S_{\mathcal{O}}$ in $\mathcal{F}$ realizing that minimum. Note that there is no reason why $S_{\mathcal{O}}$ would equal $S^\star$.

Let us assume for now that the user has chosen the global method (2.3). The only thing he knows is what his criterion has found is the best support, namely

$$\widehat{S} = \operatorname{Argmin} \left\{ \mathrm{IC}(Y, S),\ S \in \mathcal{F} \right\}.$$

Baraud shows in [3] that the user did not take too much risks in the sense that

$$\mathbb{E}\left[\left\| f^\star - \widetilde{f}_{\widehat{S}} \right\|_w^2 \right] \leq C \mathcal{O}_{\mathcal{F}}(f^\star). \tag{4.4}$$

where C is a constant depending on θ appearing in the penalty (4.5) but neither on n nor on $f^\star$. This result stands if the penalty in IC is of the form :

$$\alpha(n) = (1+\theta)\sigma^2|S|,\ \theta > 0. \tag{4.5}$$

Now, as stressed in part 2.4, the global method has exponential complexity. Our aim in the sequel is to show that the descending method, that has polynomial complexity, also gives an oracle inequality of the type (4.4).

4.2.3. A family of nested deterministic supports

We define here a sequence of decreasing unknown supports $S^{\star(k)}$, $k = 0, \dots, d$ all with cardinality $d - k$. Firstly set $S^{\star(0)} = \{1, \dots, d\}$. Then, when $S^{\star(k)}$ is defined, set

$$S^{\star(k+1)} = \text{Argmin}\left(R^\star(S),\ S \subset S^{\star(k)},\ |S| = d - (k+1)\right) \tag{4.6}$$

where $R^\star$ is defined in (4.2).

We thus obtain a sequence of risks $R^\star(S^{\star(k)})$, $k = 0, \dots, d$. Each of those represents the minimum risk achieved by removing a single function in the previous support.

In the sequel, we are constrained to make the following assumption :

$$R^\star(S^{\star(k)}) \neq R^\star(S^{\star(k+1)}),\ k = 0, \dots, d-1. \tag{4.7}$$

This holds most of the time regarding the expression of $R^\star$ (4.2). It is indeed very unlikely that the potential increase of the first term in $R^\star$ is exactly compensated by the decrease σ^2/n of the second. However, if that happens, it suffices to add one or several points to the sample to deal with the problem. Assumption (4.7) ensures that the first index $1 \leq k^\star \leq d-1$ such that

$$R^\star(S^{\star(k^\star-1)}) > R^\star(S^{\star(k^\star)}) \text{ and } R^\star(S^{\star(k^\star+1)}) > R^\star(S^{\star(k^\star)}) \tag{4.8}$$

is correctly defined. In the case where the sequence $R^\star(S^{\star(k)})$, $k = 0, \dots, d$ is always decreasing, we set $k^\star = d$. In the case where $R^\star(S^{\star(1)}) > R^\star(S^{\star(0)})$, we set $k^\star = 0$.

This way, $S^{\star(k^\star)}$ is the first support that does not immediatly include a support achieving a smaller risk. The quantity $R^\star(S^{\star(k^\star)})$ is an oracle risk, not among any risks possible as in (4.3) with $\mathcal{F} = \mathcal{P}(\{1, \dots, d\})$, but among a smaller, nested, family of risks.

The deterministic family of supports $S^{\star(k)}$ (4.6) is related to the random family $S^{(k)}$ produced by the descending comparative method in part 2.3.5. Ideally, one would like the method to choose good supports and stop at the right step in the sense that

$$S^{(0)} = S^{\star(0)},\ S^{(1)} = S^{\star(1)}, \dots, S^{(k^\star)} = S^{\star(k^\star)}, \text{ and } k_f = k^\star.$$

Even though the forecoming theorem 4.2 deals with an oracle inequality, its proof also shows that this happens except on a event the probability of which decreases as $o(1/n)$.

4.2.4. The oracle inequality

Theorem 4.2 *Assume that (H8) holds. Consider an information criteria of the form (2.2) whose penalty term writes as*

$$\alpha(n) = (1+\theta)\sigma^2|S|$$

with $\theta > 0$. Using this criterion along with the descending comparative method described in part 2.3.5, one produces $\widetilde{f}_{S^{(k_f)}}$ as an estimation of the unknown function $f^\star$. The risk of such an estimation satisfies

$$\mathbb{E}\left[\left\|f^\star - \widetilde{f}_{S^{(k_f)}}\right\|_w^2\right] \leq C.R^\star(S^{\star(k^\star)}) + o\left(\frac{1}{n}\right) \tag{4.9}$$

where C is a constant depending on θ but neither on n nor on $f^\star$ and $R^\star(S^{\star(k^\star)})$ is the nested oracle risk defined in (4.8).

The proof of this theorem is given in appendix E

Appendix A: Expressions of the estimators $\widehat{f}_S$ and $\widetilde{f}_S$

This section introduces some notations useful in the various proofs of the theorems presented in the sequel.

A.1. Non-random objects

These objects are linked to the regression space F defined in part 1.1. It will be convenient to think of F as $\mathbb{R}^d$ via $F \to \mathbb{R}^d$, $f = \sum_{k=1}^d a_k f_k \mapsto a = (a_1, \ldots, a_d)^T$.

For $S \subset \{1, \ldots, d\}$, we consider $M_{w,S}$ the Gram matrix

$$M_{w,S}(j,k) = \langle f_j, f_k \rangle_w \,,\ j,k \in S \tag{A.1}$$

associated to F_S. Those matrices, indexed by w, are non-random objects. We shorten $M_{w,\{1,\ldots,d\}}$ to M_w. This way, the squared norm of a function writes simply as $\|f\|_w^2 = f^T M_w f$. We also denote by Π_S^F the orthogonal projector on F_S.

A.2. Random objects

These objects are linked to the observation space $\mathbb{R}^n$ defined in part 1.2. Let us use the Vandermonde-type $n \times d$ matrix V depending only of the x_i's :

$$V = \begin{pmatrix} f_1(x_1) & \ldots & f_d(x_1) \\ \vdots & \vdots & \vdots \\ f_1(x_n) & \ldots & f_d(x_n) \end{pmatrix}$$

This way, the passage from a function f to the vector $f(x)$ is a simple multiplication $f(x) = Vf$. Moreover, the Taylor expansion of l at any function f writes as

$$l(h) - l(f) = -2\left(M(f - f^\star) - V^T g\right)^T (h - f) + (h - f)^T M (h - f). \tag{A.2}$$

We set D_S the $d \times |S|$ matrix of zeros and ones such that $V_S := V D_S$ contains columns j of V for $j \in S$ only. We also set the Gram matrix

$$M_S = V_S^T V_S, \tag{A.3}$$

which is the random analog of the matrix $M_{w,S}$ defined in (A.1). We also shorten $M_{\{1,\dots,d\}} = V^T V$ to simply M.

The orthogonal projection of $\mathbb{R}^n$ onto E_S is denoted by Π_S^E, its matrix is :

$$\mathrm{Mat}(\Pi_S^E) = V_S M_S^{-1} V_S^T. \tag{A.4}$$

A.3. The estimator $\widehat{f}_S$

The likelihood (2.1) writes as

$$l(\widehat{f}_S) = d^2(y, E_S) = \frac{\mathrm{Gram}\left(\{y, f_j(x),\, j \in S\}\right)}{\det\left(M_S\right)}. \tag{A.5}$$

where $\mathrm{Gram}(u_1, \dots, u_k)$ denotes the Gram determinant of those vectors. The orthogonal projection Π_S^E (A.4) gives the following expression for $\widehat{f}_S$ realizing the minimum in (2.1) :

$$\widehat{f}_S = D_S M_S^{-1} V_S^T (Vf + g) \in F_S. \tag{A.6}$$

A.4. The estimator $\widetilde{f}_S$

It is important to note that, contrarily to the fixed design setting, it may happen that $\left\|\widehat{f}_S\right\|_w^2$ is not integrable, thus preventing us from calculating risks. To handle this we use at some points in the paper a truncated estimator

$$\widetilde{f}_S = \widehat{f}_S . \mathbb{1}_{\{\|nM_S^{-1}\| < C\}} \tag{A.7}$$

where $\|.\|$ is any norm on matrices and C is a constant satisfying $C > \|M_{w,S}^{-1}\|$.

A.5. An application of the law of large numbers

In our random design setting, the law of large numbers gives a simple asymptotic connection between the observation space (random objects) and the regression space (non-random objects) by relating Gram matrices (A.3) and (A.1) as follows :

$$\frac{1}{n} M_S \to M_{w,S},\ a.s.\ S \subset \{1, \dots, d\}. \tag{A.8}$$

Appendix B: Some technical lemmas

Lemma B.1 *Assume that (H8) holds.*

(i) For any constant C greater than $\|M_{w,S}^{-1}\|$ we have :

$$\mathbb{E}\left[\mathbb{1}_{\{\|nM_S^{-1}\|<C\}^c}\right] = o\left(\frac{1}{n}\right). \tag{B.1}$$

(ii) For any $\varepsilon > 0$:

$$\mathbb{P}\left(\left|\frac{1}{n}l(\widehat{f}_S) - \sigma^2 - d_w^2(f^\star, F_S)\right| > \varepsilon\right) = o\left(\frac{1}{n}\right). \tag{B.2}$$

(iii) Using any of the comparative methods described in part 2.3 and assuming that $\alpha(n) = o(n)$, we have, for any $j \in S^\star$:

$$\mathbb{P}(j \notin \widehat{S}) = o\left(\frac{1}{n}\right). \tag{B.3}$$

Proof : For any $i = 1, \ldots, n$ and $j, k \in \{1, \ldots, d\}$, set $Y_i = f_j(X_i)f_k(X_i)$. From (H8), we get $Y_i \in L^4$. Denote by $S_n \in L^4$ the sum $Y_1 + \cdots + Y_n$. The law of large numbers (A.8) amounts to write $S_n/n \to \langle f_j, f_k\rangle_w$. Choose $\varepsilon > 0$ and use Markov inequality to write

$$\mathbb{P}\left(\left|\frac{S_n}{n} - \langle f_j, f_k\rangle_w\right| > \varepsilon\right) \leq \frac{1}{n^4\varepsilon^4}\mathbb{E}\left[\left|S_n - n\langle f_j, f_k\rangle_w\right|^4\right].$$

Moreover, setting $Z := Y_1 - \langle f_j, f_k\rangle_w$, we have

$$\mathbb{E}\left[\left|S_n - n\langle f_j, f_k\rangle_w\right|^4\right] = n\mathbb{E}\left[Z^4\right] + 6n(n-1)\mathbb{E}\left[Z^2\right]^2.$$

Consequently, we may write $\mathbb{P}\left(\left|S_n/n - \langle f_j, f_k\rangle_w\right| > \varepsilon\right) = o(1/n)$ as well as

$$\mathbb{P}\left(\|M_s/n - M_{w,S}\| > \varepsilon\right) = o\left(\frac{1}{n}\right) \tag{B.4}$$

Now for the proof of (i), since $C > \|M_{w,S}^{-1}\|$, there exists an $\varepsilon_C > 0$ such that the event $\{\|nM_S^{-1}\| < C\}^c$ is included in $\{\|M_s/n - M_{w,S}\| > \varepsilon_C\}$. Control (B.1) follows from (B.4).

In order to proove (ii), first note that, by independence between X and G, the entry (1,1) of the first Gram matrix in (A.5) satisfies

$$\frac{1}{n}\langle f^\star(x) + g, f^\star(x) + g\rangle_n \to \langle f^\star, f^\star\rangle_w + \sigma^2 \ a.s. \tag{B.5}$$

Moreover, any other entry converges to the corresponding one of the Gram matrix of functions $f^\star, f_j, j \in S$. Consequently we get :

$$\frac{1}{n} l(\widehat{f}_S) \to \sigma^2 + d_w^2(f^\star, F_S) \text{ a.s.} \tag{B.6}$$

From (H8), each variable for whom the law of large numbers has been used here have a moment of order 4. Arguments similar to those leading to (B.4) give (B.2).

We now proove (*iii*). Choose $j \in S^\star$ so that $D^2 := d_w^2(f^\star, F_{-j}) > 0$. Assume that $\alpha(n) = o(n)$ and consider n large enough to make $0 \leq \alpha(n)/n < D^2/2$. Assume we use the comparative method (2.4). We may write :

$$\begin{aligned}
\mathbb{P}(j \notin \widehat{S}) &= \mathbb{P}\left(n^{-1}\left(\mathrm{IC}_{\mathrm{ref}} - \mathrm{IC}(-j)\right) \geq 0\right) \\
&\leq \mathbb{P}\left(n^{-1} l(\widehat{f}_{\{1,\dots,d\}}) - \left(n^{-1} l(\widehat{f}_{-j}) - D^2\right) > D^2/2\right) \\
&\leq \mathbb{P}\left(\left|n^{-1} l(\widehat{f}_{\{1,\dots,d\}}) - \sigma^2\right| > D^2/4\right) \\
&\quad + \mathbb{P}\left(\left|n^{-1} l(\widehat{f}_{-j}) - \sigma^2 - D^2\right| > D^2/4\right).
\end{aligned}$$

Then, (B.3) follows from (B.2). We handle the case of the other methods the same way. □

Lemma B.2 *Under (H8) and in the setting used in part 4.1.1, equation (4.1) holds.*

Proof : For any function f and any support S, $D_S f D_S^T$ sets to 0 all the components of f along the f_j's, $j \in S$. Consequently, $D_{S^\star} f^\star D_{S^\star}^T = f^\star$. Now, from (A.6) and (A.7), we may write :

$$\begin{aligned}
\widetilde{f}_S^\star &= \mathbb{1}_{\{\|nM_{S^\star}^{-1}\| < C\}} D_S^\star M_{S^\star}^{-1} V_{S^\star}^T \left(V D_{S^\star} f^\star D_{S^\star}^T + g\right) \\
&= \mathbb{1}_{\{\|nM_{S^\star}^{-1}\| < C\}} \left(f^\star + D_S^\star M_{S^\star}^{-1} V_{S^\star}^T g\right)
\end{aligned}$$

Also recall, that, for any function f, the norm $\|f\|_w^2$ writes as $f^T M_w f$. Note that the event $\{\|M_{S^\star}^{-1}\| < C\}$ appearing in (A.7) is independent of the noise g. Therefore, raw calculations give

$$\begin{aligned}
\mathbb{E}\left[\left\|\widetilde{f}_{S^\star} - f^\star\right\|_w^2\right] &= \|f^\star\|_w^2 \, \mathbb{E}\left[\mathbb{1}_{\{\|nM_{S^\star}^{-1}\| < C\}^c}\right] \\
&\quad + \sigma^2 \mathbb{E}\left[\mathbb{1}_{\{\|nM_{S^\star}^{-1}\| < C\}} \mathrm{Tr}(M_{w,S^\star} M_{S^\star}^{-1})\right].
\end{aligned}$$

The first term is handled by (B.1). For the second, recall that $M_{w,S^\star}$ is a $|S^\star| \times |S^\star|$ matrices towards which M_S/n converges a.s. ; finally, the indicator function gives a dominated convergence and allows to write (4.1).

□

Lemma B.3 *Assume that (H8) holds. We have*

$$\mathbb{E}\left[\left\|\widetilde{f}_S\right\|_w^{2p}\right] < \infty \text{ for any } p \leq 4. \tag{B.7}$$

Moreover, for any event A and any $p, q > 1$ such that $p \leq 4$, $1/p + 1/q = 1$, Hölder's inequality gives

$$\mathbb{E}\left[\left\|\widetilde{f}_S\right\|_w^2 \mathbb{1}_A\right] \leq \mathbb{E}\left[\left\|\widetilde{f}_S\right\|_w^{2p}\right]^{1/p} \mathbb{P}(A)^{1/q}. \tag{B.8}$$

Proof : Recall the definition of $\widetilde{f}_S$ in (A.7) and use the law of large numbers (A.8) to write $\mathbb{1}_{\{\|nM_S^{-1}\|<C\}} \to 1$ a.s. From (A.6), the squared norm of $\widetilde{f}_S$ writes as

$$\left\|\widetilde{f}_S\right\|_w^2 = \mathbb{1}_{\{\|nM_S^{-1}\|<C\}} \left(D_S M_S^{-1} V_S^T y\right)^T M_{w,S} \left(D_S M_S^{-1} V_S^T y\right)$$

with $y = Vf^\star + g$. In that expression, matrices D_S and $M_{w,S}$ are deterministic and the indicator function ensures the boundedness of M_S^{-1}. The integrability we have requested on the f_j's in (H8) along with the fact that G is Gaussian ensure that give (B.7). □

Appendix C: Proof of theorem 3.1

C.1. Proof of assertion (i)

Let us split it into two parts.

⋆ **First part** : the case $j \in S^\star$. Here, use (B.6) to write :

$$\frac{1}{n}\left(\mathrm{IC}_{\mathrm{ref}} - \mathrm{IC}(-j) - \alpha(n)\right) = -d_w^2(f^\star, F_{-j}) + o(1) \text{ a.s.}, \tag{C.1}$$

where that latter distance $D^2 := d_w^2(f^\star, F_{-j})$ does not vanish since $j \in S^\star$.

Again because $j \in S^\star$, one would like to select j as a part of $\widehat{S}$; in other terms one would like $\mathrm{IC}_{\mathrm{ref}} - \mathrm{IC}(-j)$ to be nonpositive. The fact that $\alpha(n) = o(n)$ from assumption (H_α) ensures that this happens a.s. for n large enough.

⋆⋆ **Second part** : the case $j \notin S^\star$. Recall the definitions of the matrix M_w in (A.1), the random matrix M in (A.3) and the law of large numbers (A.8) linking them :

$$M/n \longrightarrow M_w \text{ a.s.},$$

that latter limit being invertible and positive definite. Consequently each entry of nM^{-1} is bounded a.s. at least above a certain rank, which we denote by

$$M^{-1} = O\left(\frac{1}{n}\right) \text{ a.s.} \tag{C.2}$$

More precisely, if m_j is the j-th diagonal coefficient of M^{-1}, then $n.m_j$ converges a.s. to the j-th diagonal coefficient of M_w^{-1} which is positive and

$$\frac{1}{m_j} = O(n) \text{ a.s.} \tag{C.3}$$

Let us call $\widehat{f}$ and $\widehat{f}_{-j}$ the functions in F and F_{-j} respectively maximizing the likelihood on those spaces as in (2.1). We are interested in the difference of likelihood which, from (A.2), writes as :

$$l\big(\widehat{f}_{-j}\big) - l\big(\widehat{f}\big) = \big(\widehat{f}_{-j} - \widehat{f}\big)^T M \big(\widehat{f}_{-j} - \widehat{f}\big) \tag{C.4}$$

since $\widehat{f}$ satisfies grad $l\big(\widehat{f}\big) = 0$.

Now for $\widehat{f}_{-j}$, it satisfies

$$\Big(\text{grad } l(\widehat{f}_{-j})\Big)^T = M\widehat{f}_{-j} - M\widehat{f} = (0, \dots, 0, \lambda_j, 0, \dots, 0)^T \tag{C.5}$$

where λ_j is a Lagrange coefficient set at the j-th place in the latter vector. Since the j-th coefficient of $\widehat{f}_{-j}$, must vanish, that Lagrange coefficient necessarily satisfies

$$\lambda_j = -\frac{\widehat{f}_j}{m_j} \tag{C.6}$$

where $\widehat{f}_j$ is the j-th coefficient of $\widehat{f}$.

Plugging (C.3), (C.5) and (C.6) in expansion (C.4) gives

$$l\big(\widehat{f}^j\big) - l\big(\widehat{f}\big) = \Big(\widehat{f}_j\Big)^2 O(n) \tag{C.7}$$

Now note that $\widehat{f}$ is defined by $\widehat{f} = f^\star + M^{-1}V^T g$ which yields $\mathbb{E}[\widehat{f}] = f^\star$ keeping in mind that matrices V and M only depend on X which is independent of G. Consequently, in our case $j \notin S^\star$, we get

$$\widehat{f}_j = 0 + \big(M^{-1}V^T g\big)_j .$$

Any entry in the vector $V^T g$ is of the form $\sum_{i=1}^n f_k(x_i)g_i$ which is the sum of n independent variables with mean 0, finite variance, and thus is $O(\sqrt{n \ln\ln n})$ a.s. by the law of iterated logarithm. The order of M^{-1} given in (C.2) makes $\widehat{f}_j = O\left(\sqrt{\frac{\ln\ln n}{n}}\right)$ a.s.

Plugging in (C.7), we finally get

$$l\big(\widehat{f}^j\big) - l\big(\widehat{f}\big) = O(\ln\ln n) \text{ a.s.}$$

Now in our case $j \notin S^\star$, one would like to reject j ; in other terms, on would like $\text{IC}_{\text{ref}} - \text{IC}(-j)$ to be nonnegative. Write

$$\text{IC}_{\text{ref}} - \text{IC}(-j) = l\big(\widehat{f}\big) - l\big(\widehat{f}^j\big) + \alpha(n) = \alpha(n) + O\left(\ln\ln n\right) \text{ a.s.}$$

so that

$$\frac{1}{\ln\ln n}\left(\mathrm{IC}_{\mathrm{ref}} - \mathrm{IC}(-j) - \alpha(n)\right) = O(1) \text{ a.s.} \tag{C.8}$$

From assumption (H_α), we have $\ln\ln n = o(\alpha(n))$. Consequently, a.s. and for n large enough, $\mathrm{IC}_{\mathrm{ref}} - \mathrm{IC}(-j) > 0$.

C.2. Proof of assertion (ii)

We work here with the reverse comparative method (2.5). The function achieving the maximum likelihood for $\mathrm{IC}_{\mathrm{ref}}$ obviously vanishes. Denote by $\widehat{f}_{\{j\}}$ the function with support $\{j\}$ achieving it for $\mathrm{IC}(\{j\})$. The following control for the difference of the likelihoods holds :

$$l(0) - l(\widehat{f}_{\{j\}}) = n\left\langle f_j, f^\star\right\rangle_w^2 + \left\langle f_j, f^\star\right\rangle_w O\left(\sqrt{n\ln\ln n}\right) + O(\ln\ln n). \tag{C.9}$$

Indeed, the function $\widehat{f}_{\{j\}}$ has the following component along f_j :

$$\widehat{f}_{\{j\},j} = \frac{\langle f_j(x), y\rangle_n}{\langle f_j(x), f_j(x)\rangle_n}$$

Now the difference of the likelihoods is easier to compute than with the regular comparative method :

$$\begin{aligned} l(0) - l(\widehat{f}_{\{j\}}) &= \sum_{i=1}^n \left(y_i^2 - \left(y_i - \frac{\langle f_j(x), y\rangle_n}{\langle f_j(x), f_j(x)\rangle_n} f_j(x_i)\right)^2\right) \\ &= \frac{\langle f_j(x), y\rangle_n^2}{\langle f_j(x), f_j(x)\rangle_n} \end{aligned}$$

In order to control the asymptotics, use the law of that large numbers to write $\langle f_j(x), f_j(x)\rangle_n^{-1} = O(1/n)$ a.s. Moreover, two applications of the law of iterated logarithm give, a.s. :

$$\langle f_j(x), y\rangle_n = \langle f_j(x), f^\star(x)\rangle_n + \langle f_j(x), g\rangle_n = n\left\langle f_j, f^\star\right\rangle_w + O\left(\sqrt{n\ln\ln n}\right).$$

Equation (C.9) follows.

C.2.1. Orthonormal case and reversed comparative method

As in theorem (3.1) (ii) to be shown, we assume here that the basis $(f_j)_j$ is orthonormal relatively to the scalar product (1.1). Note that orthogonal would be enough.

Recall that $f^\star = \sum_j a_j^\star f_j$, consequently $a_j^\star = \left\langle f_j, f^\star\right\rangle_w = 0 \Leftrightarrow j \notin S^\star$ and (C.9) yields

$$\mathrm{IC}_{\mathrm{ref}} - \mathrm{IC}(j) + \alpha(n) = n{a_j^\star}^2 + a_j^\star O\left(\sqrt{n\ln\ln n}\right) + O(\ln\ln n) \text{ a.s.}$$

This formula is to be related to equations (C.1) and (C.8) concerning the regular comparative method to see that, in the orthonormal case, the reversed method behaves asymptotically as the comparative method. Assertion (ii) follows from arguments similar to those leading to (i).

C.2.2. Non-orthonormal case and reversed comparative method.

This case is not dealt with in theorem (3.1). However, we give here some comments about it.

Suppose we are in a case where $\langle f_j, f^\star \rangle_w = \sum_{k \in S^\star} a_k^\star \langle f_j, f_k \rangle_w$ vanishes for no index j. This happens most of the times if the basis $(f_k, k \in \{1, \dots, d\})$ is not orthonormal even though $j \notin S^\star$. Then formula (C.9) gives

$$\frac{1}{n}\left(\mathrm{IC}_{\mathrm{ref}} - \mathrm{IC}(\{j\}) + \alpha(n)\right) = \langle f_j, f^\star \rangle_w^2 + o(1). \tag{C.10}$$

Now assume $\alpha(n) = o(n)$, then $\mathrm{IC}_{\mathrm{ref}} - \mathrm{IC}(j) > 0$ a.s. above a certain rank and this for all j.The condition $\alpha(n) = o(n)$ which ensured we kept good indices with the regular comparative method (assertion (i)) turns out to make us keep every indices.

One should thus think about taking a penalty a bit larger. However, formula (C.10) also implies that if

$$\alpha(n) = kn \text{ where } C > \max\left\{ \langle f_j, f^\star \rangle_w^2 , \, j \in \{1, \dots, d\} \right\}$$

then $\mathrm{IC}_{\mathrm{ref}} - \mathrm{IC}(j) < 0$ a.s. above a certain rank and this for all j. This means we reject every indices $j \in \{1, \dots, d\}$.

Therefore, in order to use the reversed comparative method, one should firstly choose a penalty of order not smaller than n and not greater than kn to ensure that the criterion does not accept or reject systematically every indices. A good order for the penalty would be in the few place left :

$$\alpha(n) = \alpha.n \text{ with } \min\left\{ \langle f_j, f^\star \rangle_w^2 , \, j \in S^\star \right\} > \alpha > \max\left\{ \langle f_j, f^\star \rangle_w^2 , \, j \notin S^\star \right\}.$$

However, that α might not exist if its bounds are not ordered the correct way. In any case, it is unavailable to the user and thus not of a practical use.

C.3. Proof of assertion (iii)

We work here with the adapted reversed comparative method (2.6). Choose $j \in \{1, \dots, d\}$. Replace the basis $(f_1, \dots, f_d)$ with $\{f_k, k \neq j\} \cup \{f_j^N\}$ and decompose $f^\star$ as

$$f^\star = \sum_{k \neq j} a_k^{\star N} f_k + a_j^{\star N} f_j^N .$$

We get a formula similar to (C.9) :

$$\mathrm{IC}_{\mathrm{ref}} - \mathrm{IC}(j) + \alpha(n) = n\left(a_j^{\star N}\right)^2 + a_j^{\star N} O\left(\sqrt{n \ln\ln n}\right) + O(\ln\ln n).$$

Note that $a_j^{\star N} \neq 0 \Leftrightarrow j \in S^\star$ and apply arguments similar to those leading to (i) to obtain (iii).

C.4. Proof of assertion (iv)

We work here with the descending method described in part 2.3.5.

Applying (i) we obtain that, with probability 1 and for n large enough, the indices $C^{(1)}$ selected in the first step (2.7) are exactly $S^{\star c}$. Therefore, $J^{(1)} \notin S^\star$ and $S^\star \subset S^{(1)}$. Since the dimension d of the regression space is not allowed to depend on n, that process may be iterated enough times to eliminate all indices out of $S^\star$. Another application of (i) ensures that the following choice of $C^{(k)}$ will lead to the empty set. This concludes the proof

Appendix D: Proof of theorem 4.1

Let us consider the families of supports :

$$\mathcal{F}_1 = \{S \subset \{1,\dots,d\} \,|\, S^\star \nsubseteq S\} \text{ and } \mathcal{F}_2 = \{S \subset \{1,\dots,d\} \,|\, S^\star \subseteq S\}. \quad \text{(D.1)}$$

Theorem 3.1 as well as control (B.3), ensure that with any of the comparative method used here, we get

$$\begin{cases} \widehat{S} \to S^\star & \text{a.s.} \\ \mathbb{P}(\widehat{S} = S) \to 0 & \text{for any} \quad S \in \mathcal{F}_2 \setminus \{S^\star\} \\ \mathbb{P}(\widehat{S} = S) = o\left(\frac{1}{n}\right) & \text{for any} \quad S \in \mathcal{F}_1 \end{cases} \quad \text{(D.2)}$$

where c is a positive constant.

Our estimation procedure of $f^\star$ by $\widetilde{f} = \widetilde{f}_{\widehat{S}}$ has a risk $R(n)$ given by

$$R(n) = \mathbb{E}\left[\left\|f^\star - \widetilde{f}\right\|_w^2\right] = \sum_{S \subset \{1,\dots,d\}} \mathbb{E}\left[\left\|f^\star - \widetilde{f}_S\right\|_w^2 \mathbb{1}_{\widehat{S}=S}\right] =: \sum_{S \subset \{1,\dots,d\}} R(n,S). \quad \text{(D.3)}$$

Now distinguish between two cases.

$\star$ The case $S \in \mathcal{F}_2$. Computations similar to the ones exposed in the proof of lemma B.2 show that the quantity $R(n,S)$ equals

$$\begin{aligned} R(n,S) &= \|f^\star\|_w^2 \,\mathbb{E}\left[\mathbb{1}_{\{\|nM_S^{-1}\|<C\}^c \cap \{\widehat{S}=S\}}\right] \\ &\quad + \mathbb{E}\left[G^T A_S G \,\mathbb{1}_{\{\|nM_S^{-1}\|<C\} \cap \{\widehat{S}=S\}}\right]. \end{aligned}$$

where $A_S = V_S^T M_S^{-1} M_{w,S} M_S^{-1} V_S$. Let us denote by Z_n the new variable

$$\begin{aligned} Z_n &= nG^T A_S G \, \mathbb{1}_{\{\|nM_S^{-1}\|<C\}} \\ &= \frac{1}{n} \sum_{i,j=1}^{n} G_i G_j \left(V_S^T n M_S^{-1} M_{w,S} n M_S^{-1} V_S \right)_{i,j} \mathbb{1}_{\{\|nM_S^{-1}\|<C\}}. \end{aligned} \quad \text{(D.4)}$$

Note that, via the law of large numbers (A.8), independence between X and G, and dominated convergence given by the indicator function, we get

$$\begin{cases} \mathbb{E} Z_n \to \sigma^2 |S|, \\ \\ \mathbb{E} Z_n^2 = O(1). \end{cases} \quad \text{(D.5)}$$

Let us begin by $S = S^\star$. Then (B.1), (D.2) and (D.5) give :

$$R(n, S^\star) \sim \frac{\sigma^2 |S^\star|}{n}. \quad \text{(D.6)}$$

Now for $S \in \mathcal{F}_2 \setminus \{S^\star\}$ we use Schwarz's inequality to write

$$\mathbb{E}\left[Z_n \mathbb{1}_{\{\widehat{S}=S\}} \right] \leq \left(\mathbb{E} \, Z_n^2 \, \mathbb{P}(\widehat{S} = S) \right)^{1/2} \to 0$$

because of (D.5) and (D.2). Consequently,

$$nR(n, S) \to 0, \text{ for any } S \in \mathcal{F}_2 \setminus \{S^\star\}. \quad \text{(D.7)}$$

$\star\star$ The case $S \in \mathcal{F}_1$. Here we simply write

$$\begin{aligned} R(n, S) &= \mathbb{E}\left[\left\| f^\star - \widetilde{f}_S \right\|_w^2 \mathbb{1}_{\widehat{S}=S} \right] \\ &\leq 2 \left\| f^\star \right\|_w^2 \mathbb{P}(\widehat{S} = S) + 2\mathbb{E}\left[\left\| \widetilde{f}_S \right\|_w^{2p} \right]^{1/p} \mathbb{P}(\widehat{S} = S)^{1/q} \end{aligned}$$

where p, q are chosen as in (B.8). Recall (B.7) and (D.2) to obtain

$$nR(n, S) \to 0, \text{ for any } S \in \mathcal{F}_1. \quad \text{(D.8)}$$

Plugging (D.3), (D.6), (D.7) and (D.8) together gives theorem 4.1.

Appendix E: Proof of theorem 4.2

Define the decreasing sequence of events $(A_k)_{k=1,\dots,d}$ by

$$A_k = \left\{ S^{(k-1)} = S^{\star(k-1)}, \dots, S^{(0)} = S^{\star(0)} \right\}. \quad \text{(E.1)}$$

Note that A_k is implicitely included in the event $\{k_f \geq k-1\}$. Let us start with a lemma

Lemma E.1 *Assume that (H8) holds. For any $k = 1, \dots, k^\star$, we have*

$$\begin{aligned} \mathbb{P}\left(S^{(k)} \neq S^{\star(k)} \,|\, A_k\right) &= o\left(\frac{1}{n}\right), \\ \mathbb{P}\left(k_f = k^\star - 1 \,|\, A_{k^\star}\right) &= o\left(\frac{1}{n}\right), \\ \mathbb{P}\left(k_f > k^\star \,|\, A_{k^\star}\right) &= o\left(\frac{1}{n}\right). \end{aligned} \tag{E.2}$$

Proof : Choose $1 \leq k \leq k^\star$. We are interested in the probability that the descending comparative method selects good supports up to step $k-1$ and fails to select $S^{\star(k)}$ at step k : $\mathbb{P}\left(S^{(k)} \neq S^{\star(k)}\,,\, A_k\right)$. Write

$$\begin{aligned} \mathbb{P}\left(S^{(k)} \neq S^{\star(k)}\,,\, A_k\right) &\leq \sum_S \mathbb{P}\left(\frac{1}{n}\mathrm{IC}(S) - \sigma^2 \leq \frac{1}{n}\mathrm{IC}(S^{\star(k)}) - \sigma^2\,,\, A_k\right) \\ &\leq \sum_S \mathbb{P}\left(\left|\frac{1}{n}\mathrm{IC}(S) - \sigma^2 - R^\star(S)\right| > \varepsilon_k\,,\, A_k\right) + \\ &\quad \sum_S \mathbb{P}\left(\left|\frac{1}{n}\mathrm{IC}(S^{\star(k)}) - \sigma^2 - R^\star(S^{\star(k)})\right| > \varepsilon_k\,,\, A_k\right) \end{aligned} \tag{E.3}$$

where the sums are extended to all supports $S \subset S^{\star(k-1)}$ with cardinal $|S| = d - k$ except $S^{\star(k)}$ and

$$\varepsilon_k = \frac{1}{2}\min_S (R^\star(S) - R^\star(S^{\star(k)})) > 0,$$

the minimun being taken among the same set of supports. Let us denote by $\varepsilon > 0$ the smallest of those ε_k's, $k = 1, \dots, k^\star$. Because of the form of the penalty term (4.5), expressions appearing in the probabilities of equation (E.3) simplify to

$$\frac{1}{n}\mathrm{IC}(S) - \sigma^2 - R^\star(S) = \frac{1}{n}d^2(y, E_S) - \sigma^2 - d_w^2(f^\star, F_S) + \frac{\theta\sigma^2|S|}{n}.$$

It remains to choose n large enough to make $\theta\sigma^2|S|/n < \varepsilon/4$ and apply control (B.2) to have

$$\mathbb{P}\left(S^{(k)} \neq S^{\star(k)}\,,\, A_k\right) = o\left(\frac{1}{n}\right),\ k = 1, \dots, k^\star. \tag{E.4}$$

Now let us stress that, conditionnaly to the distribution X of the abscisses, the information criterion (2.2) has a distribution of the type non-central χ^2 ; the degrees of freedom as well as the mean parameter being related to which support S the IC is calculated on. Now, an event of the form A_k (E.1) may be expressed in term of some $\mathrm{IC}(S)$ constrained to belong to some, random but non-empty, intervals of $\mathbb{R}$. Therefore, via the positivity of the densities of

the joint laws of those IC, we get the positivity of $\mathbb{P}(A_k)$ for any k. Moreover, iterating control (E.4) also shows that $\mathbb{P}(A_k^c)$ behaves as $o\left(1/n\right)$ for some $a > 0$ and $k = 1, \ldots, k^\star$. Finally we get the existence of a positive constant c satisfying for all n :

$$\mathbb{P}(A_k) > c > 0,\ k = 1, \ldots, k^\star. \tag{E.5}$$

We are now interested in

$$\mathbb{P}\left(k_f = k^\star - 1\,,\, A_{k^\star}\right).$$

This is the probability that the method goes well up to step $k^\star - 1$ but stops here. Write :

$$\mathbb{P}\left(k_f = k^\star - 1\,,\, A_{k^\star}\right) = \mathbb{P}\left(\bigcap_S \left(\frac{1}{n}\mathrm{IC}(S) - \sigma^2 \geq \frac{1}{n}\mathrm{IC}(S^{\star(k^\star-1)}) - \sigma^2\right)\,,\, A_{k^\star}\right)$$

where the intersection is extended to all supports S of cardinal $d - k^\star$ included in $S^{\star(k^\star-1)}$. In particular, choosing $S = S^{\star(k^\star)}$:

$$\begin{aligned}
\mathbb{P}\left(k_f = k^\star - 1\,,\, A_{k^\star}\right) &\leq \mathbb{P}\left(\frac{1}{n}\mathrm{IC}(S^{\star(k^\star)}) - \sigma^2 \geq \frac{1}{n}\mathrm{IC}(S^{\star(k^\star-1)}) - \sigma^2\,,\, A_{k^\star}\right) \\
&\leq \mathbb{P}\left(\left|\frac{1}{n}\mathrm{IC}(S^{\star(k^\star-1)}) - \sigma^2 - R^\star(S^{\star(k^\star-1)})\right| > \varepsilon\,,\, A_{k^\star}\right) \\
&\quad + \mathbb{P}\left(\left|\frac{1}{n}\mathrm{IC}(S^{\star(k^\star)}) - \sigma^2 - R^\star(S^{\star(k^\star)})\right| > \varepsilon\,,\, A_{k^\star}\right)
\end{aligned}$$

where

$$\varepsilon = \frac{1}{2}\left(R^\star(S^{\star(k^\star-1)}) - R^\star(S^{\star(k^\star)})\right) > 0.$$

Again because of (4.5), expressions in those probabilities reduce to

$$\frac{1}{n}\mathrm{IC}(T) - \sigma^2 - R^\star(T) = \frac{1}{n}d^2(y, E_T) - \sigma^2 - d_w^2(f^\star, F_T) + \frac{\theta\sigma^2|T|}{n}.$$

Choose n large enough to make $\theta\sigma^2|T|/n < \varepsilon/4$ and apply control (B.2) to get

$$\mathbb{P}\left(k_f = k^\star - 1\,,\, A_{k^\star}\right) = o\left(\frac{1}{n}\right). \tag{E.6}$$

The last probability we need to control is the following :

$$\mathbb{P}\left(k_f > k^\star\,,\, A_{k^\star}\right) \leq \mathbb{P}\left(k_f > k^\star\,,\, S^{(k^\star)} \neq S^{\star(k^\star)}\,,\, A_{k^\star}\right) + \mathbb{P}\left(k_f > k^\star\,,\, A_{k^\star+1}\right).$$

The first term has already been dealt with when we obtained control (E.4). For the second one, it is handled by arguments similar to the ones we developped to justify (E.6). Consequently we get

$$\mathbb{P}\left(k_f > k^\star\,,\, A_{k^\star}\right) = o\left(\frac{1}{n}\right). \tag{E.7}$$

Put (E.4), (E.6), (E.7) together with (E.5) to end the proofs of the lemma. □

Let us now finish the proof of theorem 4.2. Recall that the descending comparative method produces $\widehat{f}_{S^{(k_f)}}$ as an estimation of $f^\star$. The loss is measured by $\left\| f^\star - \widetilde{f}_{S^{(k_f)}} \right\|_w^2$ which we shorten to L.

We condition by the event $\{k_f = k^\star\} \cap A_{k^\star}$ which means that the method has chosen good supports up to step $k^\star - 1$ and will stop at the good step $k^\star$. The comparative descending method (part 2.3.5) used here ensures that we are going to choose a support by minimization of our criterion among the family of supports of cardinal $d - k^\star$ included in $S^{\star(k^\star-1)}$. The oracle risk associated with this family is precisely $R^\star(S^{\star(k^\star)})$.

It is time to apply Baraud's result as described in section 4.2.2. Precisely, we use equation (15) following theorem 1.1 in [3] to get

$$\mathbb{E}\left[L \middle| k_f = k^\star,\, A_{k^\star}\right] \leq C.R^\star(S^{\star(k^\star)}),$$

where C depends on θ but neither on n nor on $f^\star$. We need to remove the conditionning in order to obtain the oracle inequality (4.9). Write

$$\mathbb{E}\left[L \middle| A_{k^\star}\right] \quad \leq \quad \mathbb{E}\left[L \middle| k_f = k^\star,\, A_{k^\star}\right] + \mathbb{E}\left[L\mathbb{1}_{k_f = k^\star - 1} \middle| A_{k^\star}\right] + \mathbb{E}\left[L\mathbb{1}_{k_f > k^\star} \middle| A_{k^\star}\right]$$

to handle the event $k_f = k^\star$. Moreover :

$$\mathbb{E}\left[L \middle| A_{k^\star - 1}\right] \quad \leq \quad \mathbb{E}\left[L \middle| A_{k^\star}\right] + \mathbb{E}\left[L\mathbb{1}_{S^{(k^\star)} \neq S^{\star(k^\star)}} \middle| A_{k^\star - 1}\right]$$

handles the event $S^{(k^\star-1)} = S^{\star(k^\star-1)}$ in $A_{k^\star}$. Iterating that latter argument we get

$$\begin{aligned}\mathbb{E}\left[L|A_1\right] = \mathbb{E}\left[L\right] \quad \leq \quad & C.R^\star(S^{\star(k^\star)}) + \\ & \mathbb{E}\left[L\mathbb{1}_{k_f = k^\star - 1} \middle| A_{k^\star}\right] + \mathbb{E}\left[L\mathbb{1}_{k_f > k^\star} \middle| A_{k^\star}\right] + \\ & \sum_{k=1}^{k^\star - 1} \mathbb{E}\left[L\mathbb{1}_{S^{(k)} \neq S^{\star(k)}} \middle| A_k\right].\end{aligned}$$

It suffices to apply Hölder's inequality (B.8) along with controls (E.2) to conclude the proof.

References

[1] H. Akaike. A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control*, 19:716–723, 1974.

[2] Y. Baraud. Model selection for regression on a fixed design. *Probab. Theory Related Fields*, 117(4):467–493, 2000.

[3] Y. Baraud. Model selection for regression on a random design. *ESAIM Probab. Statist.*, 6:127–146 (electronic), 2002.

[4] A. Barron, L. Birgé, and P. Massart. Risk bounds for model selection via penalization. *Probab. Theory Related Fields*, 113(3):301–413, 1999.

[5] L. Birgé. Model selection for Gaussian regression with random design. *Bernoulli*, 10(6):1039–1051, 2004.

[6] L. Birgé. Statistical estimation with model selection. *Indag. Math. (N.S.)*, 17(4):497–537, 2006.

[7] G. Castellan. Sélection d'histogrammes à l'aide d'un critère de type Akaike. *C. R. Acad. Sci. Paris Sér. I Math.*, 330(8):729–732, 2000.

[8] P. D. Grunwald, In Jae Myung, and M. A. Pitt. *Advances in Minimum Description Length: Theory and Applications (Neural Information Processing)*. The MIT Press, 2005.

[9] 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.

[10] R. Nishii. Asymptotic properties of criteria for selection of variables in multiple regression. *Ann. Statist.*, 12(2):758–765, 1984.

[11] R. Nishii. Maximum likelihood principle and model selection when the true model is unspecified. *J. Multivariate Anal.*, 27(2):392–403, 1988.

[12] C. R. Rao and Y. Wu. On model selection. In *Model selection*, volume 38 of *IMS Lecture Notes Monogr. Ser.*, pages 1–64.

[13] C. Radhakrishna Rao and Yue Hua Wu. A strongly consistent procedure for model selection in a regression problem. *Biometrika*, 76(2):369–374, 1989.

[14] J. Rissanen. Modeling by the shortest data description. *Automatica*, 14:465–471, 1978.

[15] J. Rissanen. Stochastic complexity and modeling. *Ann. Statist.*, 14(3):1080–1100, 1986.

[16] J. Rissanen. *Stochastic complexity in statistical inquiry*, volume 15 of *World Scientific Series in Computer Science*. World Scientific Publishing Co. Inc., Teaneck, NJ, 1989.

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