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

Guilhem Coq

▶ **To cite this version:**

Guilhem Coq. Alternative utilizations of Information Criteria for Gaussian regression on a random design.. 2008. hal-00367551v1

HAL Id: hal-00367551

https://hal.science/hal-00367551v1

Preprint submitted on 11 Mar 2009 (v1), last revised 4 Dec 2009 (v2)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

ESAIM: Probability and Statistics Will be set by the publisher
URL: http://www.emath.fr/ps/

ALTERNATIVE UTILIZATIONS OF INFORMATION CRITERIA FOR GAUSSIAN REGRESSION ON A RANDOM DESIGN.

Guilhem Coq[1]

Abstract. We consider the problem of estimating an unknown function $f^\star$. Our data consist in a set of points in the plane, the abscisses of which are distributed according to a known density while their ordinates are the image of those abscisses by $f^\star$ deteriorated by a Gaussian white noise. To this end, we use general Information Criteria, also called penalized likelihood criteria. We introduce several 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.

1991 Mathematics Subject Classification. 62J02, 62F12.

The dates will be set by the publisher.

Introduction

We study the problem of Gaussian regression in the case called "random design" by Baraud [3] or Birgé [5]. In this setting, more precisely described in part 1.3, 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] consider the same problem except the abscisses are deterministic.

In all that paper, we address the regression problem using information criteria. Those tools are widely used to solve model selection problem in a general way. Let us cite for reference Akaike [1] who gives the AIC criterion (Akaike Information Criterion) in the early 70's and uses it in the problem of the determination of the order of an autoregression. Following Akaike, Schwarz [17] presents BIC for Bayesian Information Criterion in a more general setting. Rissanen studies the stochastic complexity of datas relatively to a set of models during the 80's [14,15]. His work, along with the MDL principle [9,13], allows to derive RIC, Rissanen Information Criterion, which is similar to BIC. Rissanen also applies stochastic complexity to study the problem of the selection of an histogram estimating an unknown density in [16]. In a general setting, Nishii studies the asymptotic properties of models selected by a general information criterion in [11]. Regarding the results of Nishii, El-Matouat and Al. present in [8] the so-called φ_β criterion that allows Nishii's result to apply.

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, 10] give 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

Keywords and phrases: Information Criteria, model selection, Gaussian regression, oracle inequality

[1] Laboratoire de Mathématiques et Applications de Poitiers, BP 30179, 86962 Futuroscope Chasseneuil France, coq@math.univ-poitiers.fr

© EDP Sciences, SMAI 1999

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 here is to provide methods of model selection, always based on information criteria, that are both "efficient" and "fast". 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.

Firstly we present the general criterion we use throughout the paper. Then, in part 2.5, we introduce the different methods under study and explain why they are fast by giving them a complexity in part 2.6. The asymptotic study, section 3, shows why those methods are efficient by giving the asymptotic behavior of the selected model. Result of this section are inspired by the work of Nishii [11]. We give in section 4 some simulations results illustrating the convergence theorems given earlier. Note that, for those simulations, we use the φ_β criterion since it is practical. Section 5 is devoted to the computation of risks in the random design case. We stress the main differences between the fixed design setting. Finally, in section 6, we give an oracle inequality regarding the risk of the estimation of $f^\star$ resulting from a fast model selection method presented earlier.

1. Notations

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 is denoted by $\|.\|_w$

We choose $F = \text{Vect}(f_1, \ldots, f_d)$ a d-dimensional subspace of L^2 with $d \leq n$ and denote by M_w the Gram matrix of the f_j's

$$M_w(j,k) = \langle f_j, f_k\rangle_w\,,\ j,k = 1,\ldots,d.$$

For any support $S \subset [\![1,d]\!]$ we denote by F_S the $|S|$-dimensionnal subspace of F

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

and by $M_{w,S}$ the associated Gram matrix $M_{w,S}(j,k) = \langle f_j, f_k\rangle_w\,,\ j,k \in S$. In F, the orthogonal projector on F_S is denoted by Π_S^F.

In the case where $f^\star$ actually lives in F, we write $f^\star = \sum a_j^\star f_j$ and call $S^\star$ its support

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

In the sequel it will be convenient to identify F to $\mathbb{R}^d$ via $F \to \mathbb{R}^d,\ f = \sum_{k=1}^d a_k f_k \mapsto a = (a_1,\ldots,a_d)^T$. This way, the squared norm of a function may be written as $\|f\|_w^2 = f^T M_w f$.

We will also need to apply central limit theorem several times in the sequel. In order to ensure that all considered variables have a variance, we assume from now on that there exists a $\eta > 0$ such that

$$f^\star, f_j \in L^{4+\eta}(w),\ j = 1,\ldots,d. \tag{1.4}$$

Note that we are constrained to suppose that $f^\star$ presents this integrability only in the case where it does not belong to the space of regression F.

1.2. **Observations space**

We endow the observations space $\mathbb{R}^n$ with the canonical scalar product $\langle .,. \rangle_n$ and its associated norm $\|.\|_n$. When $x \in I^n$ and a support S are given we denote by E_S the subspace

$$E_S = \text{Vect}\left\{f_j(x),\ j \in S\right\} \tag{1.5}$$

and shorten $E_{[\![1,d]\!]}$ to E. In $\mathbb{R}^n$, the orthogonal projector on E_S is denoted by Π_S^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 as follows :

$$Y = f^\star(X) + G \tag{1.6}$$

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 E_S (1.5) and their projection Π_S^E become random.

For any support $S \subset [\![1, d]\!]$, let us set the following model Θ_S :

$$\Theta_S = \left\{\mathcal{L}(f(X) + G) |\ f \in F_S\right\}. \tag{1.7}$$

Those are the law of n-dimensionnal variables which write as $f(X) + G$ where $f \in F_S$. If $f^\star$ lives in F and has support $S^\star$, the law of Y belongs to the model $\Theta_{S^\star}$.

2. Information criteria and their use

2.1. **Maximum likelihood**

Relatively to a function $f \in F$, a realization (x, y) of (X, Y) has the following likelihood :

$$\frac{1}{(2\pi\sigma^2)^{n/2}} \prod_{i=1}^n w(x_i) \exp\left(-\frac{1}{2\sigma^{2n}} \sum_{i=1}^n \left(y_i - f(x_i)\right)^2\right).$$

In the sequel we shall rather work with twice the opposite of the log-likelihood which we denote by l. Since we are only concerned in minimization of l with respect to f, we drop terms not depending on it and l may be seen as

$$\begin{aligned} l : F \sim \mathbb{R}^d &\to \mathbb{R} \\ f &\mapsto \sum_{i=1}^n \left(y_i - f(x_i)\right)^2 \\ a &\mapsto \sum_{i=1}^n \left(y_i - \sum_{k=1}^d a_k f_k(x_i)\right)^2. \end{aligned}$$

Consequently, for any support S, the maximization of the likelihood of y within the model Θ_S amounts to the minimization of the quadratic error l on F_S. Then the (opposite of the) maximum (log-)likelihood of y with respect to the model Θ_S writes as :

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

2.2. **Estimators $\widehat{f}_S$, $\widetilde{f}_S$**

In order to express asymptotics of this likelihood as well as to give a matricial expression of the function $\widehat{f}_S \in F_S$ that realizes the minimum in (2.1), it will be convenient to use the Vandermonde-type $n \times d$ matrix V depending only of the x_i's :

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

as well as the $d \times d$ Gram matrix $M = V^T V$. The passage from a function f to the vector $f(x)$ is then a simple multiplication $f(x) = Vf$. Also note that, in the case where $f^\star \in F$, we may write $y = Vf^\star + g$ and the Taylor expansion of l at any function f writes as

$$\begin{aligned} l(h) - l(f) &= (\text{grad } l_f)^T (h-f) + \frac{1}{2}(h-f)^T \text{Hess } l_f (h-f) \\ &= -2\left(M(f-f^\star) - V^T g\right)^T (h-f) + (h-f)^T M (h-f). \end{aligned} \tag{2.2}$$

Now let us set D_S the $d \times |S|$ matrix of zeros and ones such that $V_S := VD_S$ contains columns j of V for $j \in S$ only. We also set the Gram matrix $M_S = V_S^T V_S$. Consequently, the likelihood (2.1) writes as

$$L(y|\Theta_S) = d^2(y, E_S) = \frac{\text{Gram}\left(\{y, f_j(x),\, j \in S\}\right)}{\det\left(M_S\right)}. \tag{2.3}$$

where $\text{Gram}(u_1, \dots, u_k k)$ denotes the Gram determinant of those vectors. Moreover the matrix of the orthogonal projection Π_S^E of $\mathbb{R}^n$ onto E_S is

$$\text{Mat}(\Pi_S^E) = V_S M_S^{-1} V_S^T \tag{2.4}$$

which gives the following expression for $\widehat{f}_S$ realizing the minimum in (2.1) :

$$\widehat{f}_S = D_S M_S^{-1} V_S^T y \in F_S. \tag{2.5}$$

It is important to note that, unlike the fixed points case, it may happen that $\left\|\widehat{f}_S\right\|_w^2$ is not integrable. To handle this problem we will use at some points in the sequel a truncated estimator

$$\widetilde{f}_S = \widehat{f}_S . \mathbb{1}_{\{\|nM_S^{-1}\| < C\}} \tag{2.6}$$

where $\|.\|$ is any norm on matrices and C is a constant satisfying $C > \|M_{w,S}^{-1}\|$ so that $\mathbb{1}_{\{\|nM_S^{-1}\|<C\}} \to 1$ a.s. Now, from (2.5), 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 $y = f^\star(x) + g$. In that expression, matrices D_S and $M_{w,S}$ are deterministic and the indicator function ensures the boundedness of M_S^{-1}. Moreover, the integrability we have requested on $f^\star$ and on the f_j's in (1.4) along with the fact that G is Gaussian ensure that

$$\mathbb{E}\left[\left\|\widetilde{f}_S\right\|_w^{2p}\right] < \infty \text{ for any } p \text{ such that } 2p \le 4 + \eta. \tag{2.7}$$

Furthermore, for an event A, Hölder's inequality gives

$$\mathbb{E}\left[\left\|\widetilde{f}_S\right\|_w^2 \mathbb{1}_A\right] \leq \mathbb{E}\left[\left\|\widetilde{f}_S\right\|_w^{2p}\right]^{1/p} \mathbb{P}(A)^{1/q}. \tag{2.8}$$

for any $p > 1$ such that $2p \leq 4 + \eta$ and $1/p + 1/q = 1$.

The convergence of M_S/n to $M_{w,S}$ results from the law of large numbers. Moreover, since all f_j's are $L^4(w)$, any variable of the form $f_j(X)f_k(X)$ has a variance V smaller than

$$V_{max} = \max\left\{V\left(f^\star(X)^2\right), V\left(G_1^2\right), V\left(f_j(X)f_k(X)\right),\ j,k = 1,\ldots,d\right\}. \tag{2.9}$$

Consequently we may apply central limit theorem to obtain a control given by, for any $\varepsilon > 0$,

$$\mathbb{P}\left(\|M_s/n - M_{w,S}\| > \varepsilon\right) = O\left(\exp\left(\frac{-n\varepsilon^2}{4V_{max}}\right)\right).$$

Since the event $\{\|nM_S^{-1}\| < C\}^c$ is included in an event of the form $\|M_s/n - M_{w,S}\| > \varepsilon_C$ for a certain $\varepsilon_C > 0$ depending on C, we obtain that

$$\mathbb{E}\left[\mathbb{1}_{\{\|nM_S^{-1}\|<C\}^c}\right] = O\left(\exp\left(\frac{-n\varepsilon_C^2}{4V_{max}}\right)\right). \tag{2.10}$$

2.3. **Asymptotics of the maximum likelihood**

Because all the f_j's are L^2, the law of large numbers ensures that each entry $M_S(i,j)$ of M_S satisfies :

$$\frac{1}{n}M_S(i,j) \to M_{w,S}(i,j),\ a.s. \tag{2.11}$$

Moreover, since $y = f^\star(x) + g$ and by independence between X and G, the entry (1,1) of the first Gram matrix in (2.3) satisfies

$$\frac{1}{n}\left\langle f^\star(x) + g, f^\star(x) + g\right\rangle_n \to \left\langle f^\star, f^\star\right\rangle_w + \sigma^2\ a.s. \tag{2.12}$$

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

$$\frac{1}{n}l(\widehat{f}_S) \to \sigma^2 + d^2(f^\star, F_S)\ \text{a.s.} \tag{2.13}$$

Now we need to control the speed of the convergence in (2.13). Every variables for whom the law of large numbers has been used in (2.11) and (2.12) have a variance because of (1.4) and because the error G is Gaussian. All those variances are smaller than V_{max} defined earlier in (2.9). Then we may apply central limit theorem to obtain that convergences in (2.11) and (2.12) have a speed given by $O\left(\exp(-n\varepsilon^2/4V_{max})\right)$, for any $\varepsilon > 0$. That speed passes to the determinant to give

$$\mathbb{P}\left(\left|\frac{1}{n}l(\widehat{f}_S) - \sigma^2 - d^2(f^\star, F_S)\right| > \varepsilon\right) = O\left(\exp\left(\frac{-n\varepsilon^2}{4V_{max}}\right)\right). \tag{2.14}$$

2.4. **The criterion**

Let us choose any function $\alpha : \mathbb{N} \to \mathbb{R}^+$ referred to in the sequel as the penalty term. Here comes the criterion to be minimized among all possible supports $S \subset [\![1,d]\!]$:

$$\mathrm{IC}(S) = l(\widehat{f}_S) + |S|\alpha(n). \tag{2.15}$$

This minimization realizes the best trade-off between the goodness of fit, measured by the likelihood term $l(\widehat{f}_S) = d^2(y, E_S)$ and the complexity of the model needed to obtain such a fit, measured by the penalty $|S|\alpha(n)$.

In this paper, we are not only interested on properly choosing the penalty $\alpha(n)$ but also in "efficient" and "fast" methods allowing to use the criterion. We present those methods now.

2.5. **Methods**

By method, we understand a sequence of computations of $\mathrm{IC}(S)$ that eventually leads to a selection of a support $\widehat{S}$. We define several of those in the sequel. The methods we use allow to select a support that may take any value in $\mathcal{P}(\llbracket 1, d \rrbracket)$. 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 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.

We will always denote the selected support by $\widehat{S}$ but in the sequel the context will allow to determine which method is used.

2.5.1. *Global method*

The estimated support is chosen as

$$\widehat{S} = \mathrm{Argmin}\left\{\mathrm{IC}(S), S \in \mathcal{P}(\llbracket 1, d \rrbracket)\right\}. \tag{2.16}$$

2.5.2. *Comparative method*

Rather than testing any support as in the global method, Nishii suggest in [11] the following method. The expression "$-j$" denotes the support $\llbracket 1, d \rrbracket \setminus \{j\}$.

$$\begin{aligned} \mathrm{IC}_{\mathrm{ref}} &= \mathrm{IC}(\llbracket 1, d \rrbracket) \\ \widehat{S} &= \left\{j \in \llbracket 1, d \rrbracket \text{ such that } \mathrm{IC}_{\mathrm{ref}} \leq \mathrm{IC}(-j)\right\} \end{aligned} \tag{2.17}$$

Here the value $\mathrm{IC}_{\mathrm{ref}}$ of the criterion where every basis functions is allowed to appear in the curve is taken as a reference. Then it is compared to the value $\mathrm{IC}(-j)$ where the specific function f_j is forbidden. If the criterion prefers that function f_j appears, *i.e.* $\mathrm{IC}_{\mathrm{ref}} \leq \mathrm{IC}(-j)$, we keep it in $\widehat{S}$, otherwise we reject it.

2.5.3. *Reversed comparative method*

Looking at method (2.17) one may ask why not to consider the following :

$$\begin{aligned} \mathrm{IC}_{\mathrm{ref}} &= \mathrm{IC}(\emptyset) \\ \widehat{S} &= \left\{j \in \llbracket 1, d \rrbracket \text{ such that } \mathrm{IC}(\{j\}) \leq \mathrm{IC}_{\mathrm{ref}}\right\} \end{aligned} \tag{2.18}$$

This way, we compare the null curve to curves where only a single function f_j is allowed and keep indices for which the second curve is prefered via $\mathrm{IC}(\{j\}) \leq \mathrm{IC}_{\mathrm{ref}}$.

The asymptotic flaws of this method will be precisely discussed in part 3.2. However, we may already give some comments about it. Suppose the basis $(f_j)_j$ is orthogonal relatively to the scalar product (1.1), that $f^\star$ belongs to F and writes $f^\star = \sum_j a_j f_j$. Then from the asymptotic behaviour of likelihood terms in our criterion given in (2.13), we obtain that the difference $n^{-1}(\mathrm{IC}(\{j\}) - \mathrm{IC}(\emptyset))$ behaves as $n^{-1}\alpha(n) - a_j^2$. Then the decision of the reversed method regarding the function f_j is related only to the coefficient a_j of $f^\star$ which vanishes if and only if f_j is to be rejected. Note that, in the same setting, the regular comparative method (2.17) is interested in the difference $n^{-1}(\mathrm{IC}_{\llbracket 1,d \rrbracket} - \mathrm{IC}(-j))$ that also behaves as $n^{-1}\alpha(n) - a_j^2$ because of (2.13). Those two methods, in the orthogonal case, are asymptotically equivalent.

Now for the non-orthogonal case, still from (2.13), the difference $n^{-1}(\mathrm{IC}(\{j\}) - \mathrm{IC}(\emptyset))$ only behaves as $n^{-1}\alpha(n) + d^2(f^\star, F_j) - \|f^\star\|_w^2$. Here, even in the case where f_j is to be rejected, the term $d^2(f^\star, F_j) - \|f^\star\|_w^2$ remains negative, requiring a stronger penalization for the reversed method to actually reject f_j. In fact, it will be shown in the asymptotic study that this need of a large penalization may almost not be fulfilled without rejecting every functions. By opposition, still in the non-orthogonal case, for the regular comparative method (2.17) the difference $n^{-1}(\mathrm{IC}_{[\![1,d]\!]} - \mathrm{IC}(-j))$ behaves as $n^{-1}\alpha(n) - d^2(f^\star, F_j)$ where the term $d^2(f^\star, F_j)$ vanishes when f_j is to be rejected. In this sense, the regular comparative method is not affected by the orthogonality of the basis.

To sum up those comments, let us just say that in the orthogonal case, reversed and regular comparative methods have the same asymptotic behaviour whereas in the non-orthogonal case the reversed method suffers flaws that will be more precisely described in the part 3.2 of the asymptotic study.

Finally, let us say that 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, as well as all others criteria needed to select $\widehat{S}$ in (2.17), are not computable in the infinite dimensionnal case.

2.5.4. *Adapted reversed comparative method*

We give here an adaptation of the reversed comparative method that will be shown to avoid issues occuring with the previous one. To this end we need to define new functions f_j^N ; the superscript N stands for normal. Indeed, each function f_j^N is chosen to be 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.

We may then define new alternative models as in (1.7) :

$$\Theta_j^N = \left\{\mathcal{L}(f(X) + G) |\ f = a_j^N f_j^N\right\},\ a_j^N \neq 0.$$

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 :

$$\begin{aligned} \mathrm{IC}_{\mathrm{ref}} &= \mathrm{IC}(\emptyset) \\ \widehat{S} &= \left\{j \in [\![1,d]\!] \text{ such that } l(\widehat{f}_S^N) + \alpha(n) \leq \mathrm{IC}_{\mathrm{ref}}\right\}. \end{aligned} \tag{2.19}$$

Note that the basis $(f_j^N)_j$ is orthogonal if and only $(f_j)_j$ is. In this case, f_j^N is colinear to f_j which makes (2.18) and (2.19) equivalent methods of selection of $\widehat{S}$.

In the sequel, it will be shown that under some assumptions on the penalty, this method satisfies a convergence theorem 3.3 . Let us stress that, even though this theorem applies, this method suffers the same flaws as the reversed comparative method (2.18) since it remains a *reversed* method. Indeed, as in comments made before, when f_j is to be rejected, the adapted reversed comparative method is strongly affected by the orthogonality of f_j^N and $f^\star$. However, calculations are not made with those function but rather with $f_j^N(X)$ and $f^\star(X)$ that are orthogonal only asymptotically. Consequently, for a fixed n, we also expect this method to require a larger penalization in order to avoid overparametrization.

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

TABLE 1. Methods and their complexities.

2.5.5. *Descending comparative method*

The descending comparative method is designed especially in the aim of using results from Baraud in [2,3]. It requires a random computation of ICs. Firstly let us set

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

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

$$\begin{aligned} C^{(1)} &= \left\{ j \in S^{(0)},\ \mathrm{IC}\left(S^{(0)} \setminus \{j\}\right) \leq \mathrm{IC}^{(0)}_{\mathrm{ref}} \right\} \\ J^{(1)} &= \mathrm{Argmin}\left\{ \mathrm{IC}\left(S^{(0)} \setminus \{j\}\right),\ j \in C^{(1)} \right\}. \end{aligned} \tag{2.20}$$

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) :

$$\begin{aligned} S^{(1)} &= S^{(0)} \setminus \{J^{(1)}\} \\ \mathrm{IC}^{(1)}_{\mathrm{ref}} &= \mathrm{IC}(S^{(1)}). \end{aligned} \tag{2.21}$$

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

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

and refresh again our reference by adding 1 to all superscripts in (2.21).

This process is repeated until the random 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 $\widehat{S}$. We say that the procedure stops at step k_f for "k final".

2.6. **Methods complexities**

We also define here an integer willing to reflect the complexity of a particular method in terms of computations. This complexity reflects how fast the method is. We define it simply as the number of ICs that one needs to compute in order to obtain the estimator $\widehat{S}$. For instance the complexity of the global method (2.16) is the number of partition of $[\![1,d]\!]$, that is 2^d.

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 in d 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 $\mathcal{P}(\llbracket 1,d \rrbracket)$. They are what we have called "fast" methods. The remainder of the paper is devoted to explain in which way they are also "efficient".

3. Asymptotic study

In all our asymptotic study, we assume that $f^\star$ lives in F and has support $S^\star$ as in (1.3). Our concern is then to determine conditions on our criterion (2.15), more precisely on its penalty term, to obtain some convergence of $\widehat{S}$ to $S^\star$ as n grows. The main results are given in theorems 3.1, 3.2, 3.3, 3.4.

3.1. **Asymptotics of the comparative method**

Here we work with the comparative method (2.17) and criterion (2.15). The main result is as follows.

Theorem 3.1. *If the penalty $\alpha(n)$ of the criterion (2.15) satisfies*

(i) $\alpha(n) = o(n)$

(ii) $\ln\ln n = o(\alpha(n))$

then the method (2.17) is strongly consistent in the sense that $\widehat{S}$ converges (stationnarily) to $S^\star$ almost surely. More precisely, conditions (i) and (ii) ensure respectively $S^\star \subset \widehat{S}$ and $\widehat{S} \subset S^\star$ a.s. above a certain rank.

Remark. The proof actually shows that we might relax condition (i) to

$$\alpha(n) = c\,n \text{ where } c < \min\left\{d^2(f^\star, F_{-j}),\ j \in S^\star\right\}$$

but that min is not available to the user.

Proof : Let us split it into two parts.

$\star$ **First part** : the case $j \in S^\star$. Here, let us use asymptotics of the likelihood term given in (2.14) to write :

$$\frac{1}{n}\left(\mathrm{IC}_{\mathrm{ref}} - \mathrm{IC}(-j) - \alpha(n)\right) = -d^2(f^\star, F_{-j}) + o(1) \text{ a.s.}, \tag{3.1}$$

where that latter distance $D^2 := d^2(f^\star, F_{-j})$ 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. It suffices to choose $\alpha(n)$ such that $\alpha(n) = o(n)$ to ensure that fact a.s. for n large enough.

Taking a not too large penalty (of order $o(n)$) thus ensures that we do not reject basis functions f_j that actually appear in the unknown function $f^\star$, this is the first statement of the theorem.

In this case, let us take n large enough to make $0 \le \alpha(n)/n < D^2/2$ to get the following control :

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

Hence, from (2.14) :

$$\mathbb{P}(j \notin \widehat{S}) = O\left(\exp\left(-\frac{nD^4}{64 V_{max}}\right)\right),\ j \in S^\star \tag{3.2}$$

which will be useful in studies of risks to come later.

$\star\star$ **Second part** : the case $j \notin S^\star$. Recall the definition of the (random) matrix M in part 2.1. Since each f_j is in L^2 the law of large numbers gives

$$M/n \longrightarrow M_w$$

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

$$M^{-1} = O\left(\frac{1}{n}\right) \text{ a.s.} \tag{3.3}$$

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

$$\frac{1}{m_j} = O(n) \text{ a.s.} \tag{3.4}$$

We call here for brievety $\widehat{f}$ and $\widehat{f^j}$ the functions in F and F_{-j} respectively maximizing the likelihood on those spaces as in (2.1). From (2.2), the quantities we are interested in are:

$$l\big(\widehat{f^j}\big) - l\big(\widehat{f}\big) = \big(\widehat{f^j} - \widehat{f}\big)^T M \big(\widehat{f^j} - \widehat{f}\big) \tag{3.5}$$

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

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

$$\Big(\text{grad } l(\widehat{f^j})\Big)^T = M\widehat{f^j} - M\widehat{f} = (0, \ldots, 0, \lambda_j, 0, \ldots, 0)^T \tag{3.6}$$

where λ_j is a Lagrange coefficient set at the j-th place in the latter vector which necessarily satisfies

$$\lambda_j = -\frac{\widehat{f}_j}{m_j} \tag{3.7}$$

since $\widehat{f}_j^j$, the j-th coefficient of $\widehat{f^j}$, must vanish.

Plugging (3.4), (3.6) and (3.7) in expansion (3.5) gives

$$l\big(\widehat{f^j}\big) - l\big(\widehat{f}\big) = \Big(\widehat{f}_j\Big)^2 O(n) \tag{3.8}$$

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 from (1.4), and thus is $O(\sqrt{n \ln\ln n})$ a.s. by the law of iterated logarithm. The order of M^{-1} given in (3.3) makes $\widehat{f}_j = O\left(\sqrt{\frac{\ln\ln n}{n}}\right)$ a.s.

Plugging in (3.8), 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

$$\frac{1}{\ln\ln n}\left(\mathrm{IC}_{\mathrm{ref}} - \mathrm{IC}(-j) - \alpha(n)\right) = O(1) \text{ a.s.} \tag{3.9}$$

Now it suffices to choose $\alpha(n)$ such that $\ln\ln n = o(\alpha(n))$ to ensure that a.s. and for n large enough, $\mathrm{IC}_{\mathrm{ref}} - \mathrm{IC}(-j) > 0$.

We have shown that taking a large enough penalty ensures that we reject basis functions f_j which do not appear in $f^\star$, this is the second statement of the theorem. □

3.2. **Asymptotics of the reversed comparative method**

We work here with the method (2.18).

3.2.1. *Asymptotics of the likelihood difference*

The function achieving the maximum likelihood for $\mathrm{IC}^r_{\mathrm{ref}}$ obviously vanishes and the function with support $\{j\}$ achieving it for $\mathrm{IC}(\{j\})$ is the new function $\widehat{f^j}$ whose component along f_j is:

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

Now the difference of the log-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, we firstly 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. :

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

Consequently, a.s. :

$$l(0) - l(\widehat{f^j}) = n\,\langle f_j, f^\star\rangle_w^2 + \langle f_j, f^\star\rangle_w O\left(\sqrt{n\ln\ln n}\right) + O(\ln\ln n). \tag{3.10}$$

3.2.2. *Non-orthonormal case*

Here appear the main problem about the reversed method (2.18). Suppose we are in a case where $\langle f_j, f^\star\rangle_w = \sum_{k\in S^\star} a_k \langle f_j, f_k\rangle_w$ vanishes for no index j. This happens most of the times if the basis $(f_k, k \in [\![1,d]\!])$ is not orthonormal even though $j \notin S^\star$. Then formula (3.10) gives

$$\frac{1}{n}\left(\mathrm{IC}^r_{\mathrm{ref}} - \mathrm{IC}(\{j\}) + \alpha(n)\right) = \langle f_j, f^\star\rangle_w^2 + o(1). \tag{3.11}$$

Now assume $\alpha(n) = o(n)$, then $\mathrm{IC}^r_{\mathrm{ref}} - \mathrm{IC}(j) > 0$ a.s. above a certain rank and this for all j. This means we keep every indices $j \in [\![1,d]\!]$. Here the condition $\alpha(n) = o(n)$ which ensured we kept good indices with the regular comparative method (see theorem 3.1) turns out to make us keep every indices.

One should thus think about taking a penalty a bit larger. However, formula (3.11) also implies that if

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

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

Therefore, in order to obtain a result similar to theorems 3.1 one should firstly choose a penalty of order not smaller than n and not greater than Cn to ensure that we do not have a criterion that accept or reject systematically every indices. In certain cases, there exists a good order for the penalty in the few place left between n and Cn as follows.

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

However, that k might not exist if its bounds are not ordered the correct way ; moreover, it is unavailable to the user and thus not of a practical use.

In the case where the basis (f_j) is orthonormal, those issues disappear and we establish in the next part a convergence theorem for the reversed comparative method.

3.2.3. *Orthonormal case*

We assume here that the basis $(f_j, j \in [\![1, d]\!])$ is orthonormal relatively to the scalar product (1.1). Note that this is often the case ; let us cite for instance the situation where one needs to decompose a measured signal on a wavelet or a Fourier basis (in this case, w is the uniform density on the corresponding interval).

Recall that $f^\star = \sum_{j\in S^\star} a_j f_j$, consequently $a_j = \langle f_j, f^\star\rangle_w = 0 \Leftrightarrow j \notin S^\star$ and (3.10) yields

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

This formula is to be related to equations (3.1) and (3.9) concerning the regular comparative method to see that, in the orthonormal case, the reversed method behaves asymptotically as the regular method. The following theorem is derived from considerations similar to the proof of theorem 3.1.

Theorem 3.2. *In the orthonormal case and under the following assumptions on the penalty :*

(i) $\alpha(n) = o(n)$
(ii) $\ln\ln n = o(\alpha(n))$

the selection of indexes by the reversed comparative method (2.18) is strongly consistent.

More precisely, conditions (i) and (ii) ensure respectively $S^\star \subset \widehat{S}$ *and* $\widehat{S} \subset S^\star$ *a.s. above a certain rank.*

The orthonormal case as treated here gives the idea of considering the adapted reversed comparative method (2.19). In the next part, we show that this method is consistent without the orthonormality hypothesis.

3.3. **Asymptotics of the adapted reversed comparative method**

We work here with the method (2.19) and fix $j \in [\![1, d]\!]$. Arguments in part 3.2.1 may be transposed here simply by replacing f_j with f_j^N and noting that $\left\langle f_j^N, f^\star\right\rangle_w = a_j^N$. We get a formula similar to (3.10) :

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

Now recalling that $a_j^N \neq 0 \Leftrightarrow j \in S^\star$, we obtain a result similar to theorem 3.2 for the adapted reversed comparative method without the orthonormality hypothesis.

Theorem 3.3. *Under the following assumptions on the penalty :*

(i) $\alpha(n) = o(n)$
(ii) $\ln\ln n = o(\alpha(n))$,

the selection of indexes by the adapted reversed comparative method (2.19) is strongly consistent.

More precisely, conditions (i) and (ii) ensure respectively $S^\star \subset \widehat{S}$ *and* $\widehat{S} \subset S^\star$ *a.s. above a certain rank.*

3.4. **Asymptotics of the descending comparative method**

We work here with the method described in part 2.5.5 and show the

Theorem 3.4. *Under the following assumptions on the penalty :*
(i) $\alpha(n) = o(n)$
(ii) $\ln \ln n = o(\alpha(n))$,
the selection of indexes by the descending comparative method in part (2.5.5) is strongly consistent.

More precisely, conditions (i) and (ii) ensure respectively $S^\star \subset \widehat{S}$ *and* $\widehat{S} \subset S^\star$ *a.s. above a certain rank.*

Proof : Under assumptions of that theorem, we may apply theorem 3.1 concerning the regular comparative method. Then, with probability 1 and for n large enough, the indices $C^{(1)}$ selected in the first step (2.20) are exactly $S^{\star c}$ so that $J^{(1)} \notin S^\star$ and $S^\star \subset S^{(1)}$.

Then, applying again theorem 3.1 gives that, with n possibly larger, $J^{(2)} \notin S^\star$. Once that process has been iterated enough times to eliminate all indices out of $S^\star$, theorem 3.1 once again ensures that the following choice of $C^{(k)}$ will lead to the empty set. This completes the proof. □

4. Simulations

In this section, we present simulation results illustrating theorems 3.1, 3.2, 3.3, 3.4 regarding convergence of the chosen supports $\widehat{S}$ selected by our alternative methods toward the true support $S^\star$.

4.1. **Setting**

The unknown function we consider is

$$\begin{array}{rcl} f^\star : [-\pi, \pi] & \to & \mathbb{R} \\ x & \mapsto & f^\star(x) = -x + \cos(2x) - \sin(2x). \end{array} \tag{4.1}$$

The distribution of the abscisses is given by the density

$$w = \frac{1}{2\pi} \mathbb{1}_{[-\pi,\pi]} \tag{4.2}$$

and the model considered is a 6-dimensionnal space $F = \mathrm{Vect}(f_1, \ldots, f_6)$ where

$$\begin{array}{lll} f_1(x) = x, & f_3(x) = \cos(x), & f_5(x) = \sin(x), \\ f_2(x) = x^2, & f_4(x) = \cos(2x), & f_6(x) = \sin(2x). \end{array} \tag{4.3}$$

Note that $f^\star \in F$.

4.2. **The φ_β criterion**

Recall that the main assumption of theorems 3.1, 3.2, 3.3, 3.4 is that the penalty $\alpha(n)$ in the criterion (2.15) satisfies

(i) $\alpha(n) = o(n)$
(ii) $\ln \ln n = o(\alpha(n))$.

Considering this, authors in [8], suggest to use the following that may be seen as a parametrization of penalties allowing previous convergence theorems to apply :

$$\alpha(n) = n^\beta \ln \ln n, \ \beta \in (0, 1).$$

For any $\beta \in (0,1)$ we thus obtain a criterion in the sense of (2.15) referred to as φ_β in the sequel :

$$\varphi_\beta(S) = d^2(y, E_S) + |S| n^\beta \ln \ln n. \tag{4.4}$$

When n is fixed, correctly choosing the value of β in (4.4) allows to recover usual information criteria. Among them, we consider the historical AIC criterion presented by Akaike [1] and the BIC criterion established by Schwarz [17] or Rissanen [14, 15] which, in our setting, write as

$$\begin{aligned} \mathrm{AIC}(S) &= d^2(y, E_S) + 2|S| \\ \mathrm{BIC}(S) &= d^2(y, E_S) + |S| \ln n \end{aligned} \tag{4.5}$$

The values of β in the φ_β criterion allowing to recover AIC and BIC are respectively

$$\begin{aligned} \beta_{\mathrm{AIC}} &= (\ln 2 - \ln \ln \ln n)/\ln n \\ \beta_{\mathrm{BIC}} &= (\ln \ln n - \ln \ln \ln n)/\ln n. \end{aligned} \tag{4.6}$$

4.3. **Results**

We generate 100 set of observations of the couple (X, Y) linked by the relation $Y = f^\star(X) + G$. On each of those observations we apply the φ_β criterion, for β ranging from 0 to 1 by step 0.05, along with the following methods :

- ⋆ Global method (2.16).
- ⋆ Comparative method (2.17)
- ⋆ Reversed comparative method (2.18)
- ⋆ Adapted reversed comparative method (2.19), shortened to "adapted"
- ⋆ Descending method (part 2.5.5).

Recall that the first one has exponential complexity while the others have polynomial complexities which make them much faster to use. In our setting, the global method took 270 seconds to provide results while any of the 4 comparative methods needed about 13 seconds.

We count a success if the selected support is exactly the one of $f^\star$ in the setting given by (4.1), (4.2) and (4.3) ; that is $S^\star = \{1, 4, 6\}$. Note that, since the basis of our setting is not orthonormal relatively to (1.1), the reversed comparative method should not give good results as seen in part 3.2. Actually, this method always give a percentage succes of 0 in all our simulations, this is why it does not appear in our results.

Figure 1 presents the percentage of succes of the different methods plugged against the value of β in (4.4) for $n = 20, 50, 200$ and 1000. The two vertical lines correspond, from the left to the right, to the values β_{AIC} and β_{BIC} given in (4.6). Note that most of the time, descending and global methods give the same percentage.

4.4. **Comments**

Firstly let us say that when β is too low, the penalization of the criterion is too weak and overparametrization occurs : $\widehat{S}$ contains too many functions, thus the failure. By opposition, when β is too large, underparametrization occurs and $\widehat{S}$ does not contain $S^\star$.

Now, as n grows, we observe an increasing rate of succes for any fixed value of β as convergence theorems of the previous section announced.

However, the AIC criterion (4.5) (corresponding to the first vertical line) does not fullfill those theorems requirements and thus present a quite low percentage of succes. The AIC criterion is known for its lack of penalization yielding overparametrization ; this is what we observe here.

The BIC criterion (4.5) (corresponding to the second vertical line) does not give here the fastest increasing rate of success. It seems it also lacks a little more penalization to reach the 100% of success sooner. Our previous use of the φ_β criterion in other model selection problems (such as autoregression order determination) also resulted in the same conclusion regarding the BIC criterion.

As announced when the adapted reversed comparative method (2.19) was introduced (part 2.5.4), it requires a bit more penalization than regular comparative method in order to avoid overparametrization. This fact appears on figure 1.

Finally let us stress that the regression problem as studied in this paper is the one, to our knowledge, that allows the biggest penalization before underparametrization occurs. Indeed, in the others model selection problems we studied with the φ_β criterion, we obtained results similar to figure 1 except that the success rate of any method fell back to 0 before β reached 0.5. This occured even with values of n much larger than 2000.

4.5. **A brief word on future applications**

Choosing $I = [-\pi, \pi]$, w the uniform density on I, $f^\star(x) = x$ and a basis consisting of $\cos(ax)$, $\sin(bx)$ where $a, b \in [\![0, 5]\!]$, we observed that comparative methods select supports containing only sinus functions and produced the following estimation of $f^\star$:

$$2.02\sin(x) - 1.02\sin(2x) + 0.71\sin(3x) - 0.47\sin(4x) + 0.46\sin(5x)$$

whereas the Fourier series of $f^\star$ starts by

$$2\sin(x) - \sin(2x) + \frac{2}{3}\sin(3x) - \frac{1}{2}\sin(4x) + \frac{2}{5}\sin(5x) + ...$$

This observations enlights the fact that linear regression helps finding the most important harmonics contained in a noised signal. We currently work on finding those harmonics on real signals, e.g. heartbeat signals measured before or after physical efforts. The same work might also be done with wavelets basis.

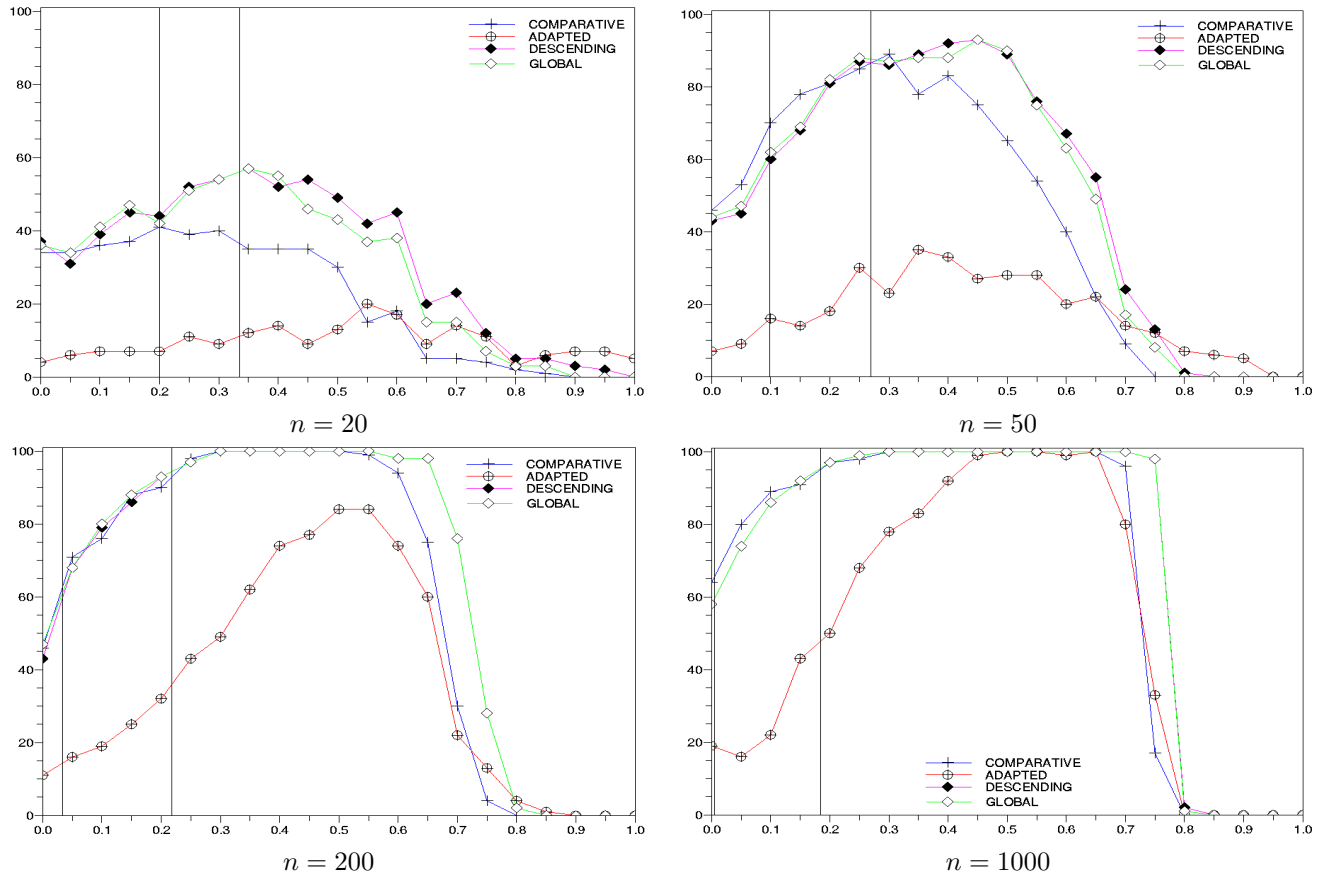

FIGURE 1. Percentage of succes of the different methods against the value of β in the φ_β criterion (4.4). Vertical lines correspond respectively to AIC and BIC criteria (4.5).

5. Study of the risks

5.1. Expression of the risks

The aim of this part is to show non-asymptotic differences between our setting and the fixed points case in terms of risk for our estimators. Because of the remark following (2.5), we are constrained here to make the following assumption :

$$\mathbb{E}\left[M_S^{-1}\right] < +\infty. \tag{5.1}$$

Note that, from (2.5), this assumption also implies that $\mathbb{E}\left[\left\|\widehat{f}_S\right\|_w^2\right] < +\infty$ and thus gives sense to following computations about the risk of $\widehat{f}_S$ as an estimator of $f^\star$. We have reason to believe that, under mild conditions on the regression space (such as every f_j presents at least a non-vanishing derivative at any points), there exists a fixed N such that (5.1) holds for $n \geq N$. Here, we are not interested in such conditions but in their consequences on expressions of the bias and risk of $\widehat{f}_S$.

We work here with a fixed support S that belongs to either of the families $\mathcal{F}_1$ and $\mathcal{F}_2$ defined by

$$\mathcal{F}_1 = \{S \in [\![1,d]\!] \,|\, S^\star \not\subseteq S\} \text{ and } \mathcal{F}_2 = \{S \in [\![1,d]\!] \,|\, S^\star \subseteq S\}. \tag{5.2}$$

In the case where $f^\star$ does not live in F, we simply set $\mathcal{F}_2 = \emptyset$ and $\mathcal{F}_1 = \mathcal{P}([\![1,d]\!])$.

Recall briefly that, in the fixed points setting studied for instance in [2], computations of losses are done directly in the space of observations $\mathbb{R}^n$ endowed with the normalized canonical norm $n^{-1}\|.\|_n$ that is chosen to satisfy $n^{-1}\|t\|_n^2 = n^{-1}\sum_i t(x_i)^2$ for all $t \in F$. We obtain :

$$\begin{aligned} \mathbb{E}\widehat{f}_S &= \Pi_S^F f^\star \\ \mathbb{E}\left[\|\widehat{f}_S - f^\star\|_n^2\right] &= \|f^\star - \Pi_S^F f^\star\|_n^2 + \sigma^2|S|/n. \end{aligned} \tag{5.3}$$

5.1.1. *The case $S \in \mathcal{F}_2$*

Remark that the multiplication $D_S D_S^T f$ sets to 0 the components of f along each f_j, $j \notin S$. Since $f^\star$ has support $S^\star$ we get $D_S D_S^T f^\star = f^\star$ and formula (2.5) gives

$$\widehat{f}_S = D_S M_S^{-1} V_S^T (V D_S D_S^T f^\star + g) = f^\star + D_S M_S^{-1} V_S^T g. \tag{5.4}$$

By independence between X and G we obtain $\mathbb{E}\widehat{f}_S = f^\star = \Pi_S^F f^\star$, that is $\widehat{f}_S$ is an unbiased estimator of $\Pi_S^F f^\star$. In order to compute the risk let us remark that

$$\left\|\widehat{f}_S - f^\star\right\|_w^2 = g^T V_S M_S^{-1} D_S^T M_w D_S M_S^{-1} V_S^T g =: g^T A_S\, g. \tag{5.5}$$

One of the main difference between the deterministic points case and our case appears in that matrix A_S. Indeed, in the former case, the computation of losses is done in the space of observations $\mathbb{R}^n$ rather than in F. In other words, the matrix $M_{w,S}$ in A_S is replaced by M_S which reduces A_S to the matrix of Π_S^E as in (2.4) and gives an exact formula for the variance. In our settings we are not able to derive such a formula and may only write :

$$\begin{aligned} \mathbb{E}\widehat{f}_S &= f^\star = \Pi_S^F f^\star \\ \mathbb{E}\left[\left\|\widehat{f}_S - f^\star\right\|_w^2\right] &= \sigma^2\mathbb{E}\left[\mathrm{Tr}(A_S)\right] = \sigma^2\mathbb{E}\left[\mathrm{Tr}(M_{w,S}M_S^{-1})\right]. \end{aligned} \tag{5.6}$$

In this case $S \in \mathcal{F}_2$, the result (5.6) is quite close to (5.3).

5.1.2. *The case $S \in \mathcal{F}_1$*

Here we may only write $y = Vf^\star + g$ and (2.5) gives

$$\mathbb{E}\widehat{f}_S = \mathbb{E}\left[D_S M_S^{-1} D_S^T M f^\star\right] =: \mathbb{E}\left[B_S f^\star\right].$$

Note that $B_S \to D_S M_{w,S}^{-1} D_S^T M_w$ a.s., that latter being the matrix of the orthogonal projector Π_S^F. However $\mathbb{E}B_S \neq D_S M_{w,S}^{-1} D_S^T M_w$. Consequently, as an estimator of $\Pi_S^F f^\star$, $\widehat{f}_S$ has a bias.

Now for the risk we write

$$\begin{aligned}
\mathbb{E}\left[\left\|\widehat{f}_S - f^\star\right\|_w^2\right] &= \mathbb{E}\left[\left\|f^\star - \Pi_S^F f^\star + \Pi_S^F f^\star - \mathbb{E}\widehat{f}_S + \mathbb{E}\widehat{f}_S - \widehat{f}_S\right\|_w^2\right] \\
&= \left\|f^\star - \Pi_S^F f^\star\right\|_w^2 + \left\|\Pi_S^F f^\star - \mathbb{E}\widehat{f}_S\right\|_w^2 \\
&\quad + \mathbb{E}\left[\left\|\mathbb{E}\widehat{f}_S - \widehat{f}_S\right\|_w^2\right],
\end{aligned}$$

moreover

$$\begin{aligned}
\mathbb{E}\left[\left\|\widehat{f}_S - \mathbb{E}\widehat{f}_S\right\|_w^2\right] &= \mathbb{E}\left[\left\|B_S f^\star + D_S M_S^{-1} V_S^T g - \mathbb{E}\left[B_S f^\star\right]\right\|_w^2\right] \\
&= \mathbb{E}\left[\left\|B_S f^\star - \mathbb{E}\left[B_S f^\star\right]\right\|_w^2\right] + \mathbb{E}\left[\left\|D_S M_S^{-1} V_S^T g\right\|_w^2\right],
\end{aligned}$$

the expectation of the scalar product vanishing since X and G are independent. Finally, non asymptotically :

$$\begin{aligned}
\mathbb{E}\widehat{f}_S &= \mathbb{E}\left[B_S f^\star\right] \\
\mathbb{E}\left[\left\|\widehat{f}_S - f^\star\right\|_w^2\right] &= \left\|f^\star - \Pi_S^F f^\star\right\|_w^2 + \sigma^2 \mathbb{E}\left[\mathrm{Tr}(M_{w,S} M_S^{-1})\right] \\
&\quad + \left\|\Pi_S^F f^\star - \mathbb{E}\left[B_S f^\star\right]\right\|_w^2 + \mathbb{E}\left[\left\|B_S f^\star - \mathbb{E}\left[B_S f^\star\right]\right\|_w^2\right] \\
&= \left\|f^\star - \Pi_S^F f^\star\right\|_w^2 + \sigma^2 \mathbb{E}\left[\mathrm{Tr}(M_{w,S} M_S^{-1})\right] \\
&\quad + \mathbb{E}\left[\left\|B_S f^\star - \Pi_S^F f^\star\right\|_w^2\right]
\end{aligned} \tag{5.7}$$

Comparing this result to (5.3), we get a new bias and a variance term $\mathbb{E}\left[\left\|B_S f^\star - \Pi_S^F f^\star\right\|_w^2\right]$. Those are created by the randomness on the X_i's since, in the fixed setting, the expression $B_S f^\star - \Pi_S^F f^\star$ vanishes.

5.2. **Asymptotics of the risks**

We no longer suppose (5.1). Our aim is now to derive asymptotic results similar to those given in [12], except we handle the random points case. More precisely, we prove that assumptions of theorems 3.1, 3.2, 3.3, 3.4 also ensure an asymptotic risk equivalent to an oracle risk. Recall that the remark following (2.5) prevents us from computing risks in a general case. We handle this issue by using the truncated estimator $\widetilde{f}_S$ defined in (2.6).

5.2.1. *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 $\mathcal{O}R(n, S^\star) = \mathbb{E}\left[\left\|\widetilde{f}_{S^\star} - f^\star\right\|_w^2\right]$.

Note that the event $\{\|M_{S^\star}^{-1}\| < C\}$ appearing in (2.6) is independent of the noise g. Therefore, following computations similar to part 5.1.1 we get

$$\mathcal{O}R(n, S^\star) = \|f^\star\|_w^2 \, \mathbb{E}\left[\mathbb{1}_{\{\|nM_{S^\star}^{-1}\|<C\}^c}\right] + \sigma^2 \mathbb{E}\left[\mathbb{1}_{\{\|nM_{S^\star}^{-1}\|<C\}} \mathrm{Tr}(M_{w,S^\star} M_{S^\star}^{-1})\right].$$

The first term is handled by (2.10). For the second, the indicator function gives a dominated convergence allowing to write

$$\mathcal{O}R(n, S^\star) \sim \frac{\sigma^2 |S^\star|}{n} \tag{5.8}$$

5.2.2. *Risk of our procedures*

We assume now that the support $\widehat{S}$ has been selected by an IC (2.15) whose penalty satisfies

(*i*) $\alpha(n) = o(n)$
(*ii*) $\ln \ln n = o(\alpha(n))$,

and using either of the following method :

- ⋆ comparative method (2.17)
- ⋆ Reversed comparative method in the orthonormal case (2.18)
- ⋆ Adapted reversed comparative method (2.19)
- ⋆ Descending comparative method (part 2.5.5)

Recall the notations of part 5.1. Theorems of the previous section as well as control (3.2), easily transposable to other selection procedures listed above, insure that in any of those cases we get

$$\begin{cases} \widehat{S} \to S^\star & \text{a.s.} \\ \mathbb{P}(\widehat{S} = S) \to 0 & \text{for any} \quad S \in \mathcal{F}_2 \setminus \{S^\star\} \\ \mathbb{P}(\widehat{S} = S) = O\left(\exp(-cn)\right) & \text{for any} \quad S \in \mathcal{F}_1 \end{cases} \tag{5.9}$$

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

$$R(n) = \mathbb{E}\left[\left\|f^\star - \widetilde{f}\right\|_w^2\right] = \sum_{S \subset [\![1,d]\!]} \mathbb{E}\left[\left\|f^\star - \widetilde{f}_S\right\|_w^2 \mathbb{1}_{\widehat{S}=S}\right] =: \sum_{S \subset [\![1,d]\!]} R(n, S). \tag{5.10}$$

Now distinguish between two cases.

5.2.3. *Asymptotics of the case $S \in \mathcal{F}_2$*

Computations similar to (5.4) and (5.5) where $\widehat{f}_S$ is replaced by $\widetilde{f}_S$ reduce $R(n, S)$ to

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

Remark that

$$\begin{aligned} Z_n &:= ng^T A_S g \, \mathbb{1}_{\{\|nM_S^{-1}\|<C\}} \\ &= \frac{1}{n} \sum_{i,j=1}^{n} g_i g_j \left(V_S^T n M_S^{-1} M_{w,S} n M_S^{-1} V_S\right)_{i,j} \mathbb{1}_{\{\|nM_S^{-1}\|<C\}} \end{aligned} \tag{5.11}$$

has an $L^2(\Omega)$ norm satisfying $\mathbb{E}\, Z_n^2 = O(1)$.

Let us begin by $S = S^\star$. Then (5.9) yields $\lim \mathbb{E}\left[Z_n \mathbb{1}_{\{\widehat{S}=S^\star\}}\right] = \lim \mathbb{E} Z_n$ which is $\sigma^2|S^\star|$ by dominated convergence. This, with (2.10), gives

$$R(n, S^\star) \sim \frac{\sigma^2|S^\star|}{n}. \tag{5.12}$$

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 (5.11) and (5.9). Consequently,

$$nR(n, S) \to 0, \text{ for any } S \in \mathcal{F}_2 \setminus \{S^\star\}. \tag{5.13}$$

5.2.4. *Asymptotics of 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 (2.8). Recall (2.7) and (5.9) in our present case $S \in \mathcal{F}_1$ to obtain

$$nR(n, S) \to 0, \text{ for any } S \in \mathcal{F}_1. \tag{5.14}$$

5.2.5. *Summary*

Plugging (5.10), (5.12), (5.13) and (5.14) together we get the following

Theorem 5.1. *Assume that the penalty of our IC (2.15) satisfies*

(*i*) $\alpha(n) = o(n)$
(*ii*) $\ln \ln n = o(\alpha(n))$,

and that either of the following method has been used to determine $\widehat{S}$.

- ⋆ *comparative method (2.17)*
- ⋆ *Reversed comparative method in the orthonormal case (2.18)*
- ⋆ *Adapted reversed comparative method (2.19)*
- ⋆ *Descending comparative method (part 2.5.5)*

Then the estimation of $f^\star$ by $\widetilde{f} = \widetilde{f}_{\widehat{S}}$ defined in (2.6) presents a risk $R(n)$ equivalent to the oracle risk $\mathcal{O}R(n, S^\star)$ (5.8) in the sense that

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

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

In this section, we no longer assume that $f^\star$ lives in F. Our main purpose is to give theorem 6.1. This theorem presents an oracle inequality on the risk achieved by the estimator of $f^\star$ resulting from the use of an information criterion of the form (2.15) along with the (fast) descending comparative method (part 2.5.5).

6.1. **Preliminary result**

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

$$\mathcal{O}_{\mathcal{F}}(f^\star) = \min_{S\in\mathcal{F}} \left\{ d^2(f^\star, F_S) + \sigma^2|S|/n \right\} =: \min_{S\in\mathcal{F}} R^\star(S) \tag{6.1}$$

Note that the quantity $R^\star(S)$ does not represent the risk resulting from the estimation of $f^\star$ within F_S which are expressed in (5.6) and (5.7). Actually, $R^\star(S)$ is the risk resulting from such an estimation in the case where the x_i's are deterministic as in (5.3). In the sequel, as Baraud in [3], we work with quantities $R^\star(S)$.

The oracle (6.1) is the minimum "risk" the user could achieve by selecting the support $S_{\mathcal{O}}$ in $\mathcal{F}$ that realizes the minimum. However, as the name oracle implies, that quantity as well as $S_{\mathcal{O}}$ is unavalaible to the user. Also note that, even though $f^\star$ would live in F, there is no reason why $S_{\mathcal{O}}$ would equal $S^\star$.

From now on, we choose a penalty of the form

$$\alpha(n) = (1+\theta)\sigma^2|S|,\ \theta > 0. \tag{6.2}$$

Let us assume briefly that this user has chosen the global method (2.16). The only thing he knows is what his criterion has found is the best support, namely

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

Baraud shows in [3] that in our setting and by using the penalty (6.2) in his criterion (2.15), the user did not take too much risks in the sense that

$$\mathbb{E}\left[\left\| f^\star - \widetilde{f}_{\widehat{S}} \right\|_w^2 \right] \leq C\mathcal{O}_{\mathcal{F}}(f^\star). \tag{6.3}$$

where C is a constant depending on θ appearing in the penalty (6.2) but neither on n nor on $f^\star$.

Now, as stressed in part 2.6, the global method has exponential complexity. Here, if one wanted to be able to select any support, he would have computed 2^d criteria. Our aim in the sequel is to show that the descending method, that has polynomial complexity, also gives an oracle inequality of the type (6.3).

6.2. **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 we set $S^{\star(0)} = [\![1,d]\!]$ then, when $S^{\star(k)}$ is defined, we set

$$S^{\star(k+1)} = \text{Argmin}\left(R^\star(S),\ S\subset S^{\star(k)},\ |S| = d-(k+1) \right) \tag{6.4}$$

where the function $R^\star$ is defined in 6.1.

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. Let us denote by $k^\star$ the first index such that

$$R^\star(S^{\star(k^\star-1)}) > R^\star(S^{\star(k^\star)}) \text{ and } R^\star(S^{\star(k^\star+1)}) \geq R^\star(S^{\star(k^\star)}). \tag{6.5}$$

In other terms, $S^{\star(k^\star)}$ is the first support that does not 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 6.1 with $\mathcal{F} = \mathcal{P}([\![1,d]\!])$, but among a smaller, nested, family of risks.

6.3. **The oracle inequality**

Let us give the main result.

Theorem 6.1. *Consider an information criteria of the form (2.15) 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 2.5.5, one produces $\widetilde{f}_{S^{(k_f)}}$ as an estimation of the unknown function $f^\star$. The risk of such an estimation satisfies

$$\mathbb{E}\left[\left\|f^\star - \widetilde{f}_{S^{(k_f)}}\right\|_w^2\right] \leq C.R^\star(S^{\star(k^\star)}) + r_n \tag{6.6}$$

where C is a constant depending on θ but neither on n nor $f^\star$; $R^\star(S^{\star(k^\star)})$ is the nested oracle risk defined in (6.5) and r_n is a deterministic term satisfying $r_n = O(\exp(-an))$ with $a > 0$.

Proof : the deterministic family of supports (6.4) is related to the random family $S^{(k)}$ produced by the descending comparative method in part 2.5.5. Recall that this method stops at a random step k_f and thus produces only supports $S^{(k)}$, $k = 0, \ldots, k_f$. One would like the descending comparative method to choose "good" supports and stop at the "right" step in the sense that

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

Equations (6.7) and (6.9) that we justify now show that this happens except on a set of exponentially decreasing probability.

For $1 \leq k \leq d$ let us set the event

$$A_k = \left\{k_f \geq k-1, S^{(k-1)} = S^{\star(k-1)}, \ldots, S^{(0)} = S^{\star(0)}\right\}.$$

Firstly, we study the probability

$$\mathbb{P}_{>k^\star} := \mathbb{P}\left(k_f > k^\star \mid A_{k^\star+1}\right)$$

where k_f is the random step where the descending comparative method stops and $k^\star$ is the deterministic step defined by (6.5). This is the probability that the method does not stop after the oracle step $k^\star$ when it has chosen all good supports up to that step. We have

$$\mathbb{P}_{>k^\star} \leq \sum_S \mathbb{P}\left(\frac{1}{n}\mathrm{IC}(S) - \sigma^2 \leq \frac{1}{n}\mathrm{IC}(S^{\star(k^\star)}) - \sigma^2 \mid A_{k^\star+1}\right)$$

where the sum is extended to all supports $S \subset S^{(k^\star)} = S^{\star(k^\star)}$ with cardinal $|S| = d - k^\star - 1$. Because of the definition of $k^\star$, any of those supports satisfies $R^\star(S) > R^\star(S^{\star(k^\star)})$. We choose

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

where the min is taken among the same set of supports. Then we have

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

Because of the form (6.2) of the penalty, expressions in those probabilities reduce to

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

Choosing n large enough to make $\theta\sigma^2|S|/n < \varepsilon/4$ and applying control (2.14) yields

$$\mathbb{P}\left(k_f > k^\star \mid A_{k^\star+1}\right) = O\left(\exp\left(-\frac{\varepsilon^2 n}{64V_{max}}\right)\right). \tag{6.7}$$

We are now interested in the probability that the method fails to select the good support $S^{\star(k)}$ when it has done well up to step $k-1$:

$$P_k := \mathbb{P}\left(S^{(k)} \neq S^{\star(k)} \mid A_k\right).$$

Similarily, we write

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

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

$$\varepsilon_k = \frac{1}{2}\min_S(R^\star(S) - R^\star(S^{\star(k)})) > 0, \tag{6.8}$$

the minimun being taken among the same set of supports. Let us denote by $\varepsilon > 0$ the smallest of the ε_k's satisfying (6.8), $k = 1, \ldots, d$.

Again because of the penalty term (6.2), expressions simplify to

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

Now it suffices to choose n large enough to make $\theta\sigma^2|S|/n < \varepsilon/64$ and apply control (2.14) to have

$$\mathbb{P}\left(S^{(k)} \neq S^{\star(k)} \mid A_k\right) = O\left(\exp\left(-\frac{\varepsilon^2 n}{64V_{max}}\right)\right). \tag{6.9}$$

Now that we have controls (6.7) and (6.9) we will apply Baraud's result. 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 d_w^2. 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 result of Baraud (6.3), more precisely : equation (15) following theorem 1.1 in [3], ensures that

$$\mathbb{E}\left[d_w^2 \middle| k_f = k^\star,\, A_{k^\star}\right] \leq C.R^\star(S^{\star(k^\star)}),$$

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 (6.6):

$$\mathbb{E}\left[d_w^2 \middle| A_{k^\star}\right] \quad \leq \quad \mathbb{E}\left[d_w^2 \middle| k_f = k^\star,\, A_{k^\star}\right] + \mathbb{E}\left[d_w^2 \mathbb{1}_{k_f > k^\star} \middle| A_{k^\star}\right]$$

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

$$\mathbb{E}\left[d_w^2 \middle| A_{k^\star - 1}\right] \quad \leq \quad \mathbb{E}\left[d_w^2 \middle| A_{k^\star}\right] + \mathbb{E}\left[d_w^2 \mathbb{1}_{S^{(k^\star)} \neq S^\star(k^\star)} \middle| A_{k^\star - 1}\right]$$

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

$$\begin{aligned}\mathbb{E}\left[d_w^2\right] \quad \leq \quad & C.R^\star(S^{\star(k^\star)}) + \\ & + \mathbb{E}\left[d_w^2 \mathbb{1}_{k_f > k^\star} \middle| A_{k^\star}\right] \\ & + \sum_{k=1}^{k^\star - 1} \mathbb{E}\left[d_w^2 \mathbb{1}_{S^{(k)} \neq S^\star(k)} \middle| A_k\right].\end{aligned}$$

It suffices to apply Hölder's inequality (2.8) along with controls (6.7) and (6.9) in the remainder terms to obtain the theorem. □

References

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

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

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

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

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

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

[7] Gwénaëlle 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] Abdelaziz El Matouat and Marc Hallin. Order selection, stochastic complexity and Kullback-Leibler information. 115:291–299, 1996.

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

[10] Pascal 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.

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

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

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

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

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

[16] Jorma Rissanen, Terry P. Speed, and Bin Yu. Density estimation by stochastic complexity. *IEEE Transactions on Information Theory*, 38(2):315–323, 1992.

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