ESAIM: PS 15 (2011) 1–29
DOI: 10.1051/ps/2009007

ESAIM: Probability and Statistics
www.esaim-ps.org

ESTIMATING A DISCRETE DISTRIBUTION *VIA* HISTOGRAM SELECTION

Nathalie Akakpo[1]

Abstract. Our aim is to estimate the joint distribution of a finite sequence of independent categorical variables. We consider the collection of partitions into dyadic intervals and the associated histograms, and we select from the data the best histogram by minimizing a penalized least-squares criterion. The choice of the collection of partitions is inspired from approximation results due to DeVore and Yu. Our estimator satisfies a nonasymptotic oracle-type inequality and adaptivity properties in the minimax sense. Moreover, its computational complexity is only linear in the length of the sequence. We also use that estimator during the preliminary stage of a hybrid procedure for detecting multiple change-points in the joint distribution of the sequence. That second procedure still satisfies adaptivity properties and can be implemented efficiently. We provide a simulation study and apply the hybrid procedure to the segmentation of a DNA sequence.

Mathematics Subject Classification. 62G05, 62C20, 41A17.

Received August 1st, 2008. Revised April 21, 2009.

1. Introduction

Let $Y_1, Y_2, \ldots, Y_n$ be independent random variables taking values in the finite set $\{1, \ldots, r\}$, where r is an integer and $r \geq 2$. Let s be the joint distribution of $(Y_1, Y_2, \ldots, Y_n)$, that we consider as the $\mathbb{R}^r$-valued function defined on $\{1, \ldots, n\}$ with lth coordinate function

$$i \in \{1, \ldots, n\} \mapsto \mathbb{P}(Y_i = l),$$

for $l = 1, \ldots, r$. The aim of this paper is to study a nonparametric estimator of the distribution s. References treating about this problem are so scarce that we can only cite three of them. Aerts and Veraverbeke [1] propose a kernel estimator, whose convergence rate is given under a Lipschitz regularity condition. More recently, Lebarbier and Nédélec [20] and then Durot *et al.* [15] have studied procedures based on the model selection principle introduced by Barron *et al.* [3]. Thus, all their results are nonasymptotic. In both cases, a family of linear spaces of real-valued functions defined on $\{1, \ldots, n\}$ is given, and the procedures allow to select from the data, by minimizing some penalized criterion, one space among that family in which all the coordinate functions of s are estimated. The choice of the penalty is supported by an oracle-type inequality. Lebarbier and Nédélec [20] consider two different penalized criteria, one based on least-squares, the other one on maximum-likelihood, and spaces of piecewise constant functions. Durot *et al.* [15] consider only a penalized

Keywords and phrases. Adaptive estimator, approximation result, categorical variable, change-point detection, minimax estimation, model selection, nonparametric estimation, penalized least-squares estimation.

[1] Laboratoire de Probabilités et Statistiques, Université Paris Sud XI, Bâtiment 425, 91405 Orsay Cedex, France; nathalie.akakpo@math.u-psud.fr

Article published by EDP Sciences

least-squares criterion, but provide an oracle-type inequality that is valid for almost all finite families of linear spaces. Moreover, they are particularly interested in three families of spaces. The so-called exhaustive indicator strategy corresponds with the family made up of all spaces of functions piecewise constant on some partition of $\{1,\dots,n\}$, a family already encountered in [20]; the exhaustive Haar and non-exhaustive Haar (or neH) strategies are based on families made up of spaces generated by some Haar wavelets. In these three cases, the resulting estimator is proved to have adaptivity properties. Due to the richness of the underlying families of spaces, both exhaustive strategies yield estimators that only satisfy an oracle-type inequality up to a $\ln(n)$ factor, but the non-exhaustive one does not have the same drawback. Besides, implementing the first strategy requires $\mathcal{O}(n^3)$ computations, against only $\mathcal{O}(n\ln(n))$ for the other two.

In this paper, we study the penalized least-squares estimator defined as in [15] but based on a fourth family of linear spaces: in our case, each space is composed of functions piecewise constant on a partition of $\{1,\dots,n\}$ into *dyadic* intervals. Thus, we will refer to our estimator as the d-estimator. The collection of linear spaces we consider has been chosen for its potential qualities of approximation, as suggested by approximation results for real-valued functions due to DeVore and Yu [14] and DeVore (*cf.* [6]). Adapting the proofs to our framework, we prove that our collection of spaces has indeed good approximation qualities with respect to $\mathbb{R}^r$-valued functions defined on $\{1,\dots,n\}$ that either belong to Besov bodies – some discrete analogues of balls in a Besov space – or have bounded variation. On the other hand, the number of spaces per dimension is low enough to yield an oracle-type inequality with no extra logarithmic factor. The conjunction of both properties of our collection allows to prove adaptivity results in the minimax sense. From a theoretical point of view, the d-estimator thus satisfies properties similar to those of the neH-estimator, and is also proved to be adaptive for functions with bounded variation. Moreover, the d-estimator can be implemented with only $\mathcal{O}(n)$ computations. Notice that a similar collection of linear spaces has lately been used by Birgé [5,6] and Baraud and Birgé [2] for estimation by model selection in various statistical frameworks.

As an application of our estimation procedure, we address the problem of multiple change-point detection in the distribution s. Our aim is then to estimate s by a function that is piecewise constant on some partition of $\{1,\dots,n\}$ with a number of intervals much smaller than n. That issue has attracted much attention due to its application to the segmentation of DNA sequences into regions of homogeneous composition (*cf.* the review [8] by Braun and Müller). Owing to the length of sequences such as DNA ones, a special attention must be paid to the computational complexity of the statistical procedures. Braun *et al.* [9] prove consistency results for the estimation of the change-points and the number of change-points when using a penalized quasi-deviance criterion, but their estimator suffers from a heavy computational complexity, of order $\mathcal{O}(n^3)$. The two-stage procedure proposed by Gey and Lebarbier [17] in a Gaussian regression framework can be adapted to the framework considered here (*cf.* [19], Chap. 7). The preliminary stage uses CART algorithm to select a partition. In order to reduce the size of the partition, the second stage consists in selecting a partition among the rougher partitions built on the previous one, by minimizing a penalized least-squares criterion. In the best case, the number of computations falls down to only $\mathcal{O}(n\ln(n))$ for the first stage of the procedure. Last, a few linear time procedures exist, such as the one proposed by Fu and Curnow [16] (*cf.* [11] for the implementation) and the one studied by Szpankowski *et al.* [23]. We propose in this paper a hybrid procedure similar to that of [17], where the first stage consists this time in selecting a partition into dyadic intervals. In practice, our hybrid procedure can be implemented quite efficiently. Moreover, unlike the CART-based hybrid estimator, our hybrid estimator is proved to enjoy some adaptivity properties, which are similar to those of the d-estimator, up to a multiplicative constant. Notice that, contrary to [9], our aim is not to detect all the change-points, but only the most relevant ones.

The paper is organized as follows. In Section 2, we describe the statistical framework and introduce notation used throughout the paper. The next section is devoted to the theoretical study of the d-estimator . Then, we present the subsequent hybrid procedure. The performance of these procedures are illustrated in Section 5 through a simulation study. In particular, we discuss there the practical choice of the penalties constants. Besides, we compare the d-estimator with the neH-estimator introduced in [15], and apply the hybrid procedure

to a DNA sequence. The paper ends with the proof of the approximation result needed to derive one of the adaptivity properties.

2. Framework and notation

2.1. **Framework**

We observe n independent random variables $Y_1, \dots, Y_n$ defined on the same probability space $(\Omega, \mathcal{A}, \mathbb{P})$ and with values in $\{1, \dots, r\}$, where r is an integer and $r \geq 2$. We assume that n is a power of 2, $n \geq 2$, and write $n = 2^N$. The distribution of the n-uple $(Y_1, \dots, Y_n)$ is represented by the $r \times n$ matrix s whose ith column is

$$s_i = \big(\mathbb{P}(Y_i = 1) \dots \mathbb{P}(Y_i = r)\big)^T, \text{ for } i = 1, \dots, n.$$

Observing $(Y_1, \dots, Y_n)$ is equivalent to observing the random $r \times n$ matrix X whose ith column is

$$X_i = \big(\mathbb{1}_{Y_i=1} \dots \mathbb{1}_{Y_i=r}\big)^T, \text{ for } i = 1, \dots, n.$$

It should be noted that the distribution s to estimate is in fact the expectation of X.

2.2. **Notation**

All along the paper, we identify real-valued functions defined on $\{1, \dots, n\}$ with $\mathbb{R}^n$-vectors, so that $u = (u_1 \dots u_n) \in \mathbb{R}^n$ represents the function $u : i \in \{1, \dots, n\} \mapsto u_i$. In particular, for any subset I of $\{1, \dots, n\}$, we call indicator function of I, and denote by $\mathbb{1}_I$, the $\mathbb{R}^n$-vector whose ith coordinate is equal to 1 if $i \in I$, and null otherwise. In the same way, we identify $\mathbb{R}^r$-valued functions defined on $\{1, \dots, n\}$ with elements of $\mathscr{M}(r, n)$, the set of real matrices with r rows and n columns. Given an element $t \in \mathscr{M}(r, n)$, we denote by $t^{(l)}$ its lth row and by t_i its ith column. Thus $t \in \mathscr{M}(r, n)$ represents the function, also denoted by t, defined on $\{1, \dots, n\}$, whose value in i is the $\mathbb{R}^r$-vector t_i, while $t^{(1)}, \dots, t^{(r)}$ are the coordinate functions of t.

The space $\mathscr{M}(r, n)$ is endowed with the inner product defined by

$$\langle t, u \rangle = \sum_{i=1}^{n} \sum_{l=1}^{r} t_i^{(l)} u_i^{(l)}.$$

That product is linked with the standard inner products on $\mathbb{R}^r$ and $\mathbb{R}^n$, denoted respectively by $\langle ., . \rangle_r$ and $\langle ., . \rangle_n$, by the relations

$$\langle t, u \rangle = \sum_{i=1}^{n} \langle t_i, u_i \rangle_r = \sum_{l=1}^{r} \langle t^{(l)}, u^{(l)} \rangle_n.$$

The norms induced by these products on $\mathscr{M}(r, n)$, $\mathbb{R}^r$ and $\mathbb{R}^n$ are respectively denoted by $\|.\|$, $\|.\|_r$ and $\|.\|_n$. Another norm on $\mathscr{M}(r, n)$ appearing in this paper is

$$\|t\|_\infty := \max\big\{|t_i^{(l)}|; 1 \leq i \leq n, 1 \leq l \leq r\big\}.$$

Let us now define some subsets of $\mathscr{M}(r, n)$ of special interest. The set composed of the $r \times n$ matrices whose columns are probability distributions on $\{1, \dots, r\}$ is denoted by $\mathscr{P}$. Given a linear subspace S of $\mathbb{R}^n$, the notation $\mathbb{R}^r \otimes S$ stands for the linear subspace of $\mathscr{M}(r, n)$ composed of the matrices whose rows all belong to S.

When the distribution of $(Y_1, \dots, Y_n)$ is given by s, we denote respectively by $\mathbb{P}_s$ and $\mathbb{E}_s$ the underlying probability distribution on $(\Omega^{\otimes n}, \mathcal{A}^{\otimes n})$ and the associated expectation.

Last, in the many inequalities we shall encounter, the letters $C, C_1, c_1, \dots$ stand for positive constants. Sometimes, their dependence on one or several parameters will be indicated. For instance, the notation $C(\alpha, p)$ means that C only depends on α and p. The only constant whose value is allowed to change from one line to another is denoted by C, with no index.

3. The d-estimator

We study in this section the d-estimator of the distribution s, thus called because it takes values in the set of piecewise constant functions on some partition of $\{1,\dots,n\}$ into *dyadic* intervals. We begin with the definition of the estimator, explain the underlying model selection principle and justify the form of the involved penalty thanks to [15]. Then, we present the main result of this paper, about the adaptivity of the d-estimator. They greatly rely on an approximation result that will be proved later in the article. Last, we describe the algorithm used to implement that procedure and give its computational complexity.

3.1. Definition of the d-estimator

A partition of $\{1,\dots,n\}$ into dyadic intervals is a partition of $\{1,\dots,n\}$ into sets of the form $\big\{kn2^{-j}+1,\dots,(k+1)n2^{-j}\big\}$, where $j\in\{0,\dots,N\}$ is allowed to change from one interval of the partition to another, and $k\in\{0,\dots,2^j-1\}$. We denote by $\mathcal{M}$ the family of all such partitions of $\{1,\dots,n\}$. We consider the collection of linear spaces of the form $\mathbb{R}^r\otimes S_m$, where $m\in\mathcal{M}$ and S_m is the linear subspace of $\mathbb{R}^n$ generated by the indicator functions $\{\mathbb{1}_I, I\in m\}$. In the sequel, the term 'model' refers indifferently to such a subspace of $\mathscr{M}(r,n)$ or to the associated partition in $\mathcal{M}$. For all $m\in\mathcal{M}$, the least-squares estimator of s in $\mathbb{R}^r\otimes S_m$ is defined by

$$\hat{s}_m=\operatorname*{argmin}_{t\in\mathbb{R}^r\otimes S_m}\|X-t\|^2.$$

Over each interval $I\in m$, $\hat{s}_m$ is constant and equal to the mean of the $\mathbb{R}^r$-vectors $(X_i)_{i\in I}$.

Ideally, we would like to choose a model among the collection $\mathcal{M}$ such that the risk of the associated estimator is minimal. However, determining such a model requires the knowledge of s. Therefore the challenge is to define a procedure $\hat{m}$, based solely on the data, that selects a model for which the risk of $\hat{s}_{\hat{m}}$ almost reaches the minimal one. In other words, the estimator $\hat{s}_{\hat{m}}$ should satisfy a so-called oracle inequality

$$\mathbb{E}_s\big[\|s-\hat{s}_{\hat{m}}\|^2\big]\le C\inf_{m\in\mathcal{M}}\mathbb{E}_s\big[\|s-\hat{s}_m\|^2\big].$$

Besides, as often, the risk of each estimator $\hat{s}_m$ breaks down into an approximation error and an estimation error roughly proportional to the dimension of the model. Indeed, for all $m\in\mathcal{M}$, the estimator $\hat{s}_m$ satisfies

$$\|s-s_m\|^2+\big(1-\|s\|_\infty\big)D_m\le\mathbb{E}_s\big[\|s-\hat{s}_m\|^2\big]\le\|s-s_m\|^2+\Big(1-\frac{1}{r}\Big)D_m, \tag{3.1}$$

where s_m is the orthogonal projection of s on $\mathbb{R}^r\otimes S_m$ and D_m is the dimension of S_m (*cf.* [15], proof of Cor. 1). Reaching the minimal risk among the estimators of the collection thus amounts to realizing the best trade-off between the approximation error and the dimension of the model, which vary in opposite ways. Therefore, we consider the procedure

$$\hat{m}=\operatorname*{argmin}_{m\in\mathcal{M}}\big\{\|X-\hat{s}_m\|^2+\mathrm{pen}(m)\big\},$$

where $\mathrm{pen}:\mathcal{M}\to\mathbb{R}^+$ is called penalty function. The d-estimator $\tilde{s}$ of s is then defined as

$$\tilde{s}=\hat{s}_{\hat{m}}.$$

Our choice of penalty, that relies on results proved in [15], is justified by an oracle inequality, up to a quantity that depends on $\|s\|_\infty$ (*cf.* inequality (3.4) below).

Proposition 3.1. *Let* $\mathrm{pen}:\mathcal{M}\to\mathbb{R}^+$ *be a penalty of the form*

$$\mathrm{pen}(m)=c_0D_m, \tag{3.2}$$

where, for $m \in \mathcal{M}$, D_m is the dimension of S_m. If c_0 is positive and large enough, then

$$\mathbb{E}_s\big[\|s - \tilde{s}\|^2\big] \leq C(c_0) \inf_{m \in \mathcal{M}} \big\{\|s - s_m\|^2 + D_m\big\}. \tag{3.3}$$

Moreover, if $\|s\|_\infty < 1$, then

$$\mathbb{E}_s\big[\|s - \tilde{s}\|^2\big] \leq C(c_0)(1 - \|s\|_\infty)^{-1} \inf_{m \in \mathcal{M}} \mathbb{E}_s\big[\|s - \hat{s}_m\|^2\big]. \tag{3.4}$$

Proof. For all $1 \leq D \leq n$, we introduce the subcollection of models of dimension D:

$$\mathcal{M}_D = \{m \in \mathcal{M} \text{ s.t. } D_m = D\}.$$

In order to evaluate the cardinal of $\mathcal{M}_D$, let us describe $\mathcal{M}$ in a more constructive way. Let $\mathcal{T}$ be the complete binary tree with root $(0,0)$ such that:

- for all $j \in \{1, \ldots, N\}$, the nodes at level j are indexed by the elements of the set $\Lambda(j) = \{(j,k), k = 0, \ldots, 2^j - 1\}$;
- for all $j \in \{0, \ldots, N-1\}$ and all $k \in \{0, \ldots, 2^j - 1\}$, the left branch that stems from node (j,k) leads to node $(j+1, 2k)$, and the right one, to node $(j+1, 2k+1)$.

The node set of $\mathcal{T}$ is $\mathcal{N} = \cup_{j=0}^{N} \Lambda(j)$, where $\Lambda(0) = \{(0,0)\}$. The dyadic intervals of $\{1, \ldots, n\}$ are the sets

$$I_{(j,k)} = \big\{k2^{N-j} + 1, \ldots, (k+1)2^{N-j}\big\}$$

indexed by the elements of $\mathcal{N}$. Hence we deduce a one-to-one correspondence between the partitions of $\{1, \ldots, n\}$ that belong to $\mathcal{M}$ and the subsets of $\mathcal{N}$ composed of the leaves of any complete binary tree resulting from an elagation of $\mathcal{T}$. The cardinal of $\mathcal{M}_D$ is thus equal to the number of complete binary trees with D leaves resulting from an elagation of $\mathcal{T}$. So it is given by the Catalan number $D^{-1}\binom{2(D-1)}{D-1}$ and upper-bounded by 4^D. Therefore, choosing for instance $L = \ln(8)$, we get

$$\sum_{m \in \mathcal{M}} \exp(-LD_m) = \sum_{D=1}^{n} |\mathcal{M}_D| \exp(-LD) \leq 1.$$

Inequality (3.3) thus follows from Theorem 1 in [15]. Inequality (3.4) results from the upper-bound (3.3) and the lower-bound given in (3.1). □

From now on, we will always assume that the d-estimator derives from a penalty of the form $\mathrm{pen}(m) = c_0 D_m$, where the constant c_0 is positive and large enough to yield an oracle-type inequality. Choosing in practice an adequate value of c_0 is an issue that will be treated in Section 5. By way of comparison, let us mention that the *neH*-procedure studied in [15] satisfies the same kind of oracle-type inequality (*cf.* [15], Prop. 3). But the similar procedure based on the exhaustive collection of partitions of $\{1, \ldots, n\}$ only satisfies an oracle-type inequality such as (3.4) within a $\ln(n)$ factor, owing to the greater number of models per dimension (*cf.* [15], Prop. 1).

Last, notice that $\tilde{s}$ does not necessarily belong to $\mathscr{P}$. Nevertheless, since the vector $(1 \ldots 1)$ belongs to any S_m, for $m \in \mathcal{M}$, the elements in a same row of $\tilde{s}$ sum up to 1. In order to get an estimator of s with values in $\mathscr{P}$, we may consider the orthogonal projection of $\tilde{s}$ on the closed convex $\mathscr{P}$, whose risk is even smaller than that of $\tilde{s}$.

3.2. **Adaptivity of the d-estimator**

Though the oracle-type inequality (3.4) ensures that, under a minor constraint on s, the estimator $\tilde{s}$ is almost as good as the best estimator in the collection $\{\hat{s}_m\}_{m \in \mathcal{M}}$, it does not allow to compare $\tilde{s}$ with other estimators of s. Therefore, we now pursue the study of $\tilde{s}$ adopting a minimax point of view. We consider a large family

of subsets of $\mathscr{P}$, to be defined in the next paragraph. Let us denote by $\mathcal{S}$ some subset in that family. Our aim is to compare the maximal risk of $\tilde{s}$ when s belongs to $\mathcal{S}$ to the minimax risk over $\mathcal{S}$. We may rewrite the upper-bound (3.3) for the risk of $\tilde{s}$ as

$$\mathbb{E}_s\big[\|s-\tilde{s}\|^2\big] \leq C(c_0) \inf_{1\leq D\leq n}\Big\{ \inf_{m\in\mathcal{M}_D} \|s-s_m\|^2 + D\Big\}, \tag{3.5}$$

where we recall that $\mathcal{M}_D = \{m\in\mathcal{M} \text{ s.t. } D_m = D\}$ and s_m is the orthogonal projection of s on $\mathbb{R}^r\otimes S_m$. Thus, the approximation qualities of our family of models with respect to each subset $\mathcal{S}$ remain to be evaluated. More precisely, for each subset $\mathcal{S}$, and each dimension D, we shall provide upper-bounds for the approximation error $\inf_{m\in\mathcal{M}_D}\|s-s_m\|^2$ when $s\in\mathcal{S}$.

On the one hand, we consider subsets of $\mathscr{P}$ introduced in [15], whose definition is inspired from the characterization in terms of wavelet coefficients of balls in Besov spaces. In order to define them, we equip $\mathbb{R}^n$ with an orthonormal wavelet basis, the Haar basis.

Definition 3.2. Let $\varphi:\mathbb{R}\to\{-1,1\}$ be the function with support $(0,1]$ that takes value 1 on $(0,1/2]$ and -1 on $(1/2,1]$. Let $\Lambda = \cup_{j=-1}^{N-1}\Lambda(j)$, where $\Lambda(-1)=\{(-1,0)\}$ and

$$\Lambda(j) = \{(j,k), k=0,\dots,2^j-1\}, \text{ for } j=0,\dots,N-1.$$

If $\lambda=(-1,0)$, ϕ_λ is the $\mathbb{R}^n$-vector whose coordinates are all equal to $1/\sqrt{n}$.

If $\lambda=(j,k)$, where $j\neq -1$ and $k\in\Lambda(j)$, ϕ_λ is the $\mathbb{R}^n$-vector whose ith coordinate is

$$\phi_{\lambda\, i} = \frac{2^{j/2}}{\sqrt{n}}\varphi\Big(2^j\frac{i}{n}-k\Big), \text{ for } i=1,\dots,n.$$

The functions $\{\phi_\lambda\}_{\lambda\in\Lambda}$ are called the Haar functions. They form an orthonormal basis of $\mathbb{R}^n$ called the Haar basis.

Any element $t\in\mathscr{M}(r,n)$ can be decomposed into

$$t = \sum_{j=-1}^{N-1}\sum_{\lambda\in\Lambda(j)}\beta_\lambda\phi_\lambda$$

where, for all $\lambda\in\Lambda$, β_λ is the column-vector in $\mathbb{R}^r$ whose lth coefficient is $\beta_\lambda^{(l)} = \langle t^{(l)},\phi_\lambda\rangle_n$, for $l=1,\dots,r$. So, we improperly refer to the β_λ's as the wavelet coefficients of t. Besov bodies are then defined as follows.

Definition 3.3. Let $\alpha>0$, $p>0$ and $R\geq 0$. The set composed of all the elements $t\in\mathscr{M}(r,n)$ such that

$$\frac{1}{\sqrt{n}}\left(\sum_{j=0}^{N-1}2^{jp(\alpha+1/2-1/p)}\sum_{\lambda\in\Lambda(j)}\|\beta_\lambda\|_r^p\right)^{1/p}\leq R,$$

where, for $l=1,\dots,r$, $\beta_\lambda^{(l)} = \langle t^{(l)},\phi_\lambda\rangle_n$, is denoted by $\mathscr{B}(\alpha,p,R)$ and called a Besov body. The set of all the elements of $\mathscr{P}$ that belong to $\mathscr{B}(\alpha,p,R)$ is denoted by $\mathscr{BP}(\alpha,p,R)$.

We also consider subsets of $\mathscr{P}$ whose definition is inspired from functions of bounded α-variation.

Definition 3.4. Let $\alpha>0$ and $R\geq 0$. For $t\in\mathscr{M}(r,n)$, let

$$V_\alpha(t) = \sup_{1\leq i\leq n-1}\ \sup_{\substack{x_0<\dots<x_i\\ \text{s.t. } 1\leq x_0<x_i\leq n}}\left\{\sum_{j=1}^{i}\|t_{x_j}-t_{x_{j-1}}\|_r^{1/\alpha}\right\}^\alpha.$$

The set composed of all the elements $t \in \mathscr{M}(r,n)$ such that $V_\alpha(t) \leq R$ is denoted by $\mathscr{V}(\alpha,R)$. The set of all the elements of $\mathscr{P}$ that belong to $\mathscr{V}(\alpha,R)$ is denoted by $\mathscr{V}\mathscr{P}(\alpha,R)$.

Notice that, for all $t \in \mathscr{M}(r,n)$, when $\alpha \geq 1$,

$$V_\alpha(t) = \left\{ \sum_{i=2}^{n} \|t_i - t_{i-1}\|_r^{1/\alpha} \right\}^\alpha$$

so that $V_\alpha(t)$ may be interpreted as the $\ell_{1/\alpha}$-norm of the "jumps" of t.

For a wide range of values of the parameters (α,p,R) or (α,R), we are able to bound the approximation errors appearing in (3.5) uniformly over $\mathscr{B}\mathscr{P}(\alpha,p,R)$ and $\mathscr{V}\mathscr{P}(\alpha,R)$.

Theorem 3.5. *Let $p \in (0,2]$, $\alpha > 1/p - 1/2$ and $R \geq 0$. For all $s \in \mathscr{B}\mathscr{P}(\alpha,p,R)$ and all $D \in \{1,\dots,n\}$, there exists a partition $m \in \mathcal{M}$ such that $D_m = D$ and*

$$\|s - s_m\|^2 \leq C(\alpha,p) n R^2 D^{-2\alpha}.$$

That result will be proved in Section 6.

Theorem 3.6. *Let $\alpha > 0$ and $R \geq 0$. Let $k_1(\alpha) = \big(1 - 2^{-(1+2\alpha)/(2\alpha)}\big)/\big(1 - 2^{-1/(2\alpha)}\big)$. For all $s \in \mathscr{V}\mathscr{P}(\alpha,R)$ and all $j \in \{0,\dots,N\}$, there exists a partition $m \in \mathcal{M}$ such that $1 \leq D_m \leq k_1(\alpha) 2^j$ and*

$$\|s - s_m\|^2 \leq C(\alpha) n R^2 2^{-2\alpha j}.$$

Proof. The proof of Proposition 3 in [6] can be readily adapted to our framework, whatever $\alpha > 0$. In the proof of that proposition, the assumption $\alpha \in (0,1]$ is only used to bound $k_1(\alpha)$ and $C(\alpha)$. □

Let us now come back to our initial problem, that is comparing the performance of $\tilde{s}$ with that of any other estimator of s. For $\alpha > 0$, $p > 0$ and $R \geq 0$, the minimax risk over $\mathscr{B}\mathscr{P}(\alpha,p,R)$ is given by

$$\mathcal{R}_{\mathscr{B}}(\alpha,p,R) = \inf_{\hat{s}} \sup_{s \in \mathscr{B}\mathscr{P}(\alpha,p,R)} \mathbb{E}_s\big[\|s - \hat{s}\|^2\big]$$

where the infimum is taken over all the estimators $\hat{s}$ of s. We denote by $\mathcal{R}_{\mathscr{V}}(\alpha,R)$ the minimax risk over $\mathscr{V}\mathscr{P}(\alpha,R)$. Thanks to the above approximation results, we obtain, as stated below, that, for a whole range of values of (α,p,R) or (α,R), the estimator $\tilde{s}$ reaches the minimax risk over $\mathscr{B}\mathscr{P}(\alpha,p,R)$ and $\mathscr{V}\mathscr{P}(\alpha,R)$ within a multiplicative constant. Therefore, $\tilde{s}$ is adaptive in the minimax sense not only over the same range of Besov bodies as the *neH*-estimator (*cf.* [15], Cor. 4) but also on a wide range of sets in the scale $\{\mathscr{V}\mathscr{P}(\alpha,R)\}_{\alpha>0,R\geq 0}$.

Theorem 3.7. *For all $p \in (0,2]$ and $\alpha > 1/p - 1/2$, if $n^{-1/2} \leq R < n^\alpha$, then*

$$\sup_{s \in \mathscr{B}\mathscr{P}(\alpha,p,R)} \mathbb{E}_s\big[\|s - \tilde{s}\|^2\big] \leq C(c_0,\alpha,p) \mathcal{R}_{\mathscr{B}}(\alpha,p,R).$$

For all $\alpha > 0$, there exists a real $k_2(\alpha) \in (0,1)$ such that, if $R \geq n^{-1/2}$ and $R \leq k_2(\alpha) n^\alpha$, then

$$\sup_{s \in \mathscr{V}\mathscr{P}(\alpha,R)} \mathbb{E}_s\big[\|s - \tilde{s}\|^2\big] \leq C(c_0,\alpha) \mathcal{R}_{\mathscr{V}}(\alpha,R).$$

Proof. Let us fix $p \in (0,2], \alpha > 1/p - 1/2$ and $n^{-1/2} \leq R < n^\alpha$. Combining inequality (3.5) and Theorem 3.5 leads to

$$\sup_{s \in \mathscr{B}\mathscr{P}(\alpha,p,R)} \mathbb{E}_s\big[\|s - \tilde{s}\|^2\big] \leq C(c_0,\alpha,p) \inf_{1 \leq D \leq n} \big\{ n R^2 D^{-2\alpha} + D \big\}. \tag{3.6}$$

In order to realize approximately the best trade-off between the terms $nR^2D^{-2\alpha}$ and D, which vary in opposite ways when D increases, we choose D as large as possible under the constraint $D \leq nR^2D^{-2\alpha}$. Let us denote by $D^\star$ the largest integer D such that $D \leq (nR^2)^{1/(1+2\alpha)}$. One can easily check that, given the hypotheses linking n and R, $D^\star$ does belong to $\{1,\dots,n\}$. Since $2D^\star > (nR^2)^{1/(1+2\alpha)}$, we deduce from inequality (3.6) the upper-bound

$$\sup_{s\in\mathscr{B}\mathscr{P}(\alpha,p,R)} \mathbb{E}_s\left[\|s-\tilde{s}\|^2\right] \leq C(c_0,\alpha,p)(nR^2)^{1/(2\alpha+1)}.$$

The matching lower-bound for the minimax risk over $\mathscr{B}\mathscr{P}(\alpha,p,R)$ is proved in [15] (Thm. 3).

In the same way, for $\alpha > 0$ and $n^{-1/2} \leq R \leq n^\alpha$, we get

$$\sup_{s\in\mathscr{V}\mathscr{P}(\alpha,R)} \mathbb{E}_s\left[\|s-\tilde{s}\|^2\right] \leq C(c_0,\alpha)(nR^2)^{1/(2\alpha+1)}.$$

The definition of $k_2(\alpha)$ and the matching lower-bound for the minimax risk over $\mathscr{V}\mathscr{P}(\alpha,R)$, for $n^{-1/2} \leq R \leq k_2(\alpha)n^\alpha$, are given in Proposition 7.1. □

3.3. **Computing the d-estimator**

Since the penalty only depends on the dimension of the models, we denote by $\mathrm{pen}(D)$ the penalty assigned to all models in $\mathcal{M}_D$, for $1 \leq D \leq n$. A way to compute $\tilde{s}$ could rely on the equality

$$\min_{m\in\mathcal{M}}\left\{\|X-\hat{s}_m\|^2+\mathrm{pen}(m)\right\} = \min_{1\leq D\leq n}\left\{\min_{m\in\mathcal{M}_D}\|X-\hat{s}_m\|^2+\mathrm{pen}(D)\right\}.$$

We should thus compute the best estimator for each dimension $D \in \{1,\dots,n\}$, and choose one among them by taking into account the penalty term, as in [19] (Chap. 3) or [9]. But, even with Bellman's algorithm, that requires polynomial time. Here, we shall see that we can avoid such a computationaly intensive way by taking advantage of the form of the penalty.

Let us express more explicitly the criterion to minimize. The dyadic intervals of a given partition $m \in \mathcal{M}$ are denoted by $\{i_k,\dots,i_{k+1}-1\}$, $k=1,\dots,D_m$, with $1=i_1<i_2\dots<i_{D_m+1}=n+1$. For all $1\leq k\leq D_m$, any column of $\hat{s}_m$ whose index belongs to $\{i_k,\dots,i_{k+1}-1\}$ is equal to the mean $\bar{X}(i_k:i_{k+1})$ of the columns of X whose indices belong to the interval $\{i_k,\dots,i_{k+1}-1\}$. Owing to the form of the penalty, and to the additivity of the least-squares criterion, the whole criterion to minimize breaks down into a sum:

$$\|X-\hat{s}_m\|^2+\mathrm{pen}(m) = \sum_{k=1}^{D_m}\mathcal{L}(i_k,i_{k+1}), \tag{3.7}$$

where, for all $1\leq k\leq D_m$,

$$\mathcal{L}(i_k,i_{k+1}) = c_0 + \sum_{i=i_k}^{i_{k+1}-1}\|X_i-\bar{X}(i_k:i_{k+1})\|_r^2.$$

By comparison with the method suggested in the previous paragraph, we are left with only one minimization problem, with no dimension constraint, instead of n. We now turn to graph theory where our minimization problem finds a natural interpretation. We consider the weighted directed graph G having $\{1,\dots,n+1\}$ as vertex set and whose edges are the pairs (i,j) such that $\{i,\dots,j-1\}$ is a dyadic interval of $\{1,\dots,n\}$ assigned with the weight $\mathcal{L}(i,j)$. We say that a vertex j is a successor to a vertex i if (i,j) is an edge of the graph G and we associate to each vertex i its successor list Γ_i. For all $1\leq D\leq n$, a $D+1$-uple $(i_1,i_2,\dots,i_{D+1})$ of vertices of G such that $i_1=1$, $i_{D+1}=n+1$ and each vertex is a successor to the previous one, will be called a path leading from 1 to $n+1$ in D steps. The length of such a path is defined as $\sum_{k=1}^{D}\mathcal{L}(i_k,i_{k+1})$. Determining $\hat{m}$ thus amounts to finding a shortest path leading from 1 to $n+1$ in the graph G. That problem can be solved by using a simple shortest-path algorithm dedicated to acyclic directed graphs, presented for instance in [10]

TABLE 1. Algorithm for computing $\tilde{s}$.

***Step 1**: Initialization*
Set $d(1) = 0$ and $p(1) = +\infty$.
For $i = 2, \ldots, n+1$,
  set $d(i) = +\infty$ and $p(i) = +\infty$.

***Step 2**: Determining the lengths of the shortest paths with origin 1*
For $i = 1, \ldots, n$,
  for $j \in \Gamma_i$,
    if $d(j) > d(i) + \mathcal{L}(i,j)$,
      then do $d(j) \leftarrow d(i) + \mathcal{L}(i,j)$ and $p(j) \leftarrow i$.

***Step 3**: Determining a shortest path P from 1 to $n+1$*
Set $pred = p(n+1)$ and $P = (n+1)$.
While $pred \neq +\infty$,
  replace P with the concatenation of $pred$ followed by P,
  do $pred \leftarrow p(pred)$.

***Step 4**: Computing the d-estimator*
Set $\tilde{D} = length(P) - 1$.
For $k = 1, \ldots, \tilde{D}$,
  for $i = P(k), \ldots, P(k+1) - 1$,
    set $\tilde{s}_i = \bar{X}(P(k) : P(k+1))$.

(Sect. 24.2). For the sake of completeness, we also describe it in Table 1. We have to underline that there are only $2n-1$ dyadic intervals of $\{1, \ldots, n\}$. Therefore, the graph G, with $n+1$ vertices and $2n-1$ edges, can be represented by only $\mathcal{O}(n)$ data: the weights $\mathcal{L}(i,j)$, for $1 \leq i \leq n$ and $j \in \Gamma_i$, and the successor lists Γ_i, for $1 \leq i \leq n$. In the key step of the algorithm, *i.e.* step 2, each edge is considered only once. When the time comes to consider the edges with origin i, the variables $d(i)$ and $p(i)$ respectively contain the length of a shortest path from 1 to i and a predecessor of i in such a path. Just before the edge (i,j), where $j \in \Gamma_i$, be processed, the variables $d(j)$ and $p(j)$ contain respectively the length of a shortest path leading from 1 to j and a predecessor of j in such a path, based solely on the edges that have already been encountered. Then dealing with the edge (i,j) consists in testing whether the length of the path leading from 1 to j can be shortened by going *via* i and updating, if necessary, $d(j)$ and $p(j)$. What clearly appears from the above description of the algorithm is that its complexity is only *linear* in the length n of the sequence.

4. HYBRID PROCEDURE

We shall now apply the previous procedure to the detection of multiple change-points in the distribution s. Let us give a first glimpse of what can be expected from the d-estimator for that problem. In Figure 1, we plot the first coordinate function of a distribution $s_a \in \mathscr{M}(2, 1024)$ that is piecewise constant over a partition with only 3 segments together with the first coordinate of a realization of $\widetilde{s_a}$. The value of c_0 has been chosen so as to minimize the distance between s_a and its estimator. If both change-points in s_a are indeed detected, the selected partition, due to its special nature, also points at irrelevant ones. In order to get rid of them, we propose a two-stage procedure, that we name hybrid procedure. After describing it, we provide an adaptivity result for that procedure and end this section with computational issues.

In the sequel, we suppose that $n \geq 4$. We shall work with the set $\mathscr{M}(r, n/2)$ of $r \times (n/2)$ real matrices and introduce other notation. For all $t \in \mathscr{M}(r,n)$, we denote by $t^\bullet$ (resp. t°) the element of $\mathscr{M}(r, n/2)$ composed of the columns of t whose indices are even (resp. odd). We equip $\mathscr{M}(r, n/2)$ with the norm analogous to the

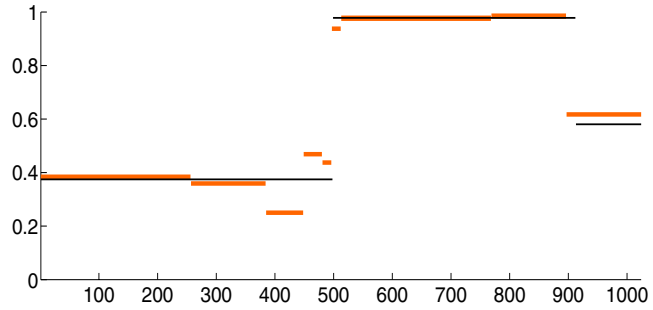

FIGURE 1. (Color online) First coordinate functions of the distribution s_a (thin black line) and of its d-estimator $\widetilde{s_a}$ (thick yellow line).

norm $\|.\|$ on $\mathscr{M}(r,n)$. For the sake of simplicity, we also denote by $\|.\|$ that norm on $\mathscr{M}(r,n/2)$. For a partition m of $\{1,\dots,n/2\}$, we denote by S'_m the linear subspace of $\mathbb{R}^{n/2}$ generated by the indicator functions of the intervals $I \in m$ and by D'_m its dimension. We are now able to describe the hybrid procedure. First, the previous procedure based on $X^\bullet$ provide us with a random partition of $\{1,\dots,n/2\}$ into dyadic intervals denoted by $\hat{m}^\bullet$. Then, we consider the random collection $\widehat{\mathcal{M}}^\bullet$ of all the partitions of $\{1,\dots,n/2\}$ that are built on $\hat{m}^\bullet$. For each partition m of $\{1,\dots,n/2\}$, we define the least-squares estimator of s° in $\mathbb{R}^r \otimes S'_m$ by

$$\widehat{s^\circ}_m = \operatorname*{argmin}_{t \in \mathbb{R}^r \otimes S'_m} \|X^\circ - t\|^2.$$

Then we select

$$\hat{m}^\circ = \operatorname*{argmin}_{m \in \widehat{\mathcal{M}}^\bullet} \left\{ \|X^\circ - \widehat{s^\circ}_m\|^2 + \widehat{\mathrm{pen}}^\circ(m) \right\},$$

where the penalty $\widehat{\mathrm{pen}}^\circ$ will be chosen in the next paragraph. That partition provide us with the estimated change-points in the distribution s. As a matter of fact, we define the hybrid estimator $\tilde{s}_{hyb}$ of s as the random $r \times n$ matrix whose submatrices composed respectively of columns with even indices and of columns with odd indices are both equal to $\widehat{s^\circ}_{\hat{m}^\circ}$. The application of this procedure to s_a is illustrated by Figure 4. Notice that other ways of splitting the sample could be considered. This one has been chosen for ease of notation.

We obtain the following upper-bound for the risk of $\tilde{s}_{hyb}$.

Theorem 4.1. *Let $\widehat{D}$ be the cardinal of $\hat{m}^\bullet$ and $\widehat{\mathrm{pen}}^\circ : \widehat{\mathcal{M}}^\bullet \to \mathbb{R}^+$ be a penalty of the form*

$$\widehat{\mathrm{pen}}^\circ(m) = \left(c_1 + c_2 \ln\left(\widehat{D}/D'_m\right) \right) D'_m, \tag{4.8}$$

where c_1 and c_2 are positive. If c_0, c_1 and c_2 are large enough, then

$$\mathbb{E}_s\left[\|s - \tilde{s}_{hyb}\|^2\right] \le C(c_0,c_1,c_2)\left[\inf_{m \in \mathcal{M}} \left\{ \|s - s_m\|^2 + D_m \right\} + \|s^\circ - s^\bullet\|^2 \right].$$

Thus, if s also satisfies $\|s^\circ - s^\bullet\|^2 \le \lambda \inf_{m\in\mathcal{M}} \left\{ \|s - s_m\|^2 + D_m \right\}$, then

$$\mathbb{E}_s\left[\|s - \tilde{s}_{hyb}\|^2\right] \le C(c_0,c_1,c_2,\lambda) \inf_{m \in \mathcal{M}} \left\{ \|s - s_m\|^2 + D_m \right\}. \tag{4.9}$$

Inequality (4.9) must be compared with inequality (3.3). In particular, provided s° and $s^\bullet$ are close enough, the adaptivity properties of the hybrid estimator are similar to those of the d-estimator. The constant $C(c_0,c_1,c_2,\lambda)$ in (4.9) is expected to be larger than the constant $C(c_0)$ in (3.3), but we will see in Section 5.3 that, in practice, provided the penalty constants are well chosen, the risk of $\tilde{s}_{hyb}$ is not so far from that of $\tilde{s}$.

Proof. For all $1 \leq D \leq \widehat{D}$, the number $\widehat{N}_D$ of partitions in $\widehat{\mathcal{M}}^\bullet$ with D pieces satisfies

$$\widehat{N}_D = \binom{\widehat{D}-1}{D-1} \leq \left(\frac{e\widehat{D}}{D}\right)^D.$$

The above inequality results from a property of binomial coefficients that may be found in [21] (Prop. 2.5) for instance. So the weights defined by

$$\widehat{L}(D) = \ln(2e) + \ln(\widehat{D}/D), \text{ for } 1 \leq D \leq \widehat{D},$$

are such that

$$\sum_{D=1}^{\widehat{D}} \widehat{N}_D \exp(-D\widehat{L}(D)) \leq 1.$$

Moreover, given $X^\bullet$, the penalty $\widehat{\text{pen}}^\circ$ given by (4.8) fulfills the hypotheses of Theorem 1 in [15] provided c_1 and c_2 are large enough. With a slight abuse of notation, for any partition m of $\{1,\dots,n/2\}$, we still denote by t_m the orthogonal projection of an element $t \in \mathscr{M}(r, n/2)$ on $\mathbb{R}^r \otimes S'_m$. Working conditionally to $X^\bullet$, the collection $\widehat{\mathcal{M}}^\bullet$ is deterministic, so we deduce from Theorem 1 of [15] applied to the estimator $\widehat{s}^\circ_{\hat{m}^\circ}$ of s° that

$$\mathbb{E}_{s^\circ}\left[\|s^\circ - \widehat{s}^\circ_{\hat{m}^\circ}\|^2 | X^\bullet\right] \leq C(c_1,c_2)\left[\|s^\circ - s^\circ_{\hat{m}^\bullet}\|^2 + \widehat{\text{pen}}^\circ(\hat{m}^\bullet)\right]. \tag{4.10}$$

We recall that the *d*-estimator of $s^\bullet$ is $\widetilde{s}^\bullet = \widehat{s}^\bullet_{\hat{m}^\bullet}$. So, thanks to the triangle inequality, and since an orthogonal projection is a shrinking map, we get

$$\|s^\circ - s^\circ_{\hat{m}^\bullet}\|^2 \leq C\big(\|s^\circ - s^\bullet\|^2 + \|s^\bullet - \widetilde{s}^\bullet\|^2\big).$$

By definition, $\widehat{D} = D'_{\hat{m}^\bullet}$, so

$$\widehat{\text{pen}}^\circ(\hat{m}^\bullet) = c_1\widehat{D}.$$

Taking into account the last two inequalities and integrating with respect to $X^\bullet$ leads from (4.10) to

$$\mathbb{E}_s\left[\|s^\circ - \widehat{s}^\circ_{\hat{m}^\circ}\|^2\right] \leq C(c_1,c_2)\Big[\|s^\circ - s^\bullet\|^2 + \mathbb{E}_{s^\bullet}\left[\|s^\bullet - \widetilde{s}^\bullet\|^2\right] + \mathbb{E}_{s^\bullet}(\widehat{D})\Big].$$

Besides, it follows from the definition of $\tilde{s}_{hyb}$ that

$$\|s - \tilde{s}_{hyb}\|^2 = \|s^\bullet - \widehat{s}^\circ_{\hat{m}^\circ}\|^2 + \|s^\circ - \widehat{s}^\circ_{\hat{m}^\circ}\|^2.$$

Applying the triangle inequality, we then get

$$\|s - \tilde{s}_{hyb}\|^2 \leq C\big(\|s^\bullet - s^\circ\|^2 + \|s^\circ - \widehat{s}^\circ_{\hat{m}^\circ}\|^2\big).$$

Consequently,

$$\mathbb{E}_s\left[\|s - \tilde{s}_{hyb}\|^2\right] \leq C(c_1,c_2)\Big[\|s^\circ - s^\bullet\|^2 + \mathbb{E}_{s^\bullet}\left[\|s^\bullet - \widetilde{s}^\bullet\|^2\right] + \mathbb{E}_{s^\bullet}(\widehat{D})\Big]. \tag{4.11}$$

Let us denote by $\mathcal{M}'$ the set of all partitions of $\{1,\dots,n/2\}$ into dyadic intervals. For the risk of $\widetilde{s}^\bullet$, inequality (3.3) provides

$$\mathbb{E}_{s^\bullet}\left[\|s^\bullet - \widetilde{s}^\bullet\|^2\right] \leq C(c_0) \inf_{m\in\mathcal{M}'}\left\{\|s^\bullet - s^\bullet_m\|^2 + D'_m\right\}. \tag{4.12}$$

In order to bound the term $\mathbb{E}_{s^\bullet}(\widehat{D})$, we need to go back to the proof of Theorem 1 in [15] (Sect. 8.1). As already seen during the proof of Proposition 3.1, we can choose a positive constant L such that $\sum_{m\in\mathcal{M}'}\exp(-LD'_m) \leq 1$.

Let us fix a partition $m \in \mathcal{M}'$ and $\xi > 0$. Using the same notation as in [15], we deduce from the proof of Theorem 1 in [15] that there exists an event $\Omega_\xi(m)$ such that $\mathbb{P}_{s^\bullet}(\Omega_\xi(m)) \geq 1 - \exp(-\xi)$ and on which

$$c_0 \widehat{D} \leq C_1 \|s^\bullet - s_m^\bullet\|^2 + C_2(c_0) D'_m + C_3 \widehat{D} + C_4 \xi.$$

Therefore, if $c_0 > C_3$, then

$$\widehat{D} \leq C(c_0)\big(\|s^\bullet - s_m^\bullet\|^2 + D'_m + \xi\big).$$

Integrating this inequality and taking the infimum over $m \in \mathcal{M}'$ then yields

$$\mathbb{E}_{s^\bullet}(\widehat{D}) \leq C(c_0) \inf_{m \in \mathcal{M}'} \big\{\|s^\bullet - s_m^\bullet\|^2 + D'_m\big\}. \tag{4.13}$$

Moreover, one can check that

$$\inf_{m \in \mathcal{M}'} \big\{\|s^\bullet - s_m^\bullet\|^2 + D'_m\big\} \leq \inf_{m \in \mathcal{M}} \big\{\|s - s_m\|^2 + D_m\big\}. \tag{4.14}$$

Combining Inequalities (4.11) to (4.14), we finally get

$$\mathbb{E}_s\big[\|s - \tilde{s}_{hyb}\|^2\big] \leq C(c_0, c_1, c_2)\Big[\|s^\circ - s^\bullet\|^2 + \inf_{m \in \mathcal{M}} \big\{\|s - s_m\|^2 + D_m\big\}\Big]. \qquad \square$$

Regarding the computation of $\tilde{s}_{hyb}$, we know from Section 3.3 that determining $\widetilde{s^\bullet}$ only requires $\mathcal{O}(n)$ computations. On the other hand, since $\widehat{\text{pen}}^\circ$ is not linear in the dimension of the models, $\hat{m}^\circ$ has to be determined following the method suggested at the beginning of Section 3.3 and using Bellman's algorithm. Thus, the second stage requires $\mathcal{O}(\hat{D}^3)$ computations. However, if s belongs to $\mathscr{BP}(\alpha, p, R)$ or $\mathscr{VP}(\alpha, p, R)$, it follows from Inequalities (4.13) and (4.14) and the proof of Theorem 3.7 that the expectation of $\hat{D}$ is of order $n^{1/(1+2\alpha)}$. In such a case, the second stage of the hybrid procedure is thus expected to require much less than $\mathcal{O}(n^3)$ computations.

5. Simulation study

In the previous sections, our main concern has been to propose a form of penalty yielding, in theory, a performant estimator. In this section, we study some practical choice of the penalty for each procedure. Besides, we compare the d-estimator with the neH-estimator proposed in [15] on several simulated examples. We also compare on a DNA sequence our hybrid procedure with that based on CART (*cf.* [19]) and that based on the neH-procedure (*cf.* [15], Sect. 7).

5.1. Choosing the penalty constant for the d-estimator

We have examined some examples for $r = 2$ and $r = 4$, with different values of $n = 2^N$. For $r = 2$, the distribution s is entirely determined by its first coordinate function, that is the only one to be plotted (*cf.* Fig. 2). For $r = 4$, examples s_d to s_f are plotted in Figure 3 (left column).

As already said in Section 3.1, the d-estimator has been designed for satisfying an oracle inequality, what it almost does according to Proposition 3.1. Therefore, the risk of the oracle, *i.e.* $\inf_{m \in \mathcal{M}} \mathbb{E}_s\big[\|s - \hat{s}_m\|^2\big]$, serves as a benchmark in order to judge of the quality of $\tilde{s}$, and also of the quality of a method for choosing a penalty constant. The different quantities introduced in the sequel have been estimated over 500 simulations. Denoting by $\tilde{s}(c)$ the d-estimator when c_0 takes the value c, we have first estimated

$$c^\star(s) := \operatorname*{argmin}_c \ \mathbb{E}_s\big[\|s - \tilde{s}(c)\|^2\big],$$

where, in practice, we have varied c from 0 to 4, by step 0.1, and from 4 to 6 by step 0.5. We plot in Table 2 an estimation of $c^\star$ and the ratio $Q^\star$ between an estimation of $\mathbb{E}_s\big[\|s - \tilde{s}(c^\star)\|^2\big]$ and the estimated risk of the

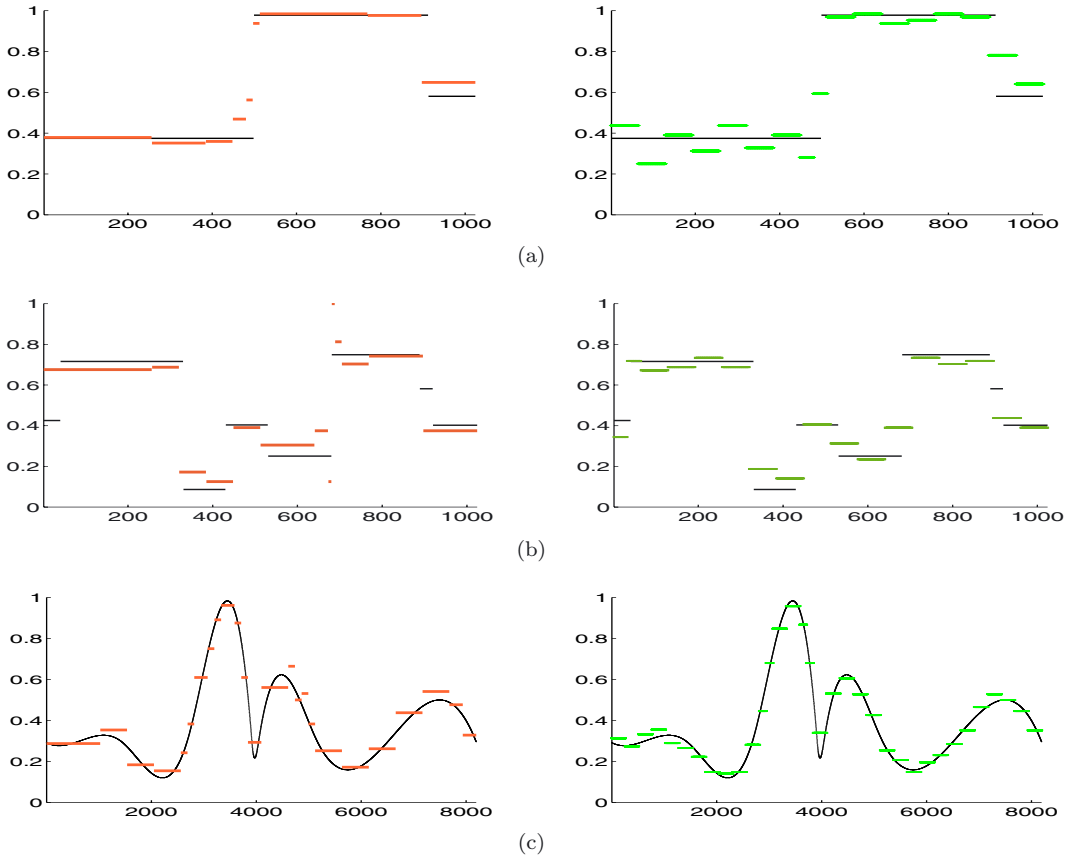

FIGURE 2. (Color online) Left column: first coordinate functions of s (thin black line) and $\tilde{s}$ (thick orange line) for $s \in \{s_a, s_b, s_c\}$; Right column: first coordinate functions of s (thin black line) and $\tilde{s}_{neH}$ (thick green line) for $s \in \{s_a, s_b, s_c\}$.

TABLE 2. Performance of the d-estimator for different choices of the penalty constant.

s	r	N	$c^\star$	$Q^\star$	$\bar{c}_j$	σ_j	Q_j
s_a	2	10	1.7	2.4	1.9	0.2	2.7
s_b	2	10	1.7	1.9	2.0	0.2	2.1
s_c	2	13	2.2	1.7	2.0	0.1	1.8
s_d	4	10	2.1	1.4	2.4	0.2	1.4
s_e	4	12	2.5	1.3	2.3	0.1	1.3
s_f	4	13	2.7	1.3	2.5	0.1	1.3

oracle. In view of the results obtained here, it seems difficult to propose a value of c_0 that would be convenient for any s. Therefore, as in [15], Section 8, we have tried a data-driven method, inspired from results proved by Birgé and Massart in a Gaussian framework (*cf.* [7]). Given a simulation of $(Y_1, \ldots, Y_n)$, the procedure we have followed can be decomposed into three steps:

- determine the dimension $\hat{D}(c)$ of the selected partition for each value c of the penalty constant c_0, where c increases from 0, by step 0.1, until $\hat{D}(c) = 1$;

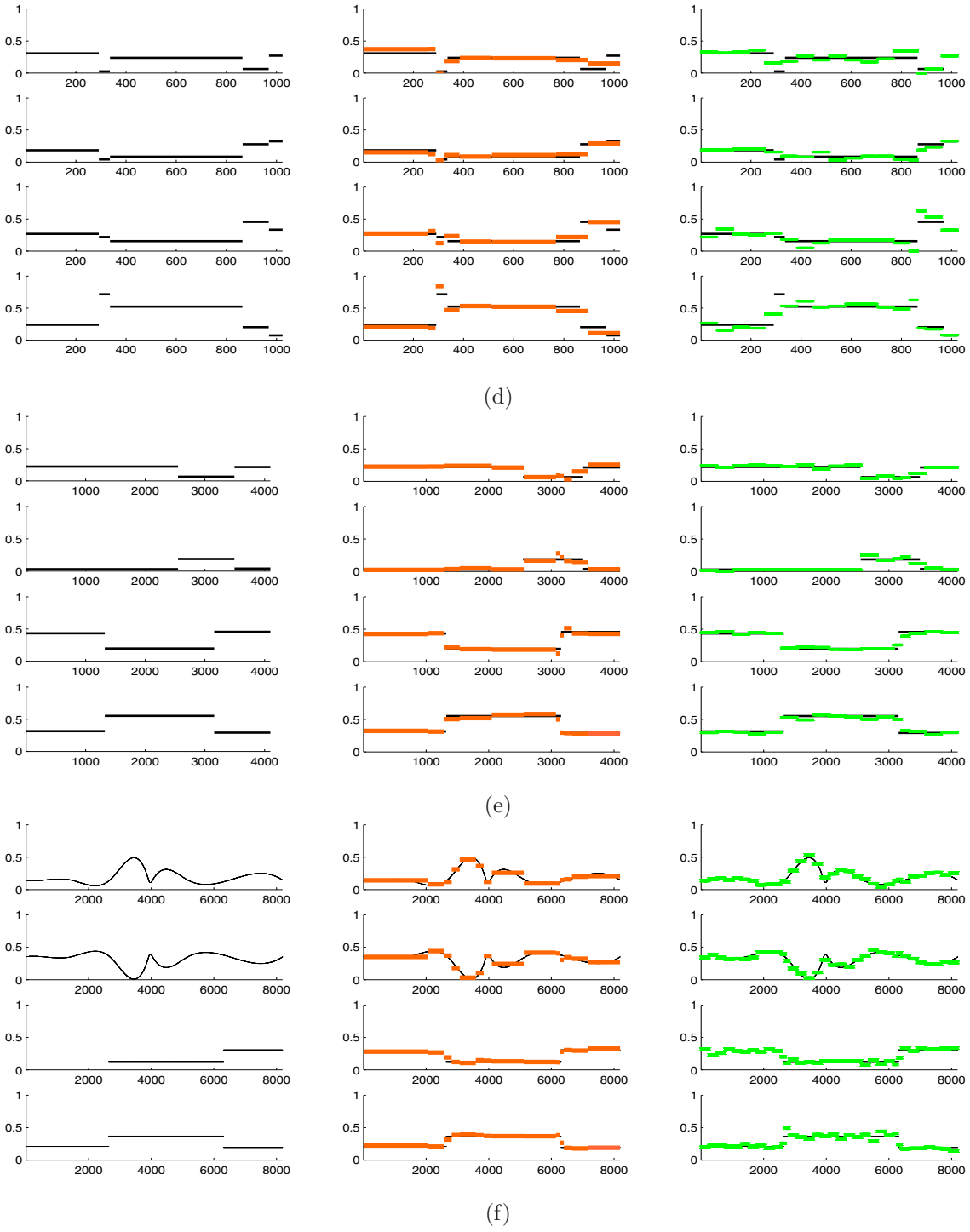

FIGURE 3. (Color online) For $s \in \{s_d, s_e, s_f\}$. Left column: four coordinate functions of s; center column: four coordinate functions of s (thin black line) and $\tilde{s}$ (thick orange line); right column: four coordinate functions of s (thin black line) and $\tilde{s}_{neH}$ (thick green line).

TABLE 3. Comparison between the d-estimator and the neH-estimator.

s	r	N	risk_d	risk_{neH}	$\text{risk}_d/\text{risk}_{neH}$	$\text{time}_d/\text{time}_{neH}$
s_a	2	10	11.5	16.4	0.7	8.6
s_b	2	10	22.1	26.5	0.8	1.1
s_c	2	13	40.2	36.2	1.1	0.5
s_d	4	10	14.0	19.4	0.7	1.2
s_e	4	12	18.3	23.8	0.8	0.6
s_f	4	13	33.0	37.7	0.9	0.3

- compute the difference between the dimensions of the selected partitions for two consecutive values of c_0 and retain the value $\hat{c}$ corresponding to the biggest jump in dimension under the constraint $\hat{D}(\hat{c}) \leq D_{\max}$, where $D_{\max}$ is a prescribed maximal dimension;
- set $\hat{c}_j = 2\hat{c}$ and compute the d-estimator with $\text{pen}(D) = \hat{c}_j D$.

Here we have taken $D_{\max} = 60$ when $N = 10$, $D_{\max} = 200$ when $N = 12$ and $D_{\max} = 300$ when $N = 13$. We give in Table 2 the ratio Q_j between the estimated risk of $\tilde{s}$ for that procedure and the estimated risk of the oracle. We also give estimations of the mean value and standard-error of $\hat{c}_j$, denoted respectively by $\bar{c}_j$ and σ_j. One realization of each d-estimator computed with that method is plotted in Figures 2 (left column) and 3 (center column).

Let us analyze the results of the simulations. The data-driven method really seems to adapt to the unknown distribution s: in terms of risk, it is almost as good as if we knew the constant that minimizes the risk of $\tilde{s}$. Let us now compare the different values of $Q^\star$ (or Q_j). As foreseen by the oracle-type inequality (3.4), the ratio between the risk of the d-estimator and that of the oracle depends on s. In particular, the ratios $Q^\star$ or Q_j reach their highest value for s_a. It should be noted that the first coordinate function of this example takes values very close to 1 on a large segment (*cf.* Fig. 2), a critical case according to the oracle-type inequality. However, for all examples studied here, the values of those ratios remain quite low, inferior or close to 2, except for s_a.

5.2. Comparing the d-estimator with the neH-estimator

For examples s_a to s_f, we have realized 500 simulations of the d-estimator and the neH-estimator, using a data-driven penalty (*cf.* [15], Sect. 8, and the previous paragraph). We provide in Table 3 the estimated risks of each procedure, denoted by risk_d and risk_{neH}. Thanks to MATLAB "tic" and "toc" functions, we have measured the computational time of those 500 simulations for each estimator, denoted by time_d and time_{neH}. The ratio of those computational times is given in Table 3. The neH-estimators of examples s_a to s_f are plotted in Figures 2 and 3 (right columns).

Those results confirm that both procedures have about the same quality of estimation, with a slight advantage though for the d-procedure for almost all the examples. As to their computational time, let us recall that the neH-procedure requires $\mathcal{O}(n \ln(n))$ computations, against only $\mathcal{O}(n)$ for the d-procedure. That difference clearly appears through our simulations. The d-estimator seems faster to compute if n is large enough, and else requires roughly the same computational time as the neH-estimator. The only exception here occurs with s_a, but 500 simulations of $\tilde{s}_a$ can be computed within a few minutes only.

5.3. Choosing the penalty for the hybrid procedure

For the first stage of the hybrid procedure, the d-estimator has been computed using the data-driven penalty. For the second stage, the practical choice of an adequate penalty is more delicate, since the theoretical penalty depends in this case on two constants and on the dimension $\widehat{D}$ of the partition selected during the first stage.

TABLE 4. Comparison between the hybrid estimator and the d-estimator.

s	D	$\hat{D}_d$	$\hat{D}_{hyb}$	$Q_{hyb:d}$
s_a	3	9.5 (3.3)	2.9 (0.4)	1.6
s_b	8	16.1 (3.1)	4.5 (1.0)	1.6
s_d	5	8.9 (2.0)	3.0 (0.8)	1.7
s_e	5	12.3 (2.1)	4.9 (1.1)	1.7

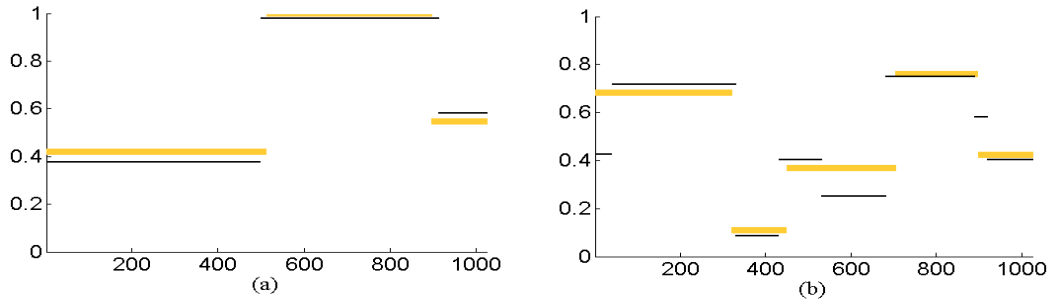

FIGURE 4. (Color online) First coordinate functions of s (thin black line) and $\tilde{s}_{hyb}$ (thick yellow line) for (a) $s = s_a$ and (b) $s = s_b$.

Since those two constants seem difficult to determine, we have assigned to all partitions of $\{1, \dots, n/2\}$ into D intervals the penalty

$$\widehat{\text{pen}}^{\circ}(D) = \hat{\beta} D.$$

The value of $\hat{\beta}$ is determined once again according to the same process as $\hat{c}_j$ (*cf.* Sect. 5.1), varying the value of the constant by step 1, and taking $D_{\max} = \hat{D}$. Since that penalty is a linear funtion of D, the second stage of the hybrid procedure can be implemented in that case with the same algorithm as the d-procedure (*cf.* Sect. 3.3). As the graph associated with all the partitions built on a partition with $\hat{D}$ intervals has $\mathcal{O}(\hat{D}^2)$ vertices, the second stage thus requires $\mathcal{O}(\hat{D}^2)$ computations, instead of $\mathcal{O}(\hat{D}^3)$ if we had used a penalty with two constants.

We have tested the hybrid procedure on examples s_a, s_b, s_d and s_e. The hybrid estimators of these examples are plotted in Figures 4 and 5. In order to draw a comparison between the hybrid procedure and the d-procedure, we give in Table 4 the following information for distributions s_a, s_b, s_d and s_e, still computed over 500 simulations. We first recall the dimension D of the partition on which s is built. Then we indicate the estimated mean of the dimensions $\hat{D}_d$ and $\hat{D}_{hyb}$ of the partitions selected respectively by the d-procedure and the hybrid procedure with data-driven penalties, and give between parentheses their estimated standard errors. We also give the ratio $Q_{hyb:d}$ between the estimated risk of the hybrid estimator and the estimated risk of the d-estimator. The estimated means of $\hat{D}_{hyb}$ and $\hat{D}_d$ indicate that the dimension of the partition selected by the hybrid procedure is much closer to the true one. Moreover, Figures 4 and 5 show that the most significant change-points are still detected and quite close to the true ones, and that irrelevant change-points are much fewer with the hybrid procedure. The only price to pay is an increase in risk, but only by a factor of the order of 2.

5.4. **Application to the segmentation of a DNA sequence**

A DNA sequence of length n can be considered as a realization of a n-uple $(Y_1, \dots, Y_n)$ of independent categorical variables with values in $\{1, \dots, 4\}$, when coding the set of bases {A,C,G,T} by $\{1, \dots, 4\}$ for instance. We have tested our hybrid procedure on a DNA sequence taken from the *Bacillus subtilis* genome. The whole genome of that bacterium (available on the NCBI website, under accession number NC_000964 in the Genome database) is composed of two complementary strands, each counting approximately 4 millions of bases. We have applied our procedure on the DNA sequence composed of the first 2^{21} =2 097 152 bases of the strand

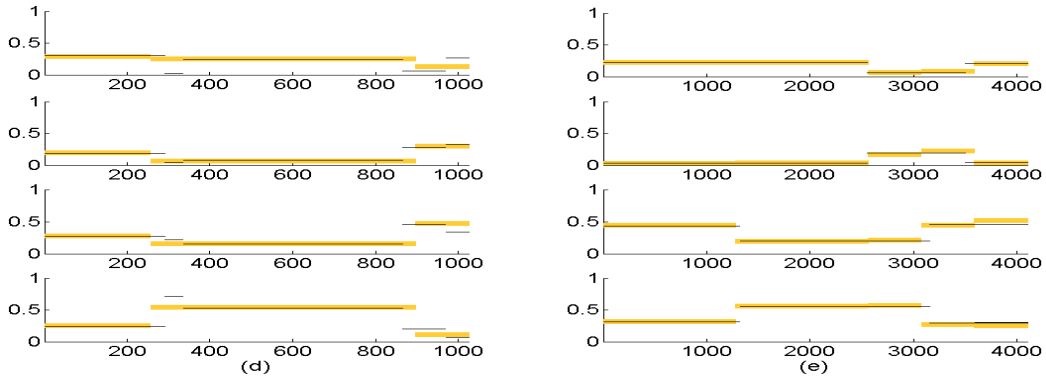

FIGURE 5. (Color online) Coordinate functions of s (thin black line) and $\tilde{s}_{hyb}$ (thick yellow line) for (d) $s = s_d$ and (e) $s = s_e$.

usually referred to as the (+) strand. For the sake of readability, we only represent in Figure 6, realized with MuGeN software [18], the genes corresponding to the first 178 000 base pairs (bp) of *B. subtilis* genome. We shall mainly distinguish between two kinds of genes: those coding for proteins and those coding for structural RNA. The first ones are represented by cyan or magenta arrows, depending on their orientation, an unfilled arrow indicating that the protein function is still unknown. The other ones are represented by red arrows if they code for ribosomal RNA (rRNA), dark blue arrows if they code for transfer RNA (tRNA), and by an empty box if they code for a small cytoplasmic RNA (scRNA). The rest of the sequence corresponds to intergenic regions, that do not contain any gene.

Let us first analyze our results for the subsequence represented in Figure 6. Our hybrid procedure delineates 19 segments: in Figure 6, the 18 corresponding change-points are represented by the highest vertical bars, and numbered from 2 to 19 so that the number i indicates the beginning of the ith segment. The estimated proportions of bases A,C,G,T in each segment are given in Table 5. Segments 2, 8, 12 and 18 clearly correspond with the 4 regions of the sequence composed at the same time of genes coding for rRNA and of genes coding for tRNA. Table 5 shows that these segments have almost the same composition, that differs from the composition of any other segment. Segments 3, 5, 16 and 17 correspond with 4 regions mainly composed of protein coding genes oriented in the negative sense. We detect all such regions except for the smallest one of about 300 bases (near 45 000 bp). All the other segments are mainly composed of protein coding genes oriented in the positive sense. In particular, segment 15 includes all the genes known to code for ribosomal proteins. Let us also underline that segments 9 and 13 have similar compositions and are both situated just after one of the 4 segments coding for rRNA and tRNA. But, as the function of the protein coded by gene csfB in segment 9 is unknown, we do not know whether such similarities are related to a biological feature.

Let us now compare our results to the aforementionned procedures. The hybrid procedures based on CART and on the *neH*-procedures have been tested on the subsequences composed respectively of the first 200 000 bases of *B. subtilis* (+) strand in [19] (Sect. 7.2.3) and of the first 2^{21} bases of that same strand in [15] (Sect. 7.2). In Figure 6, the resulting change-points are represented by the smallest vertical bars, numbered from 2 to 10, for the former procedure, and by the medium height bars, numbered from 2 to 17, for the latter procedure. As [19] and [15], we detect all the regions composed of genes coding for rRNA and tRNA. We recover the same changes of orientation as [15], except for the shortest region, and also detect another change (near 160 000 bp). The 15th segment obtained with our dyadic based hybrid procedure can be compared with the 13th segment obtained by [15], that contains all genes known to code for ribosomal proteins except for the one following gene sigH. Consequently, as [15], we slightly improve on the results obtained by [19]. Besides, unlike [19] or [15], we detect two segments that might be relevant to the biologist. Moreover, our method is expected to be the fastest since its first stage has the lowest computational complexity.

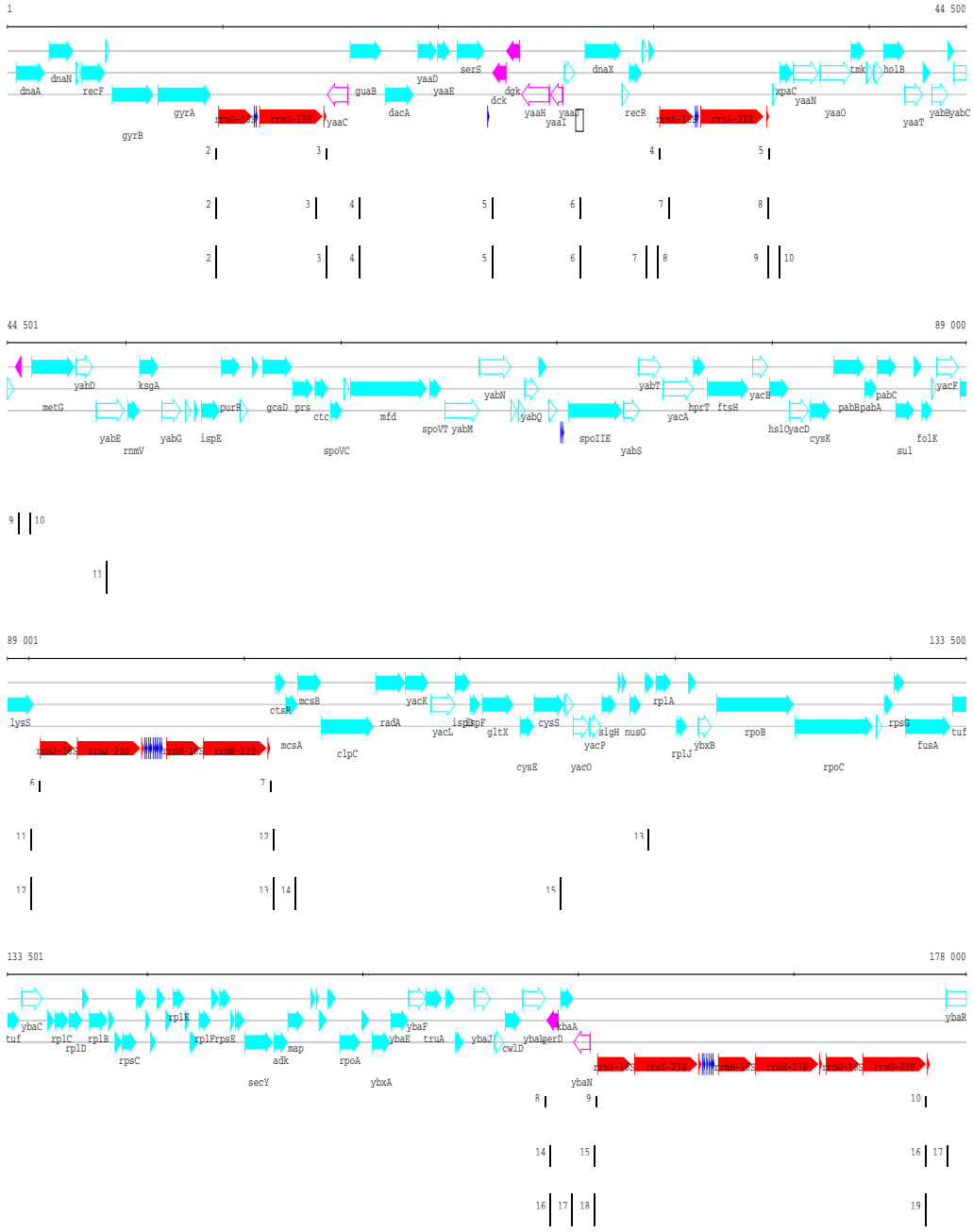

FIGURE 6. (Color online) Annotation of the first 178 000 base pairs (bp) of *B. subtilis* genome, and change-points detected by the hybrid procedures based on CART (small bars), on the *neH*-procedure (medium bars) and on the dyadic procedure (tall bars). Protein coding genes are represented by cyan or magenta arrows, depending on their orientation, unfilled when the protein function is unknown. Red and dark blue arrows represent genes coding respectively for rRNA and tRNA. The empty box stands for a gene coding for scRNA.

TABLE 5. Estimated proportions of bases A, C, G, T (in percentage) in the 18 first segments obtained by applying our hybrid procedure to *B. subtilis* (+) strand.

Segment	1	2	3	4	5	6	7	8	9	10
Begin	1	9730	14850	16386	22530	26626	29698	30210	35330	35842
A	32	26	30	31	27	33	30	26	37	33
C	22	31	18	25	22	24	22	30	19	23
G	26	21	33	26	31	24	33	21	29	26
T	20	23	18	18	20	18	15	23	15	18

Segment	11	12	13	14	15	16	17	18
Begin	49154	90114	101378	102402	114690	158722	159746	160770
A	31	26	37	31	31	24	26	26
C	24	30	20	25	23	20	21	30
G	26	22	27	26	27	40	34	21
T	19	22	16	18	19	17	19	22

Let us end with a comparison of our results with those obtained in [22] by using hidden Markov chain models on the whole (+) strand of *B. subtilis* genome. At the level of gene detection, our procedure, that relies on the assumption that the bases are independent, cannot rival with that used in [22]. But we can compare the biological features of the groups of genes that our procedure highlights with those associated with the hidden states of the most complex model fitted by [22] (see their Fig. 3). Our method does not seem to detect neither intergenic regions and protein coding genes having a similar composition (called atypical genes in [22]), nor genes coding for hydrophobic proteins. But we detect the other four features, since we delineate large groups of genes coding for structural RNA, groups of protein coding genes with negative orientation, groups of protein coding genes with positive orientation, and among them the group of genes coding for ribosomal proteins, described in [22] as the main region composed of highly expressed genes. Notice also that the distinction between those four features is not made by any of the less complex models tested by [22].

6. PROOF OF THE APPROXIMATION RESULT OVER BESOV BODIES

This section is devoted to the proof of Theorem 3.5, that extends the approximation result of DeVore and Yu [14] (Sect. 3) to the approximation of $\mathbb{R}^r$-valued functions defined on $\{1,\dots,n\}$ by piecewise constant functions. For $r=1$, that extension simply results from [14] (Cor. 3.2) and Proposition 6.5 (*cf.* Sect. 6.2), which is not the case anymore for $r \geq 2$. We first describe the approximation algorithm adapted from [14]. Then, we give the main lines of the proof and also demonstrate the key result, which is a direct consequence of the approximation algorithm. The proofs of more technical points are postponed to the next subsections.

6.1. **Approximation algorithm**

Let us fix $p \in (0,2]$, $\alpha > 1/p - 1/2$, $R > 0$ and $D \in \{1,\dots,n\}$. In order to prove Theorem 3.5, we look for an upper bound for

$$\inf_{m \in \mathcal{M}_D} \|t - t_m\|^2$$

uniformly over $t \in \mathscr{B}(\alpha,p,R)$. Let I be a dyadic interval of $\{1,\dots,n\}$. The restriction of the norm $\|.\|$ to I is denoted by $\|.\|_I$. Let U be the linear subspace of $\mathbb{R}^n$ generated by the vector $(1\dots1)$, we denote by $\mathcal{E}_2(t,I)$ the error in approximating t on I by an element of $\mathbb{R}^r \otimes U$, *i.e.*

$$\mathcal{E}_2(t,I) = \inf_{c \in \mathbb{R}^r \otimes U} \|t - c\|_I.$$

Besides, both intervals obtained by dividing I into two intervals of same length are called the children of I. The algorithm proposed by DeVore and Yu [14] (Sect. 2) proceeds as follows. We fix a threshold $\epsilon > 0$. At the beginning, the set $\mathcal{I}^1(t,\epsilon)$ contains $I_{(0,0)} = \{1,\dots,n\}$. If $\mathcal{E}_2\big(t, I_{(0,0)}\big) \leq \epsilon$, then the algorithm stops. Else, $I_{(0,0)}$ is replaced in the partition $\mathcal{I}^1(t,\epsilon)$ with his children, hence a new partition $\mathcal{I}^2(t,\epsilon)$ of $\{1,\dots,n\}$. In the same way, the kth step starts with a partition $\mathcal{I}^k(t,\epsilon)$ of $\{1,\dots,n\}$ into k dyadic intervals. If $\sup_{I\in\mathcal{I}^k(t,\epsilon)} \mathcal{E}_2(t,I) \leq \epsilon$, then the algorithm stops, else an interval I such that $\mathcal{E}_2(t,I) > \epsilon$ is chosen in $\mathcal{I}^k(t,\epsilon)$ and replaced with his children, hence a new partition $\mathcal{I}^{k+1}(t,\epsilon)$ of $\{1,\dots,n\}$ into $k+1$ dyadic intervals. The algorithm finally stops, giving a partition $\mathcal{I}(t,\epsilon)$. Denoting by $S(t,\epsilon)$ the linear space composed of the functions that are piecewise constant on $\mathcal{I}(t,\epsilon)$, the approximation $A(t,\epsilon)$ of t associated with this partition is defined as the orthogonal projection of t on $\mathbb{R}^r \otimes S(t,\epsilon)$. So, the approximation error of t by $A(t,\epsilon)$ satisfies

$$\|t - A(t,\epsilon)\|^2 = \sum_{I\in\mathcal{I}(t,\epsilon)} \big(\mathcal{E}_2(t,I)\big)^2 \leq |\mathcal{I}(t,\epsilon)|\epsilon^2.$$

For any $\epsilon > 0$ such that the algorithm stops at the latest at step D, the approximation of t that we get belongs to the collection $\{\mathbb{R}^r \otimes S_m\}_{m\in\mathcal{M}_D}$. Therefore

$$\inf_{m\in\mathcal{M}_D} \|t - t_m\|^2 \leq |\mathcal{I}(t,\epsilon)|\epsilon^2.$$

Let us denote by $\mathcal{E}_D(t)$ the infimum of $|\mathcal{I}(t,\epsilon)|\epsilon^2$ taken over all $\epsilon > 0$ satisfying $|\mathcal{I}(t,\epsilon)| \leq D$. This is in fact the quantity that we shall bound, as indicated in Theorem 6.1 below.

Theorem 6.1. *Let $p \in (0,2]$, $\alpha > 1/p - 1/2$ and $R > 0$. For all $D \in \{1,\dots,n\}$ and $t \in \mathscr{B}(\alpha,p,R)$,*

$$\mathcal{E}_D(t) \leq C(\alpha,p) n R^2 D^{-2\alpha}.$$

We then get Theorem 3.5 as a straightforward consequence of Theorem 6.1.

6.2. Proof of Theorem 6.1: the main lines

We shall prove Theorem 6.1 by following the path of DeVore and Yu in [14] (Sect. 3). Here are the notions and notation that we will need along the proof. Let $p > 0$, $\alpha > 0$ and $t \in \mathscr{M}(r,n)$. For every subset I of $\{1,\dots,n\}$, let

$$\mathcal{E}_p(t,I) = \inf_{v\in\mathbb{R}^r} \left(\sum_{k\in I} \|t_k - v\|_r^p \right)^{1/p}.$$

We define the vector $t^{\sharp,\alpha,p}$ in $\mathbb{R}^n$ whose coordinates are

$$t_i^{\sharp,\alpha,p} = \sup_{I\ni i} |I|^{-(\alpha+1/p)} \mathcal{E}_p(t,I), \text{ for } i = 1,\dots,n,$$

where the supremum is taken over all the dyadic intervals I of $\{1,\dots,n\}$ that contain i. We denote by $\|.\|_{\ell_p}$ the (quasi-)norm defined on $\mathbb{R}^n$ by

$$\|u\|_{\ell_p} = \left(\sum_{i=1}^n |u_i|^p \right)^{1/p}$$

(that is a norm only for $p \geq 1$) and by $\|.\|_{\ell_p,I}$ its restriction to a subset I of $\{1,\dots,n\}$. We define on $\mathbb{R}^n$ the discrete Hardy-Littlewood maximal function M_p by

$$\big(M_p(u)\big)_i = \sup_{I\ni i} |I|^{-1/p} \|u\|_{\ell_p,I}, \text{ for } i = 1,\dots,n,$$

where the supremum is taken over all the dyadic intervals I of $\{1,\dots,n\}$ containing i. Last, we recall that every vector $u \in \mathbb{R}^n$ is identified with the function $u : i \in \{1,\dots,n\} \mapsto u_i$, hence the meaning of notation such as $u \leq v$ or u^q, for $u \in \mathbb{R}^n$, $v \in \mathbb{R}^n$ and $q > 0$.

The beginning of the proof directly results from the way the algorithm works out. A dimension D being fixed, choosing $\epsilon > 0$ as small as possible such that the algorithm generates a partition with at most D intervals leads to a first comparison between the quantities $\mathcal{E}_D(t)$ and $D^{-2\alpha}$, without making use of any particular hypothesis on t.

Proposition 6.2. *Let $\alpha > 0$ and $p(\alpha) = (\alpha + 1/2)^{-1}$. For all $D \in \{1,\dots,n\}$ and $t \in \mathscr{M}(r,n)$,*

$$\mathcal{E}_D(t) \leq C(\alpha) \|t^{\sharp,\alpha,2}\|_{\ell_{p(\alpha)}}^2 D^{-2\alpha}.$$

Proof. If $t^{\sharp,\alpha,2} = 0$, then, whatever $\epsilon > 0$, $\mathcal{E}_2\big(t, I_{(0,0)}\big) \leq \epsilon$, so $\mathcal{E}_D(t) = 0$, which completes the proof in that case. Let us now assume that $t^{\sharp,\alpha,2}$ is non-null, and let $\epsilon > 0$. If $\mathcal{E}_2\big(t, I_{(0,0)}\big) \leq \epsilon$, then $|\mathcal{I}(t,\epsilon)| = 1$. Else, let I be a dyadic interval that belongs to $\mathcal{I}(t,\epsilon)$, then I is a child of a dyadic interval $\tilde{I}$ such that

$$\epsilon < \mathcal{E}_2\big(t, \tilde{I}\big).$$

Using the definition of $t^{\sharp,\alpha,2}$, we get, for all $i \in \tilde{I}$,

$$\mathcal{E}_2\big(t, \tilde{I}\big) \leq \big|\tilde{I}\big|^{\alpha+1/2} t_i^{\sharp,\alpha,2}.$$

Since $I \subset \tilde{I}$, $|\tilde{I}| = 2|I|$ and $p(\alpha) = (\alpha + 1/2)^{-1}$, the last two inequalities lead, for all $i \in I$, to

$$\epsilon < 2^{1/p(\alpha)} |I|^{1/p(\alpha)} t_i^{\sharp,\alpha,2},$$

hence

$$\epsilon^{p(\alpha)} < 2 \sum_{i \in I} \big(t_i^{\sharp,\alpha,2}\big)^{p(\alpha)}.$$

Then we deduce by summing over all the intervals I in the partition $\mathcal{I}(t,\epsilon)$ that

$$|\mathcal{I}(t,\epsilon)| \leq 2 \|t^{\sharp,\alpha,2}\|_{\ell_{p(\alpha)}}^{p(\alpha)} \epsilon^{-p(\alpha)}.$$

Whether $\mathcal{E}_2\big(t, I_{(0,0)}\big) \leq \epsilon$ or not, by choosing $\epsilon = 2^{1/p(\alpha)} \|t^{\sharp,\alpha,2}\|_{\ell_{p(\alpha)}} D^{-1/p(\alpha)}$, we get a partition $\mathcal{I}(t,\epsilon)$ that contains at most D elements and satisfies

$$|\mathcal{I}(t,\epsilon)| \epsilon^2 \leq D^{1-2/p(\alpha)} 2^{2/p(\alpha)} \|t^{\sharp,\alpha,2}\|_{\ell_{p(\alpha)}}^2.$$

As $p(\alpha) = (\alpha + 1/2)^{-1}$, we conclude that

$$|\mathcal{I}(t,\epsilon)| \epsilon^2 \leq 4^{\alpha+1/2} \|t^{\sharp,\alpha,2}\|_{\ell_{p(\alpha)}}^2 D^{-2\alpha}. \qquad \square$$

The proof of Theorem 6.1 now relies upon three inequalities. The first one allows to draw a comparison between $\mathcal{E}_D(t)$ and $D^{-2\alpha}$ *via* a term that does not depend on $t^{\sharp,\alpha,2}$ anymore but on $t^{\sharp,\alpha,p(\alpha)}$. It is the discrete analogue of a particular case of Theorem 4.3. of [13].

Proposition 6.3. *Let $\alpha > 0$ and $p(\alpha) = (\alpha + 1/2)^{-1}$. For all $t \in \mathscr{M}(r,n)$,*

$$t^{\sharp,\alpha,2} \leq C(\alpha) M_{p(\alpha)}\big(t^{\sharp,\alpha,p(\alpha)}\big).$$

For $\alpha > 0$, $p(\alpha) = (\alpha + 1/2)^{-1}$ and $D \in \{1, \dots, n\}$, Propositions 6.2 and 6.3 immediately lead to

$$\mathcal{E}_D(t) \leq C(\alpha) \big\| M_{p(\alpha)}\big(t^{\sharp,\alpha,p(\alpha)}\big)\big\|^2_{\ell_{p(\alpha)}} D^{-2\alpha}.$$

Let us now fix $p \in (0, 2]$. By Jensen's inequality, we have

$$\big\| M_{p(\alpha)}\big(t^{\sharp,\alpha,p(\alpha)}\big)\big\|_{\ell_{p(\alpha)}} \leq n^{1/p(\alpha)-1/p} \big\| M_{p(\alpha)}\big(t^{\sharp,\alpha,p(\alpha)}\big)\big\|_{\ell_p}$$

and

$$t^{\sharp,\alpha,p(\alpha)} \leq t^{\sharp,\alpha,p},$$

hence

$$\mathcal{E}_D(t) \leq C(\alpha) n^{2(\alpha+1/2-1/p)} \big\| M_{p(\alpha)}(t^{\sharp,\alpha,p})\big\|^2_{\ell_p} D^{-2\alpha}.$$

The following maximal inequality (inequality (6.15) below) ensures a control of u over its maximal functions. It is in fact the discrete version of the Hardy-Littlewood maximal inequality, that may be found in [4] (Thm. 3.10, p. 125).

Proposition 6.4. *Let $q > 1$. For all $u \in \mathbb{R}^n$,*

$$\|M_1(u)\|_{\ell_q} \leq C(q) \|u\|_{\ell_q}.$$

Since the maximal function M_q, $q > 0$, is related to M_1 by the property

$$M_q(u) = \big(M_1(u^q)\big)^{1/q}, \text{ for all } u \in \mathbb{R}^n,$$

Proposition 6.4 yields, for all $r > q > 0$ and $u \in \mathbb{R}^n$,

$$\|M_q(u)\|_{\ell_r} \leq C(r, q) \|u\|_{\ell_r}. \tag{6.15}$$

Thus, when applied with $u = t^{\sharp,\alpha,p}$, $r = p$ and $q = p(\alpha)$, this inequality leads to

$$\mathcal{E}_D(t) \leq C(\alpha, p) n^{2(\alpha+1/2-1/p)} \|t^{\sharp,\alpha,p}\|^2_{\ell_p} D^{-2\alpha}.$$

Last, Proposition 6.5 below provides the adequate control of the ℓ_p-(quasi-)norm of $t^{\sharp,\alpha,p}$ by the size of the wavelet coefficients of t and allows to complete immediately the proof of Theorem 6.1.

Proposition 6.5. *Let $p \in (0, 2]$ and $\alpha > 1/p - 1/2$. For all $t \in \mathscr{M}(r, n)$,*

$$\|t^{\sharp,\alpha,p}\|_{\ell_p} \leq C(\alpha, p) n^{-(\alpha+1/2-1/p)} \left(\sum_{j=0}^{N-1} 2^{jp(\alpha+1/2-1/p)} \sum_{\lambda \in \Lambda(j)} \|\beta_\lambda\|_r^p \right)^{1/p},$$

where, for all $\lambda \in \Lambda$, β_λ is the column-vector of $\mathbb{R}^r$ with lth line $\beta_\lambda^{(l)} = \langle t^{(l)}, \phi_\lambda \rangle_n$, for $l = 1, \dots, r$.

6.3. Proof of Proposition 6.3

The proof of Proposition 6.3 mostly relies on a lemma that we demonstrate after introducing some notation. Let I be a dyadic interval of $\{1, \dots, n\}$, $t \in \mathscr{M}(r, n)$, and $p > 0$. By a compactness argument, there exists at least one vector in $\mathbb{R}^r$, denoted by $v_p(t, I)$, satisfying

$$\mathcal{E}_p(t, I) = \left(\sum_{k \in I} \|t_k - v_p(t, I)\|_r^p \right)^{1/p}.$$

We define the vectors $u_p(t,I)$ and $t^{\sharp,\alpha,p,I}$ in $\mathbb{R}^n$ whose coordinates are null outside of I and given otherwise respectively by

$$\big(u_p(t,I)\big)_i = \|t_i - v_p(t,I)\|_r, \text{ for } i \in I,$$

and

$$t_i^{\sharp,\alpha,p,I} = \sup_{I \supset J \ni i} |J|^{-(\alpha+1/p)} \mathcal{E}_p(t,J), \text{ for } i \in I,$$

where the supremum is taken over all dyadic intervals J of $\{1,\dots,n\}$ that are contained in I and contain i. Last, for $u \in \mathbb{R}^n$, we denote by $u^\star$ its decreasing rearrangement, *i.e.* the $\mathbb{R}^n$-vector satisfying

$$u_1^\star \geq u_2^\star \geq \ldots \geq u_n^\star \text{ and } \{u_i^\star; 1 \leq i \leq n\} = \{|u_i|; 1 \leq i \leq n\}.$$

Lemma 6.6. *Let $\alpha > 0$, $p > 0$ and $t \in \mathscr{M}(r,n)$. Let I be a dyadic interval of $\{1,\dots,n\}$ containing at least two elements. For all $j \in \{1,\dots,|I|/2\}$,*

$$\big(u_p(t,I)\big)_j^\star \leq C(\alpha,p)\left(\sum_{k=j}^{|I|/2} k^{\alpha-1}\big(t^{\sharp,\alpha,p,I}\big)_k^\star + j^\alpha \big(t^{\sharp,\alpha,p,I}\big)_j^\star\right).$$

Proof. We fix $j \in \{1,\dots,|I|/2\}$. Let E be the set composed of all the indices i in $\{1,\dots,n\}$ satisfying $(t^{\sharp,\alpha,p,I})_i > (t^{\sharp,\alpha,p,I})_j^\star$. As $|E| \leq j-1$, we only have to prove that

$$\big(u_p(t,I)\big)_i \leq C(\alpha,p)\left(\sum_{k=j}^{|I|/2} k^{\alpha-1}\big(t^{\sharp,\alpha,p,I}\big)_k^\star + j^\alpha \big(t^{\sharp,\alpha,p,I}\big)_j^\star\right) \tag{6.16}$$

for all the indices $i \in \{1,\dots,n\}$, except maybe for those belonging to E. Consider $i \in \{1,\dots,n\}$ such that $i \notin E$. If $i \notin I$, then $\big(u_p(t,I)\big)_i = 0$, so inequality (6.16) is trivial. Suppose now that $i \in I$ and $i \notin E$, and let $\{I_l\}_{1 \leq l \leq m}$ be the sequence of dyadic intervals defined by

$$I_1 = I,\ I_{l+1} \text{ is the child of } I_l \text{ containing } i, \text{ and } I_m = \{i\},$$

where $m \geq 2$ because $|I| \geq 2$. Notice that, for all $l \in \{0,\dots,m-1\}$, $|I_{l+1}| = 2^{-l}|I|$. Let q be the strictly positive integer such that

$$2^{-(q+1)}|I| < j \leq 2^{-q}|I|.$$

That definition implies, in particular, that $2^{-q}|I| \geq 1$, hence $q < m$. From the triangle inequality,

$$\big(u_p(t,I)\big)_i \leq \sum_{l=2}^{q} \|v_p(t,I_{l-1}) - v_p(t,I_l)\|_r + \sum_{l=q+1}^{m} \|v_p(t,I_{l-1}) - v_p(t,I_l)\|_r, \tag{6.17}$$

with the convention that the first sum in inequality (6.17) is null for $q = 1$. Let us fix $l \in \{2,\dots,m\}$ and determine an upper-bound for the term $\|v_p(t,I_{l-1}) - v_p(t,I_l)\|_r$. We recall that $I_l \subset I_{l-1}$ and $|I_{l-1}| = 2|I_l|$. Besides, for all $p > 0$, the (quasi-)norm $\|.\|_{\ell_p}$ satisfies a triangle inequality within a multiplicative constant $C(p)$, where we can take $C(p) = 1$ for $p \geq 1$, and $C(p) = 2^{1/p}$ for $0 < p < 1$. Therefore, we get

$$\|v_p(t,I_{l-1}) - v_p(t,I_l)\|_r \leq C(p)|I_l|^{-1/p}\Big(\mathcal{E}_p(t,I_{l-1}) + \mathcal{E}_p(t,I_l)\Big),$$

which leads to

$$\|v_p(t,I_{l-1}) - v_p(t,I_l)\|_r \leq C(\alpha,p)|I_l|^\alpha \min_{k \in I_l} t_k^{\sharp,\alpha,p,I}. \tag{6.18}$$

Let us bound the first sum appearing in (6.17). For all $l \in \{2, \dots, m\}$, we have

$$\min_{k \in I_l} t_k^{\sharp,\alpha,p,I} \leq \left(t^{\sharp,\alpha,p,I}\right)^\star_{|I_l|} = \min_{1 \leq k \leq |I_l|} \left(t^{\sharp,\alpha,p,I}\right)^\star_k,$$

and, as $|I_{l+1}| = |I_l|/2$,

$$|I_l|^\alpha = C(\alpha) \int_{|I_{l+1}|}^{|I_l|} x^{\alpha-1} \, \mathrm{d}x \leq C(\alpha) \sum_{k=|I_{l+1}|}^{|I_l|} k^{\alpha-1}.$$

Consequently, when $q \geq 2$, inequality (6.18) yields

$$\begin{aligned} \sum_{l=2}^{q} \|v_p(t, I_{l-1}) - v_p(t, I_l)\|_r &\leq C(\alpha, p) \sum_{l=2}^{q} \sum_{k=|I_{l+1}|}^{|I_l|} k^{\alpha-1} \left(t^{\sharp,\alpha,p,I}\right)^\star_k \\ &\leq C(\alpha, p) \sum_{k=j}^{|I|/2} k^{\alpha-1} \left(t^{\sharp,\alpha,p,I}\right)^\star_k. \end{aligned}$$

Regarding the second sum appearing in (6.17), we now use inequality (6.18) combined with the following remarks. For all l such that $q+1 \leq l \leq m$, we have $\min_{k \in I_l} t_k^{\sharp,\alpha,p,I} \leq t_i^{\sharp,\alpha,p,I}$, since I_l contains i, and we recall that $|I_l| = 2^{-(l-1)}|I|$. Therefore,

$$\sum_{l=q+1}^{m} \|v_p(t, I_{l-1}) - v_p(t, I_l)\|_r \leq C(\alpha, p) |I|^\alpha \left(t^{\sharp,\alpha,p,I}\right)_i \sum_{l=q+1}^{m} 2^{-(l-1)\alpha}.$$

Furthermore, remember that $2^{-(q+1)}|I| < j$ and $i \notin E$, so we finally obtain

$$\sum_{l=q+1}^{m} \|v_p(t, I_{l-1}) - v_p(t, I_l)\|_r \leq C(\alpha, p) j^\alpha \left(t^{\sharp,\alpha,p,I}\right)^\star_j.$$

We have thus proved inequality (6.16) and Lemma 6.6. □

Let us now prove Proposition 6.3. Let $\alpha > 0$, $p(\alpha) = (\alpha + 1/2)^{-1}$, $t \in \mathscr{M}(r, n)$ and $i \in \{1, \dots, n\}$. From the definition of $\mathcal{E}_2(t, I)$ for a subset I of $\{1, \dots, n\}$, and due to the fact that $\mathcal{E}_2(t, \{i\}) = 0$, we have

$$t_i^{\sharp,\alpha,2} \leq \sup_{I \ni i} |I|^{-1/p(\alpha)} \|u_{p(\alpha)}(t, I)\|_{\ell_2},$$

where the supremum is taken over all the dyadic intervals I of $\{1, \dots, n\}$ that contain i, except for $\{i\}$. We fix such an interval I. The sequence $\left\{\left(u_{p(\alpha)}(t, I)\right)^\star_j\right\}_{1 \leq j \leq n}$ decreases and is null for $j \geq |I| + 1$, hence

$$\left\|u_{p(\alpha)}(t, I)\right\|_{\ell_2}^2 \leq 2 \sum_{j=1}^{|I|/2} \left(\left(u_{p(\alpha)}(t, I)\right)^\star_j\right)^2.$$

For $0 < p, q < +\infty$, we denote by $\|.\|_{\ell_{p,q}}$ the Lorentz (quasi-)norm defined on $\mathbb{R}^n$ by

$$\|u\|_{\ell_{p,q}} = \left(\sum_{i=1}^{n} i^{-1} (i^{1/p} u_i^\star)^q\right)^{1/q}.$$

For all subset I of $\{1,\dots,n\}$, we denote by $\|.\|_{\ell_{p,q},I}$ the restriction of $\|.\|_{\ell_{p,q}}$ to I. In particular, notice that, for all $u\in\mathbb{R}^n$ and $0<p,q<+\infty$,

$$\|u\|_{\ell_{p,p}}=\|u\|_{\ell_p}\ \text{ and }\ \|u^\star\|_{\ell_{p,q}}=\|u\|_{\ell_{p,q}}.$$

From Lemma 6.6 and the definition of $p(\alpha)$, we get

$$\big\|u_{p(\alpha)}(t,I)\big\|_{\ell_2}^2\le C(\alpha)\left(\sum_{j=1}^{|I|/2}j^{-1}\left(j^{1/2}\sum_{k=j}^{|I|/2}k^{\alpha-1}\big(t^{\sharp,\alpha,p(\alpha),I}\big)_k^\star\right)^2+\big\|\big(t^{\sharp,\alpha,p(\alpha),I}\big)^\star\big\|_{\ell_{p(\alpha),2}}^2\right).$$

Using a discrete version of Hardy's inequality (3.19), in [4] (p. 124), and noticing that $t^{\sharp,\alpha,p(\alpha),I}\le t^{\sharp,\alpha,p(\alpha)}$, we are led to

$$\big\|u_{p(\alpha)}(t,I)\big\|_{\ell_2}\le C(\alpha)\|t^{\sharp,\alpha,p(\alpha)}\|_{\ell_{p(\alpha),2},I}.$$

Last, since $p(\alpha)<2$, we conclude thanks to classical inequalities between Lorentz (quasi-)norms (*cf.* [4], Prop. 4.2, p. 217) that

$$t_i^{\sharp,\alpha,2}\le C(\alpha)\sup_{I\ni i}|I|^{-1/p(\alpha)}\|t^{\sharp,\alpha,p(\alpha)}\|_{\ell_{p(\alpha),I}}$$

where the supremum is taken over all the dyadic intervals I of $\{1,\dots,n\}$ that contain i.

6.4. **Proof of Proposition 6.4**

Let $q>1$ and $u\in\mathbb{R}^n$. As $M_1(u)=M_1(|u|)$, we can suppose that u has positive or null coordinates. Let us first demonstrate that, for all $i\in\{1,\dots,n\}$,

$$(M_1(u))_i^\star\le C\left(i^{-1}\sum_{k=1}^{i}u_k^\star\right).\tag{6.19}$$

If $i=1$, then this inequality easily follows from the definitions of $(M_1(u))_1^\star$ and $u_1^\star$. Let us now fix $i\in\{2,\dots,n\}$. We can write u as $u=v+w$, where v and w are the $\mathbb{R}^n$-vectors whose respective coordinates are

$$v_k=\max\{u_k-u_i^\star,0\}\ \text{ and }\ w_k=\min\{u_k,u_i^\star\},\ \text{ for } k=1,\dots,n.$$

From the triangle inequality, we deduce that $M_1(u)\le M_1(v)+M_1(w)$. Properties of discrete decreasing rearrangements analogous to Inequalities (1.14) and (1.16), in [4] (p. 41), then lead to

$$(M_1(u))_i^\star\le(M_1(v))_{\lceil i/2\rceil}^\star+(M_1(w))_{\lfloor i/2\rfloor}^\star.$$

Moreover,

$$(M_1(w))_{\lfloor i/2\rfloor}^\star\le\|M_1(w)\|_{\ell_\infty}\le\|w\|_{\ell_\infty},$$

and, from the discrete version of Theorem 3.3, in [4] (p. 119),

$$(M_1(v))_{\lceil i/2\rceil}^\star\le 2i^{-1}\|v\|_{\ell^1}.$$

Consequently,

$$(M_1(u))_i^\star\le C\big(i^{-1}\|v\|_{\ell^1}+\|w\|_{\ell^\infty}\big).\tag{6.20}$$

Let I be the set of all the indices $l\in\{1,\dots,n\}$ such that $u_l>u_i^\star$. From the definitions of v and w, we get

$$\|v\|_{\ell^1}+i\|w\|_{\ell^\infty}\le\sum_{k=1}^{|I|}u_k^\star+(i-|I|)u_i^\star=\sum_{k=1}^{i}u_k^\star,$$

which, given inequality (6.20), completes the proof of (6.19). We now have

$$\|(M_1(u))^\star\|_{\ell_q}^q \leq C(q)\sum_{i=1}^n \left(i^{-1}\sum_{k=1}^i u_k^\star\right)^q. \tag{6.21}$$

Let us denote by q' the conjugate exponent of q and write, for all $1 \leq k \leq n$, $u_k^\star = k^{-1/qq'}k^{1/qq'}u_k^\star$. We deduce from Hölder's inequality that

$$\sum_{i=1}^n \left(i^{-1}\sum_{k=1}^i u_k^\star\right)^q \leq \sum_{i=1}^n \left(q' i^{-1/q}\right)^{q/q'} \left(i^{-1}\sum_{k=1}^i k^{1/q'}(u_k^\star)^q\right).$$

Interchanging the order of the summations, we obtain

$$\sum_{i=1}^n \left(i^{-1}\sum_{k=1}^i u_k^\star\right)^q \leq C(q)\sum_{k=1}^n (u_k^\star)^q.$$

Consequently,

$$\|(M_1(u))^\star\|_{\ell_q} \leq C(q)\|u^\star\|_{\ell_q},$$

hence Proposition 6.4.

6.5. **Proof of Proposition 6.5**

Let $p \in (0,2]$, $\alpha > 1/p - 1/2$ and $t \in \mathscr{M}(r,n)$. For all $i \in \{1,\dots,n\}$ and all $0 \leq J \leq N$, we denote by $I(J,i)$ the only dyadic interval of length $n2^{-J}$ that is contained in $\{1,\dots,n\}$ and contains i. From the definition of $t^{\sharp,\alpha,p}$, we deduce

$$\|t^{\sharp,\alpha,p}\|_{\ell_p}^p \leq \sum_{J=0}^{N-1}(n^{-1}2^J)^{\alpha p+1}\sum_{i=1}^n \Big(\mathcal{E}_p\big(t, I(J,i)\big)\Big)^p. \tag{6.22}$$

Let us first suppose that $0 < p \leq 1$. From the definition of $\mathcal{E}_p(t, I(J,i))$, we have

$$\Big(\mathcal{E}_p\big(t, I(J,i)\big)\Big)^p \leq \sum_{k\in I(J,i)} \|t_k - t_i\|_r^p.$$

For all $-1 \leq j \leq N-1$, the functions $\{\phi_\lambda\}_{\lambda\in\Lambda(j)}$ are constant over any dyadic interval of length $n2^{-(j+1)}$. Therefore, if k belongs to $I(J,i)$, then

$$t_k - t_i = \sum_{j=J}^{N-1}\sum_{\lambda\in\Lambda(j)} \beta_\lambda(\phi_{\lambda\,k} - \phi_{\lambda\,i}).$$

As $0 < p \leq 1$, we deduce from the classical inequality between ℓ_p-quasi-norm and ℓ_1-norm

$$\sum_{i=1}^n \Big(\mathcal{E}_p\big(t, I(J,i)\big)\Big)^p \leq 2n^{2-p/2}2^{-J}\sum_{j=J}^{N-1} 2^{jp(1/2-1/p)}\sum_{\lambda\in\Lambda(j)}\|\beta_\lambda\|_r^p.$$

Interchanging the order of the summations, we get

$$\|t^{\sharp,\alpha,p}\|_{\ell_p}^p \leq C(\alpha,p)n^{1-p(\alpha+1/2)}\sum_{j=0}^{N-1} 2^{jp(\alpha+1/2-1/p)}\sum_{\lambda\in\Lambda(j)}\|\beta_\lambda\|_r^p.$$

Let us now consider the case $1 < p \leq 2$. We fix $0 \leq J \leq N-1$ and define

$$T(J) = \sum_{j=J}^{N-1} \sum_{\lambda \in \Lambda(j)} \beta_\lambda \phi_\lambda.$$

As $t - T(J)$ is constant over any dyadic interval of length $n2^{-J}$,

$$\mathcal{E}_p\big(t, I(J,i)\big) = \mathcal{E}_p\big(T(J), I(J,i)\big).$$

This equality and the definition of $\mathcal{E}_p\big(T(J), I(J,i)\big)$ lead to

$$\begin{aligned}\sum_{i=1}^{n} \Big(\mathcal{E}_p\big(t, I(J,i)\big)\Big)^p &\leq \sum_{i=1}^{n} \sum_{k \in I(J,i)} \big\|\big(T(J)\big)_k\big\|_r^p \\ &\leq n2^{-J} \sum_{k=1}^{n} \left(\sum_{j=J}^{N-1} \sum_{\lambda \in \Lambda(j)} \|\beta_\lambda\|_r |\phi_{\lambda\, k}| \right)^p.\end{aligned}$$

From (6.22) and this last inequality, we get

$$\|t^{\sharp,\alpha,p}\|_{\ell_p}^p \leq n^{-\alpha p} \sum_{k=1}^{n} \sum_{J=0}^{N-1} \left(2^{J\alpha} \sum_{j=J}^{N-1} \sum_{\lambda \in \Lambda(j)} \|\beta_\lambda\|_r |\phi_{\lambda\, k}| \right)^p.$$

Then, using one of Hardy's inequalities (*cf.* [12], Lem. 3.4, p. 27) and remembering that, for all $j \in \{-1, \ldots, N-1\}$, the functions $\{\phi_\lambda\}_{\lambda \in \Lambda(j)}$ have disjoint supports, we conclude that

$$\|t^{\sharp,\alpha,p}\|_{\ell_p}^p \leq C(\alpha,p) n^{-\alpha p} \sum_{j=0}^{N-1} 2^{j\alpha p} \sum_{\lambda \in \Lambda(j)} \|\beta_\lambda\|_r^p \sum_{k=1}^{n} |\phi_{\lambda\, k}|^p,$$

hence Proposition 6.5.

7. Lower bound for the minimax risk over $\mathscr{V}\mathscr{P}(\alpha, R)$

Proposition 7.1. *For all $\alpha > 0$, there exists a real $k_2(\alpha) \in (0,1)$ such that, if $R \geq n^{-1/2}$ and $R \leq k_2(\alpha) n^\alpha$, then*

$$\mathcal{R}_{\mathscr{V}}(\alpha, R) \geq C(\alpha)(nR^2)^{1/(2\alpha+1)}.$$

Proof. Let $R \geq 0$ and $0 \leq J \leq N$. For $\theta \in \{0,1\}^{2^J}$, we use the notation $\theta = (\theta_0, \ldots, \theta_{2^J-1})$. The Hamming distance between two elements θ and θ' in $\{0,1\}^{2^J}$ is $\delta(\theta, \theta') = \sum_{k=0}^{2^J-1} |\theta_k - \theta'_k|$. The Kullback-Leibler divergence is denoted by K. We shall construct a family $\{s_\theta\}_{\theta \in \Theta}$ of elements of $\mathscr{V}\mathscr{P}(\alpha, R)$ indexed by a properly chosen subset Θ of $\{0,1\}^{2^J}$. According to Birgé's version of Fano's lemma (*cf.* [21], Cor. 2.19), if $\max_{\theta,\theta' \in \Theta} K(s_\theta, s_{\theta'}) \leq \kappa \ln(|\Theta|)$, where κ is a universal constant that belongs to $(0,1)$, then

$$\mathcal{R}_{\mathscr{V}}(\alpha, R) \geq \frac{1-\kappa}{4} \eta^2$$

where $\eta^2 = \min_{\theta,\theta' \in \Theta, \theta \neq \theta'} \|s_\theta - s_{\theta'}\|^2$. Let us fix $\theta \in \{0,1\}^{2^J}$. We define $s_\theta \in \mathscr{M}(r,n)$ by

$$\begin{cases} s_\theta^{(1)} = 1/2 + a g_\theta \\ s_\theta^{(2)} = 1/2 - a g_\theta \\ s_\theta^{(l)} = 0 \text{ for } 3 \leq l \leq r, \text{ if } r \geq 3 \end{cases}$$

where $a = \sqrt{2} R 2^{-\alpha J}/4$ and $g_\theta = \sum_{k=0}^{2^J-1} (2\theta_k - 1) \mathbb{1}_{I_{(J,k)}}$. For all $1 \leq i < j \leq n$,

$$\|s_{\theta\, j} - s_{\theta\, i}\|_r = a\sqrt{2} |g_\theta(j) - g_\theta(i)|.$$

Since g_θ is constant on any dyadic interval of length $n/2^J$ and takes values in $\{-1,1\}$,

$$V_\alpha^{1/\alpha}(s_\theta) \leq 2^J (2\sqrt{2}a)^{1/\alpha} \leq R^{1/\alpha},$$

so $s_\theta \in \mathscr{V}(\alpha, R)$. Besides, as $\|g_\theta\|_\infty \leq 1$, if we assume that $R 2^{-\alpha J} \leq \sqrt{2}$, then $s_\theta \in \mathscr{P}$. Let us rather assume that $\sqrt{2} R 2^{-\alpha J} \leq 1$. We can then apply Lemma 4 of [15] and get, for all $\theta, \theta' \in \{0,1\}^{2^J}$,

$$K(s_\theta, s_{\theta'}) \leq 4 \|s_\theta - s_{\theta'}\|^2.$$

Now, according to Varshamov-Gilbert's lemma (*cf.* [21], Lem. 4.7, for instance), we can choose $\Theta \subset \{0,1\}^{2^J}$ such that, for any two distinct elements $\theta, \theta' \in \Theta$,

$$\delta(\theta, \theta') > 2^J/4 \tag{7.23}$$

and $\ln(|\Theta|) > 2^J/8$. For all $\theta, \theta' \in \{0,1\}^{2^J}$, $\|s_\theta - s_{\theta'}\|^2 = 8a^2 n 2^{-J} \delta(\theta, \theta')$, so, for all $\theta, \theta' \in \Theta$,

$$K(s_\theta, s_{\theta'}) \leq 2^8 a^2 n 2^{-J} \ln(|\Theta|)$$

and

$$\eta^2 \geq 2na^2.$$

Let us now set $k_2(\alpha) = \min\left(\sqrt{\kappa/2^5}, (2^5/\kappa)^\alpha 2^{-(\alpha+1/2)}\right)$ and assume that $n^{-1/2} \leq R \leq k_2(\alpha) n^\alpha$. We can then define J as the smallest integer in $\{0, \ldots, N\}$ such that $2^8 a^2 n 2^{-J} \leq \kappa$, *i.e.*

$$J = \min\left\{0 \leq j \leq N \text{ s.t. } (\kappa^{-1} 2^5 n R^2)^{1/(2\alpha+1)} \leq 2^j\right\}.$$

Such an integer J does satisfy $\sqrt{2} R 2^{-\alpha J} \leq 1$ and leads to

$$\mathcal{R}_{\mathscr{V}}(\alpha, R) \geq C(\kappa, \alpha) (nR^2)^{1/(2\alpha+1)}. \qquad \square$$

Acknowledgements. The author wishes to thank C. Durot for her advice along this work, and A.-S. Tocquet and E. Lebarbier for help in realizing Figure 6.

References

[1] M. Aerts and N. Veraverbeke, Bootstrapping a nonparametric polytomous regression model. *Math. Meth. Statist.* **4** (1995) 189–200.

[2] Y. Baraud and L. Birgé, Estimating the intensity of a random measure by histogram type estimators. *Prob. Theory Relat. Fields* **143** (2009) 239–284.

[3] A. Barron, L. Birgé and P. Massart, Risk bounds for model selection *via* penalization. *Prob. Theory Relat. Fields* **113** (1999) 301–413.
[4] C. Bennett and R. Sharpley, *Interpolation of operators*, volume 129 of Pure and Applied Mathematics. Academic Press Inc., Boston, M.A. (1988).
[5] L. Birgé, Model selection *via* testing: an alternative to (penalized) maximum likelihood estimators. *Ann. Inst. H. Poincaré Probab. Statist.* **42** (2006) 273–325.
[6] L. Birgé, Model selection for Poisson processes, in *Asymptotics: Particles, Processes and Inverse Problems, Festschrift for Piet Groeneboom. IMS Lect. Notes Monograph Ser.* **55**. IMS, Beachwood, USA (2007) 32–64.
[7] L. Birgé and P. Massart, Minimal penalties for Gaussian model selection. *Prob. Theory Relat. Fields* **138** (2007) 33–73.
[8] J.V. Braun and H.-G. Müller, Statistical methods for DNA sequence segmentation. *Stat. Sci.* **13** (1998) 142–162.
[9] J.V. Braun, R.K. Braun and H.-G. Müller, Multiple changepoint fitting *via* quasilikelihood, with application to DNA sequence segmentation. *Biometrika* **87** (2000) 301–314.
[10] T.H. Cormen, C.E. Leiserson, R.L. Rivest and C. Stein, *Introduction to algorithms*. Second edition. MIT Press, Cambridge, MA (2001).
[11] M. Csűrös, Algorithms for finding maximum-scoring segment sets, in *Proc. of the 4th international workshop on algorithms in bioinformatics 2004. Lect. Notes Comput. Sci.* **3240**. Springer, Berlin, Heidelberg (2004) 62–73.
[12] R.A. DeVore and G.G. Lorentz, *Constructive approximation*. Springer-Verlag, Berlin, Heidelberg (1993).
[13] R.A. DeVore and R.C. Sharpley, Maximal functions measuring smoothness. *Mem. Amer. Math. Soc.* **47** (1984) 293.
[14] R.A. DeVore and X.M. Yu, Degree of adaptive approximation. *Math. Comp.* **55** (1990) 625–635.
[15] C. Durot, E. Lebarbier and A.-S. Tocquet, Estimating the joint distribution of independent categorical variables *via* model selection. *Bernoulli* **15** (2009) 475–507.
[16] Y.-X. Fu and R.N. Curnow, Maximum likelihood estimation of multiple change points. *Biometrika* **77** (1990) 562–565.
[17] S. Gey S. and E. Lebarbier, *Using CART to detect multiple change-points in the mean for large samples*. SSB preprint, Research report No. 12 (2008).
[18] M. Hoebeke, P. Nicolas and P. Bessières, MuGeN: simultaneous exploration of multiple genomes and computer analysis results. *Bioinformatics* **19** (2003) 859–864.
[19] E. Lebarbier, *Quelques approches pour la détection de ruptures à horizon fini*. Ph.D. thesis, Université Paris Sud, Orsay, 2002.
[20] E. Lebarbier and E. Nédélec, *Change-points detection for discrete sequences via model selection*. SSB preprint, Research Report No. 9 (2007).
[21] P. Massart, *Concentration inequalities and model selection*. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003. *Lect. Notes Math.* **1896**. Springer, Berlin, Heidelberg (2007).
[22] P. Nicolas *et al.*, Mining *Bacillus subtilis* chromosome heterogeneities using hidden Markov models. *Nucleic Acids Res.* **30** (2002) 1418–1426.
[23] W. Szpankowski, L. Szpankowski and W. Ren, An optimal DNA segmentation based on the MDL principle. *Int. J. Bioinformatics Res. Appl.* **1** (2005) 3–17.