Change-points detection for discrete sequences via model selection

by

Emilie Lebarbier and Elodie Nédélec

Research Report No. 9
April 2007

Statistics for Systems Biology Group
Jouy-en-Josas/Paris/Evry, France
http://genome.jouy.inra.fr/ssb/

Change-points detection for discrete sequences via model selection

Émilie Lebarbier * Elodie Nédélec †

Abstract

We propose a method based on a penalized contrast criterion for estimating the change-points in a dicrete distribution of independant variables. The number of change-points and their locations are unknown. We consider two minimum contrast estimation: the maximum likelihood one and the least-squares one. In the two contexts we define the penalty function involved in our corresponding criterion such that the resulting estimator minimizes non asymptotically the associated risk.

Keyword : Non parametric estimation – regression – density – Model selection

1 Introduction

Our motivation comes from DNA analysis, in particular from the segmentation of a discrete sequence of letters Y_1, ..., Y_n taking their values in the finite DNA alphabet $\{A, T, C, G\}$. Biologists observe in DNA sequences areas with a stability of frequencies of the four letters which correspond to areas biologically significant. The aim of this paper is to provide statistical methods proposing an automatic segmentation of the sequence. This problem is abundantly treated in the literature (see Braun and Müller [3] for a complete bibliography). Churchill [6], Boys et al. [8] and Muri [7] consider a hidden Markov chain model assuming that the different areas of the DNA sequence can be classified into a fixed set of hidden states.
The DNA segmentation can be also put into the framework of multiple

*UMR INA P-G/ENGREF/INRA-MIA 518, 16 rue Claude Bernard, 75231 Paris cedex 05, France (e-mail: lebarbie@inapg.fr)

†Laboratoire de mathématiques - U.M.R. 8628 Université Paris XI, 91405 Orsay

change-points detection for a discrete distribution of variables. We define m a partition of the set $\{1, ..., n\}$ if there exist $k, a_1, a_2, ..., a_k \in \mathbb{N}$ satisfying $1 \leq a_1 < a_2 < ... < a_k \leq n$ such that

$$m = \{[1, a_1], [a_1 + 1, a_2], ..., [a_{k-1} + 1, a_k], [a_k + 1, n]\}. \tag{1.1}$$

The problem of selecting m such that for all $J \in m$, $(Y_i)_{i \in J}$ is a stationary sequence, is equivalent to the problem of estimating k and the change-points $a_1, a_2, ..., a_k \in \mathbb{N}$. Braun et al. [2] propose a penalized maximum likelihood estimation procedure. But this approach is asymptotic in the sense that they propose a penalty which leads to a consistent estimator of the number of change-points.
In this paper we propose a non asymptotic procedure based on the work of Birgé and Massart [1] about model selection which aims to estimate s with a small risk for a fixed n. In this context, if we suppose that the true function is piecewise constant, an estimator with less number of change-points compared to the true one could be preferred. This estimator is also obtained by a minimum penalized contrast procedure.

We consider Y_1, ..., Y_n independent variables taking their values in $\{1, 2, ..., r\}$ with $r \in \mathbb{N}$ and $r \geq 2$. We define for $t \in \{1, ...n\}$ and $i \in \{1, 2, ..., r\}$

$$P(Y_t = i) = s(t, i),$$

and we have in mind that $r = 4$ in DNA analysis. Remark that in this context, the large size of the sample n does not mean that we take place in an asymptotic approach since the underlying function s depends on n.
We are interested in the estimation of s by $\widehat{s}$ such that for all $i \in \{1, 2, ..., r\}$

$$\widehat{s}(t, i) = \widehat{s}(t', i)$$

for all t, $t' \in J$ and for all J segment of a "good" partition $\widehat{m}$ of $\{1, ..., n\}$. This is a model selection problem since we can select $\widehat{m}$ among a collection $\mathcal{M}_n$ of partitions defined in (1.1) constructed on the set $\{1, ...n\}$. On one hand we can see s as a vector of $\mathbb{R}^{nr}$ which is the mean of the vector $\left(\mathbb{1}_{\{Y_t = i\}}\right)_{1 \leq t \leq n, 1 \leq i \leq r}$ where $\mathbb{1}$ is the indicator function, therefore we have considered first the least-squares contrast. On the other hand we can see s as a density so we have then considered the contrast associated to the log-likelihood.

For each partition $m \in \mathcal{M}_n$, we compute the minimum contrast estimator of s, denoted by $\hat{s}_m$ and we define a collection of estimators $\{\hat{s}_m\}_{m \in \mathcal{M}_n}$.

If we consider the least-squares contrast, the ideal partition $m(s)$ would minimize with respect to $m \in \mathcal{M}_n$ the risk

$$R_1(s, \hat{s}_m) = \mathbb{E}_s\left[\|s - \hat{s}_m\|^2\right]$$

where $\|.\|$ is the euclidean norm on $\mathbb{R}^{nr}$, or the risk

$$R_2(s, \hat{s}_m) = nK(s, \hat{s}_m)$$

if we consider the contrast associated to the log-likelihood with K denoting the Kullback Leibler information. Unfortunately, the partition $m(s)$ depends on the unknown function s. The aim of the estimation procedure proposed in this paper is to provide a data-driven criterion that selects an estimator $\tilde{s}$ whose risk is as close as possible to the risk of $\hat{s}_{m(s)}$. Therefore, we consider some function $pen : \mathcal{M}_n \to \mathbb{R}_+$ which is called penalty function. We select

$$\hat{m} = \underset{m \in \mathcal{M}_n}{\operatorname{argmin}} \left\{\gamma(\hat{s}_m) + pen(m)\right\}, \tag{1.2}$$

where γ is one contrast and finally estimate s by the minimum penalized contrast estimator

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

The article is organised as follows : Section 2 is devoted to the presentation of the model selection procedure. In this section, we first present our model collection and give the minimum contrast estimators in a fix model for each procedure. Then the risks of these estimators are given and proved in Section 4. Finally we present our main result providing a form for the penalty function and an upper bound for the risk of the corresponding penalized estimator for the two considered contrasts. These results are proved in Section 5 and proofs used some results given in Section 3.

2 Model selection procedure

First we present the collection of models and the two contrasts, then we construct the corresponding collection of the minimum contrast estimators $\{\hat{s}_m\}_{m \in \mathcal{M}_n}$. We give their risks and select a final estimator among this collection by minimizing a penalized contrast. We propose a penalty and give an upper bound for the risk of the corresponding penalized estimator.

2.1 Presentation of models and collection of estimators

We observe Y_1, ..., Y_n independent variables taking their values in $\{1, 2, ..., r\}$ with $r \in \mathbb{N}$ and $r \geq 2$, and we define for $t \in \{1, ...n\}$ and $i \in \{1, 2, ..., r\}$

$$P(Y_t = i) = s(t, i), \tag{2.3}$$

where s is unknown. Let $\mathcal{M}_n$ be a collection of partitions of $\{1, .., n\}$ defined by (1.1) and for $m \in \mathcal{M}_n$, we define the associated model $\mathcal{S}_m$ by

$$\mathcal{S}_m = \left\{ \begin{array}{c} u : \{1, ..., n\} \times \{1, ..., r\} \to [0, 1] \text{ such that} \\ \forall J \in m, \forall i \in \{1, 2, ..., r\}, u(t, i) = u(t', i) \ = u(J, i) \quad \forall\, t, t' \in J \\ \text{and } \sum_{i=1}^{r} u(J, i) = 1 \ \forall\ J \in m \end{array} \right\},$$

and note

$$\mathcal{S} = \bigcup_{m \in \mathcal{M}_n} \mathcal{S}_m. \tag{2.4}$$

For the estimation procedure we consider two different contrasts:

- First the least-square contrast γ_1 defined by

$$\gamma_1(u) = -2 \sum_{i=1}^{r} \sum_{t=1}^{n} \mathbb{1}_{\{Y_t = i\}} u(t, i) + ||u||^2,$$

  for $u \in \mathcal{S}$ where $\mathbb{1}$ is the indicator function and $\|.\|$ is the euclidean norm on the space $\mathbb{R}^{nr}$ defined for $u \in \mathcal{S}$ by

$$||u||^2 = \sum_{t=1}^{n} \sum_{i=1}^{r} u^2(t, i).$$

- Then the log-likelihood contrast γ_2 defined for $u \in \mathcal{S}$ by

$$\begin{aligned} \gamma_2(u) &= -\log \left[\prod_{t=1}^{n} \prod_{i=1}^{r} u(t, i)^{\mathbb{1}_{\{Y_t = i\}}} \right] \\ &= -\sum_{t=1}^{n} \sum_{i=1}^{r} \mathbb{1}_{\{Y_t = i\}} \log [u(t, i)]. \end{aligned}$$

The minimum estimator, denoted $\hat{s}_m$, minimizes the contrast fonction γ over the defined model. For the log-likelihood framework, the contrast

is γ_2 and the model associated to m is $\mathcal{S}_m$ (defined just below). For the least-square framework, the contrast is γ_1 but the model is not exactly $\mathcal{S}_m$. Indeed in this case s represents the mean of the vector $\left(\mathbb{1}_{\{Y_t=i\}}\right)_{1\leq t\leq n, 1\leq i\leq r}$ and the condition that s is the probability against the counting measure on the set $\{1,\ldots,r\}$ is not requires. The model in this case is the same as $\mathcal{S}_m$ without the last condition. But for a sake of simplicity, we use the same notation $\mathcal{S}_m$ for the two models.
However, in the two frameworks, we obtain the same collection of minimum contrast estimators $(\hat{s}_m)_{m\in\mathcal{M}_n}$ where for a given partition $m \in \mathcal{M}_n$

$$\hat{s}_m(t,i) = \frac{N_J(i)}{\mid J \mid}, \text{ for } i \in \{1,...,r\}, t \in J \text{ and } J \in m, \tag{2.5}$$

with

$$N_J(i) = \sum_{t\in J} \mathbb{1}_{\{Y_t=i\}},$$

and $\mid J \mid$ is the cardinality of J.

2.2 The risk of the minimum contrast estimator on a given partition m

To analyze the risks of the estimators defined in (2.5), we introduce a loss function l associated to the contrast γ by the relation $l(u,v) = \mathbb{E}_s[\gamma(v) - \gamma(u)]$ for $u, v \in \mathcal{S}$ where s is defined in (2.3) and obtain two loss functions for our two different contrasts. First the loss function l_1 associated to γ_1 satisfies

$$\begin{aligned} l_1(s,v) &= -2\sum_{t=1}^{n}\sum_{i=1}^{r} s(t,i)\left[v(t,i) - s(t,i)\right] + ||v||^2 - ||s||^2 \\ &= ||s-v||^2 \end{aligned} \tag{2.6}$$

for $v \in \mathcal{S}$. This leads to the following first risk R_1 for the minimum contrast estimator $\widehat{s}_m$ defined in (2.5) for $m \in \mathcal{M}_n$

$$R_1(s,\widehat{s}_m) = \mathbb{E}_s\left[\|s-\widehat{s}_m\|^2\right]. \tag{2.7}$$

And then for the log-likelihood contrast γ_2, the loss function l_2 associated to γ_2 satisfies

$$l_2(s,v) = \sum_{t=1}^{n}\sum_{i=1}^{r} s(t,i)\log\left[\frac{s(t,i)}{v(t,i)}\right] \tag{2.8}$$

for $v \in \mathcal{S}$. We can remark that $l_2(s,v) = nK(s,v)$ where K is the Kullback-Leibler information between s and v, since s and v can be considered as densities against $\frac{1}{n}$ times the counting measure on the set $\{1,2,...,r\} \times \{1,2,...,n\}$. So the loss function is in this case, up to a constant n, the Kullback-Leibler information classical in a maximum likelihood estimation. This leads to the following second risk R_2 for the minimum contrast estimator $\widehat{s}_m$

$$R_2(s,\widehat{s}_m) = \mathbb{E}_s\left[nK(s,\widehat{s}_m)\right]. \tag{2.9}$$

Our purpose is now to evaluate the risks R_1 and R_2 defined respectively in (2.7) and in (2.9). Our motivation is to provide a benchmark in order to judge the performance of the final penalized estimator that will be defined and studied in the next section. Whatever the loss function l_1 or l_2 defined respectively in (2.6) and (2.8), we obtain for every $m \in \mathcal{M}_n$ the same projection $\bar{s}_m$ of s on the model $\mathcal{S}_m$ defined by

$$\begin{aligned} \bar{s}_m(t,i) &= \underset{u \in \mathcal{S}_m}{\operatorname{argmin}}\, l(s,u) \\ &= \frac{\sum_{t' \in J} s(t',i)}{\mid J \mid} \end{aligned} \tag{2.10}$$

for $i \in \{1,...,r\}$, $t \in J$ and $J \in m$.

Proposition 2.1. *Let m be a partition of the grid $\{1,...,n\}$, $\widehat{s}_m$ be the minimum contrast estimator of s defined by (2.5) and $\bar{s}_m$ be the projection of s given by (2.10). Assume that there exists some positive absolute constant ρ such that:*

$$s \geq \rho. \tag{2.11}$$

1. *Let R_1 be given by (2.7), then*

$$R_1(s,\widehat{s}_m) \leq l_1(s,\bar{s}_m) + (r-1)\,|m|,$$

and

$$R_1(s,\widehat{s}_m) \geq l_1(s,\bar{s}_m) + \rho\,(r-1)\,|m|.$$

2. *Let R_2 be given by (2.9) and suppose that for all $J \in m$*

$$\mid J \mid \geq \Gamma\left[\log(n)\right]^2$$

for an absolute constant $\Gamma > 0$. We have for all real numbers $\varepsilon > 0$ and $a > 1$

$$R_2(s,\hat{s}_m) \leq l_2(s,\bar{s}_m) + \frac{1+\varepsilon}{2\,(1-\varepsilon)^2}\; r\;|m| + \frac{C(\Gamma,\rho,a,r,\varepsilon)}{n^{a-1}}$$

and

$$R_2(s,\hat{s}_m) \geq l_2(s,\bar{s}_m) + \frac{1-\varepsilon}{4\left(1+\varepsilon\right)^2}\ \rho\ r\ |m| - \frac{C\left(\Gamma,\rho,a,r,\varepsilon\right)}{n^{a-1}},$$

where $C\left(\Gamma,\rho,a,r,\varepsilon\right)$ *is a positive constant only depending on* Γ*,* a*,* ρ*,* r *and* ε*.*

The proof is given in Section 4.

2.3 Model Selection

We have a collection of minimum contrast estimators $\{\hat{s}_m\}_{m\in\mathcal{M}_n}$ and we want to select the "best" estimator among this collection. We consider two different penalty functions denoted $pen_1 : \mathcal{M}_n \to \mathbb{R}_+$ for the least-squares procedure and $pen_2 : \mathcal{M}_n \to \mathbb{R}_+$ for the likelihood procedure. Selecting for $c \in \{1,2\}$

$$(\hat{m})_c = \arg\min_{m\in\mathcal{M}_n} \left\{\gamma_c\left(\hat{s}_m\right) + pen_c\left(m\right)\right\}, \tag{2.12}$$

where γ_c is the contrast, we finally estimate s by the two minimum penalized contrast estimators

$$\widetilde{s}_1 = \widehat{s}_{(\widehat{m})_1} \text{ and } \widetilde{s}_2 = \widehat{s}_{(\widehat{m})_2}. \tag{2.13}$$

Before giving our main result, let us see the definition of the squared Hellinger distance

Definition 2.2. *The squared Hellinger distance denoted by* h^2 *between two positive densities* p *and* q *with respect to* μ *is defined by*

$$h^2\left(p,q\right) = \frac{1}{2}\int \left(\sqrt{p}-\sqrt{q}\right)^2 d\mu. \tag{2.14}$$

The Hellinger distance and the Kullback-Leibler information are related by the well-known inequality

$$K\left(p,q\right) \geq 2h^2\left(p,q\right). \tag{2.15}$$

2.3.1 The main theorem

The following result provides a model selection criterion based on a minimum penalized contrast method. It gives the penalties and some non asymptotic risk bounds on the performances of the associated penalized estimator $\tilde{s}_c$ for $c \in \{1,2\}$ defined by (2.13). Note that in the log-likelihood case, we assume that all partitions $m \in \mathcal{M}_n$ are "constructed on a partition" m_f which means that there exists some partition m_f which is a refinement of every $m \in \mathcal{M}_n$.

Theorem 2.3. *Suppose that one observes independent variables Y_1, ..., Y_n taking their values in $\{1,2,...,r\}$ with $r \in \mathbb{N}$ and $r \geq 2$. We define for $t \in \{1,...n\}$ and $i \in \{1,2,...,r\}$*

$$P\left(Y_t = i\right) = s\left(t,i\right)$$

and consider a collection $\mathcal{M}_n$ of partitions constructed on the grid $\{1,...,n\}$. Let $(L_m)_{m\in\mathcal{M}_n}$ be some family of positive weights and define Σ as

$$\Sigma = \sum_{m\in\mathcal{M}_n} \exp\left(-L_m|m|\right) < +\infty. \tag{2.16}$$

1. Let $K > 1$. If for every $m \in \mathcal{M}_n$

$$pen_1\left(m\right) \geq K\ r\ |m|\ \left(\frac{1}{4} + 2\sqrt{L_m} + \ 2L_m\right),$$

then

$$R_1\left(s,\tilde{s}_1\right) \leq 8\left(\frac{K+1}{K-1}\right)^3 \left[\inf_{m\in\mathcal{M}_n}\left\{l_1\left(s,\bar{s}_m\right) + pen_1\left(m\right)\right\} + K\left(r-1\right)^2\ \Sigma\right].$$

2. Assume that

- *there exists some positive absolute constant ρ such that $s \geq \rho$,*
- *$\mathcal{M}_n$ is a collection of partitions constructed on a partition m_f such that*

$$|J| \geq \Gamma\left[\log\left(n\right)\right]^2 \forall J \in m_f \tag{2.17}$$

where Γ is a positive absolute constant.
Let $\lambda > 1$. If for every $m \in \mathcal{M}_n$

$$pen_2\left(m\right) \geq \lambda\ r|m|\ \left(\frac{1}{2} + 4\ \sqrt{L_m} + 4\ L_m\right), \tag{2.18}$$

then

$$\mathbb{E}_s\left[nh^2\left(s,\tilde{s}_2\right)\right] \leq \frac{2\lambda^{1/3}}{\lambda^{1/3}-1}\inf_{m\in\mathcal{M}_n}\left\{l_2\left(s,\overline{s}_m\right)+pen_2\left(m\right)\right\}+C\left(\Sigma,r,\lambda,\rho,\Gamma\right),$$

where h^2 is the squared Hellinger distance given by (2.14).

Now in order to evaluate the performance of the penalized estimator, we want to compare $R_c\left(s,\tilde{s}_c\right)=\mathbb{E}_s\left[l_c\left(s,\tilde{s}_c\right)\right]$ to $\inf_{m\in\mathcal{M}_n}\mathbb{E}_s\left[l_c\left(s,\hat{s}_m\right)\right]$ for $c\in\{1,2\}$. In the next Section we discuss this comparison according to the choice of the weights $\{L_m, m\in\mathcal{M}_n\}$.

We remark that for the log-likelihood procedure, the risk of the penalized estimator $\tilde{s}_c$ is treated in terms of Hellinger distance instead of the Kullback-Leibler information. This is due to the fact that the Kullback-Leibler is possibly infinite, so the comparison can not be obtained in the same risk. However, we have the converse inequality of (2.15) up to constants whatever $||\log(p/q)||_\infty<\infty$ so with an hypothesis on s we will obtain a relation in terms of Hellinger risk.

2.3.2 Choice of the weights $\{L_m, m\in\mathcal{M}_n\}$

The penalty function depends on the family $\mathcal{M}_n$ through the choice of the weights L_m satisfying (2.16). We distinguish two different types of collection $\mathcal{M}_n$.

- an exhaustive collection**:** $\mathcal{M}_n$ is the collection of all possible partitions constructed on a partition m_f satisfying (2.17). In this context, the number of partitions having the dimension D is bounded by $\binom{n}{D}$. Taking L_m as a function of the dimension, $L_m=L_{|m|}$ leads to

$$\begin{aligned}
\Sigma &= \sum_{m\in\mathcal{M}_n} e^{-L_m|m|}=\sum_{D=1}^{n} e^{-DL_D}Card\{m\in\mathcal{M}_n,|m|=D\}\\
&\leq \sum_{D=1}^{n} e^{-DL_D}\binom{n}{D}\\
&\leq \sum_{D=1}^{n} e^{-DL_D}\left(\frac{en}{D}\right)^D\\
&\leq \sum_{D=1}^{n} e^{-D\left(L_D-1-\log\left(\frac{n}{D}\right)\right)}.
\end{aligned}$$

The condition (2.16) is satisfied if we take $L_D = 1 + \theta + \log\left(\frac{n}{D}\right)$ with $\theta > 0$. This leads to $\Sigma < (\exp(\theta) - 1)^{-1}$and a penalty function of the form for $c \in \{1, 2\}$,

$$pen_c(m) = r|m|\left(\alpha_c \log\left(\frac{n}{|m|}\right) + \beta_c\right),$$

where α_c and β_c are some absolute constants which need to be calibrated. Proposition 2.1 and Theorem 2.3 provide the following two upper bounds:

- For the leasts-square procedure, there exists C_1 and C_1'some absolute constants such that

$$\begin{aligned} R_1(s, \tilde{s}_1) &\leq C_1 \log(n) \inf_{m \in \mathcal{M}_n} \{l_1(s, \bar{s}_m) + (r-1)|m|\} + C_1'(r-1)^2 \\ &\leq \frac{C_1 \log(n)}{\rho} \inf_{m \in \mathcal{M}_n} \{R_1(s, \widehat{s}_m)\} + C_1'(r-1)^2. \end{aligned}$$

However if the selected partition $\widetilde{m}$ defined by (1.2) is such that $|\widetilde{m}| >> r$ then there exists an absolute constant C_1'' such that

$$R_1(s, \tilde{s}_1) \leq \frac{C_1'' \log(n)}{\rho} \inf_{m \in \mathcal{M}_n} \{R_1(s, \widehat{s}_m)\}.$$

- For the log-likelihood procedure, there exists C_2 and C_2' some absolute constants such that

$$\begin{aligned} \mathbb{E}_s\left[nh^2(s, \tilde{s}_2)\right] &\leq \frac{C_2 \log(n)}{\rho} \inf_{m \in \mathcal{M}_n} \{\rho l_2(s, \bar{s}_m) + \rho r|m|\} + C(r, \rho) \\ &\leq \frac{C_2' \log(n)}{\rho} \inf_{m \in \mathcal{M}_n} \{R_2(s, \widehat{s}_m)\} + C(r, \rho) \end{aligned}$$

with $C(\rho, r)$ only depending on ρ and r. However if the selected partition $\widetilde{m}$ defined by (1.2) is such that $|\widetilde{m}|$ is large then there exists an absolute constant C_2'' such that

$$R_2(s, \tilde{s}_2) \leq \frac{C_2'' \log(n)}{\rho} \inf_{m \in \mathcal{M}_n} \{R_2(s, \widehat{s}_m)\}.$$

In the two cases the penalized estimators $\tilde{s}_c$ for $c \in \{1, 2\}$ have the best risk as possible up to a $\log(n)$ factor. This $\log(n)$ should be unavoidable as in the Gaussian framework studied by Birgé and Massart in [1].

Remark 1. Recall that $R_2\left(s,\widehat{s}_m\right)=\mathbb{E}_s\left[nK\left(s,\widehat{s}_m\right)\right]$. However $K\left(s,\overline{s}_m\right)$ is close to $4h^2\left(s,\overline{s}_m\right)$ when $||\log(s/\overline{s}_m)||_\infty<\infty$ and under this assumption the above upper bound can be stated as follows

$$\mathbb{E}_s\left[nh^2\left(s,\tilde{s}_2\right)\right]\leq C_2'\log\left(n\right)\inf_{m\in\mathcal{M}_n}\mathbb{E}_s\left[nh^2\left(s,\widehat{s}_m\right)\right].$$

- a reduced collection**:** we consider here the CART algorithm described in [4]. The collection of partitions $\mathcal{M}_n$ is constructed by a recursive procedure which consists at each step to split a considered segment into two segments by minimizing the sum of the contrast calculated on the two segments. The result of this procedure can be seen as a binary tree. The collection $\mathcal{M}_n$ is random and the results are written conditionnaly to the sample on which $\mathcal{M}_n$ is constructed.

  On a first time we keep only the variables Y_i with i even. Conditionnaly to $(Y_i)_{i\text{ even}}$ the collection is deterministic. Taking L_m as a function of the dimension, $L_m=L_{|m|}$ leads to

$$\begin{aligned}\Sigma &= \sum_{m\in\mathcal{M}}e^{-L_m|m|}=\sum_{D=1}^{n}e^{-DL_D}Card\{m\in\mathcal{M}_n,|m|=D\}\\ &\leq \sum_{D=1}^{n}e^{-DL_D}\frac{1}{D}\binom{2(D-1)}{D-1}\\ &\leq \sum_{D=1}^{n}\frac{1}{D}e^{D(2\log 2-L_D)},\end{aligned}$$

  since for any dimension D, the number of partitions having the dimension D is bounded by the number of balanced binary trees having D final nodes called Catalan's number and equal to $\frac{1}{D}\binom{2(D-1)}{D-1}$. Consequently, if we take $L_D>2\log 2$ then $\Sigma<1$ and this leads to a penalty function of the form for $c\in\{1,2\}$,

$$pen_c(m)=\eta_c\ r|m|,$$

  where η_c is an absolute positive constant. On a second time we do the selection with the variables $(Y_i)_{i\text{ odd}}$ and we could prove paraphrasing the proof of Proposition 2.1 and Theorem 2.3 the two following inequalities of oracle type:

- For the least-squares procedure, there exists C_1 and C_1'some absolute constants such that

$$\begin{aligned}\mathbb{E}_s\left[\|s-\tilde{s}_1\|^2 \mid (Y_i)_{i \text{ even}}\right] &\leq C_1 \inf_{m\in\mathcal{M}_n}\left\{||s-\bar{s}_m||^2+(r-1)|m|\right\} \\ &\quad +C_1'(r-1)^2 \\ &\leq \frac{C_1}{\rho}\inf_{m\in\mathcal{M}_n}\mathbb{E}_s\left[\|s-\widehat{s}_m\|^2 \mid (Y_i)_{i \text{ even}}\right] \\ &\quad +C_1'(r-1)^2.\end{aligned}$$

However if the selected partition $\widetilde{m}$ defined by (1.2) is such that $|\widetilde{m}| >> r$ then there exists an absolute constant C_1'' such that

$$\mathbb{E}_s\left[\|s-\tilde{s}_1\|^2 \mid (Y_i)_{i \text{ even}}\right] \leq \frac{C_1''}{\rho}\inf_{m\in\mathcal{M}_n}\mathbb{E}_s\left[\|s-\widehat{s}_m\|^2 \mid (Y_i)_{i \text{ even}}\right].$$

- For the log-likelihood procedure, there exists C_2 and C_2' some absolute constants such that

$$\begin{aligned}\mathbb{E}_s\left[nh^2(s,\tilde{s}_2) \mid (Y_i)_{i \text{ even}}\right] &\leq C_2 \inf_{m\in\mathcal{M}_n}\left\{nK(s,\bar{s}_m)+r|m|\right\} \\ &\quad +C(\rho,r) \\ &\leq \frac{C_2'}{\rho}\inf_{m\in\mathcal{M}_n}\mathbb{E}_s\left[nK(s,\widehat{s}_m) \mid (Y_i)_{i \text{ even}}\right] \\ &\quad +C(\rho,r).\end{aligned}$$

where $C(\rho,r)$ only depends on ρ and r. However if the selected partition $\widetilde{m}$ defined by (1.2) is such that $|\widetilde{m}|$ is large then there exists an absolute constant C_2'' such that

$$\mathbb{E}_s\left[nh^2(s,\tilde{s}_2) \mid (Y_i)_{i \text{ even}}\right] \leq \frac{C_2''}{\rho}\inf_{m\in\mathcal{M}_n}\mathbb{E}_s\left[nK(s,\widehat{s}_m) \mid (Y_i)_{i \text{ even}}\right].$$

As we have seen, the difference between these two algorithms is the considered collection of partitions $\mathcal{M}_n$. When the sample size n is small, the exhaustive search is preferable since it visits all possible partitions even if some partitions in this collection are not relevant. The algorithm complexity of the exhaustive search for a model of size D is $O(Dn^2)$ by using a dynamic programmation allowed here since the contrast is additive. However, when the size of sample n is too large, like in a genomic framework, this algorithm can not be runned and the CART algorithm is preferable.

3 Glossary

3.1 Bernstein concentration inequality

We recall here the classical Bernstein concentration inequality.

Theorem 3.1. *Let $X_1, ..., X_n$ be independent real valued random variables. Assume that there exist some positive numbers v and c such that for all integers $k \geq 2$*

$$\sum_{i=1}^{n} \mathbb{E}\left[|X_i|^k\right] \leq \frac{k!}{2}\ v\ c^{k-2}. \tag{3.19}$$

Let $S_n = \sum_{i=1}^{n} (X_i - E(X_i))$, then for any positive x we have

$$P\left(S_n \geq \sqrt{2vx} + cx\right) \leq \exp(-x). \tag{3.20}$$

Also we have for any positive x

$$P(S_n \geq x) \leq \exp\left(-\frac{x^2}{2(v + cx)}\right). \tag{3.21}$$

Note that if the variables X_t are bounded, $|X_t| \leq b'$, then assumption (3.19) is satisfied with

$$v = \sum_{t=1}^{n} \mathbb{E}\left[X_t^2\right] \quad \text{and} \quad c = b'/3.$$

3.2 Bounds for Kullback-Leibler information

The following lemma is useful since the loss function in the log-likelihood estimation context is up to a constant n the Kullback-Leibler information.

Lemma 3.2. *For all positive densities p and q with respect to μ, one has*

$$\frac{1}{2}\int f^2\ \left(1 \wedge e^f\right)\ p\ d\mu \leq K(p,q) \leq \frac{1}{2}\int f^2\ \left(1 \vee e^f\right)\ p\ d\mu$$

if one notes $f = \log\left(\frac{q}{p}\right)$.

See the proof of this lemma in [5].

4 Proofs of the evaluations of R_1 and R_2 on one model

This section gives the proof of Proposition 2.1 which evaluate respectively the risks R_1 and R_2. For $c \in \{1,2\}$ we have the following Pythagore type identity with $\widehat{s}_m$ defined in (2.5)

$$l_c(s,\widehat{s}_m) = l_c(s,\bar{s}_m) + l_c(\bar{s}_m,\widehat{s}_m), \tag{4.22}$$

where $l_c(s,\bar{s}_m)$ represents some approximation error and $l_c(\bar{s}_m,\hat{s}_m)$ represents some estimation error within the model $\mathcal{S}_m$ and where $\bar{s}_m$ is given by (2.10). For all $c \in \{1,2\}$, according to the decomposition of $l_c(s,\widehat{s}_m)$ in (4.22), the proof is reduced to the evaluation of $\mathbb{E}_s\left[l_c(\bar{s}_m,\widehat{s}_m)\right]$.

4.1 Evaluation of R_1

According to the definitions of $\widehat{s}_m$ in (2.5) and $\bar{s}_m$ in (2.10), we have

$$\|\bar{s}_m - \widehat{s}_m\|^2 = \sum_{J\in m}\sum_{i=1}^{r}\frac{1}{|J|}\left[\sum_{t\in J} s\left(t,i\right) - N_J\left(i\right)\right]^2.$$

And

$$\begin{aligned}\mathbb{E}_s\left[\sum_{t\in J} s\left(t,i\right) - N_J\left(i\right)\right]^2 &= Var_s\left[\sum_{t\in J}\mathbb{1}_{\{Y_t=i\}}\right]\\ &= \sum_{t\in J} s\left(t,i\right)\left[1-s\left(t,i\right)\right].\end{aligned}$$

Since $\rho \le s\left(t,i\right) \le 1\ \forall t,i$ by definition and assumption (2.11), we have the following bounds

$$\rho(r-1)|m| \le \mathbb{E}_s\left[\|\bar{s}_m - \widehat{s}_m\|^2\right] \le (r-1)|m|$$

which achieves the proof of the first part of Proposition 2.1.

4.2 Evaluation of R_2

Let us see the following definitions.

Definition 4.1. *Given some partition $m \in \mathcal{M}_n$ and $i \in \{1, ..., r\}$, one defines the statistics $\chi^2_m(i)$ by*

$$\chi^2_m(i) = \sum_{J \in m} \frac{\overline{N_J(i)}^2}{\sum_{t \in J} s(t,i)},$$

where

$$\overline{N_J(i)} = \sum_{t \in J} \left[\mathbb{1}_{\{Y_t = i\}} - s(t,i) \right] \tag{4.23}$$

for all $i \in \{1, ..., r\}$ and $J \in m$.

Definition 4.2. *For $(u, v) \in S^2$ where S is given by (2.4), one defines*

$$V_s^2(u,v) = \sum_{t=1}^{n} \sum_{i=1}^{r} s(t,i) \log^2 \left[\frac{u(t,i)}{v(t,i)} \right].$$

The following proposition gives a control of the loss function l_2 in term of V_s^2.

Proposition 4.3. *Let $m \in \mathcal{M}_n$, we recall that*

$$l_2(\bar{s}_m, \hat{s}_m) = \sum_{J \in m} \sum_{i=1}^{r} |J| \, \bar{s}_m(J,i) \log \left[\frac{\bar{s}_m(J,i)}{\hat{s}_m(J,i)} \right],$$

and

$$V_s^2(\bar{s}_m, \hat{s}_m) = \sum_{J \in m} \sum_{i=1}^{r} |J| \, \bar{s}_m(J,i) \log^2 \left[\frac{\bar{s}_m(J,i)}{\hat{s}_m(J,i)} \right], \tag{4.24}$$

where $\bar{s}_m$ is defined by (2.10) and $\hat{s}_m$ by (2.5). Moreover, for every $\varepsilon > 0$, we set

$$\begin{aligned} \Omega_m(\varepsilon) &= \bigcap_{t=1}^{n} \bigcap_{i=1}^{r} \left\{ \left| \frac{\widehat{s}_m(t,i)}{\bar{s}_m(t,i)} - 1 \right| \leq \varepsilon \right\} \qquad (4.25) \\ &= \bigcap_{J \in m} \bigcap_{i=1}^{r} \left\{ \left| N_J(i) - \sum_{t \in J} s(t,i) \right| \leq \varepsilon \sum_{t \in J} s(t,i) \right\}, \end{aligned}$$

and we have on $\Omega_m(\varepsilon)$ the following inequality

$$\frac{1-\varepsilon}{2} \, V_s^2(\bar{s}_m, \hat{s}_m) \leq l_2(\bar{s}_m, \hat{s}_m) \leq \frac{1+\varepsilon}{2} \, V_s^2(\bar{s}_m, \hat{s}_m).$$

Proof. We use the result about densities given in Lemma 3.2. The loss function is written in term of Kullback-Leibler information as follows

$$l_2\left(\bar{s}_m, \hat{s}_m\right) = \sum_{J \in m} \mid J \mid K\left(\bar{s}_m\left(J,.\right), \hat{s}_m\left(J,.\right)\right),$$

where $K\left(s\left(J,.\right), u\left(J,.\right)\right)$ denotes the Kullback-Leibler information between the two probabilities $s\left(J,.\right)$ and $u\left(J,.\right)$ on the set $\{1,2,...,r\}$. Applying Lemma 3.2, we have on $\Omega_m\left(\varepsilon\right)$

$$\begin{aligned} l_2\left(\bar{s}_m, \hat{s}_m\right) &\geq \frac{1}{2} \sum_{J \in m} \sum_{i=1}^{r} |J| \, \bar{s}_m\left(J,i\right) \left[1 \wedge \frac{\hat{s}_m\left(J,i\right)}{\bar{s}_m\left(J,i\right)}\right] \log^2 \left[\frac{\bar{s}_m\left(J,i\right)}{\hat{s}_m\left(J,i\right)}\right] \\ &\geq \frac{1-\varepsilon}{2} V_s^2\left(\bar{s}_m, \hat{s}_m\right), \end{aligned}$$

and

$$\begin{aligned} l_2\left(\bar{s}_m, \hat{s}_m\right) &\leq \frac{1}{2} \sum_{J \in m} \sum_{i=1}^{r} |J| \, \bar{s}_m\left(J,i\right) \left[1 \vee \frac{\hat{s}_m\left(J,i\right)}{\bar{s}_m\left(J,i\right)}\right] \log^2 \left[\frac{\bar{s}_m\left(J,i\right)}{\hat{s}_m\left(J,i\right)}\right] \\ &\leq \frac{1+\varepsilon}{2} \, V_s^2\left(\bar{s}_m, \hat{s}_m\right). \end{aligned}$$

□

We now begin the proof of Proposition 2.1. The term $V_s^2\left(\bar{s}_m, \hat{s}_m\right)$ defined in (4.24) can be written as follows

$$V_s^2\left(\bar{s}_m, \hat{s}_m\right) = \sum_{J \in m} \sum_{i=1}^{r} |J| \frac{\left[\hat{s}_m(J,i) - \bar{s}_m(J,i)\right]^2}{\bar{s}_m(J,i)} \left[\frac{\log\left[\hat{s}_m(J,i)/\bar{s}_m(J,i)\right]}{\hat{s}_m(J,i)/\bar{s}_m(J,i) - 1}\right]^2.$$

Since

$$\frac{1}{1 \vee x} \leq \frac{\log x}{x-1} \leq \frac{1}{1 \wedge x} \text{ for all } x > 0,$$

we get on the set $\Omega_m\left(\varepsilon\right)$ that

$$\frac{1}{\left(1+\varepsilon\right)^2} \sum_{i=1}^{r} \chi_m^2\left(i\right) \leq V_s^2\left(\bar{s}_m, \hat{s}_m\right) \leq \frac{1}{\left(1-\varepsilon\right)^2} \sum_{i=1}^{r} \chi_m^2\left(i\right), \tag{4.26}$$

where $\chi_m^2\left(i\right)$ is given in Definition 4.1 for $i \in \{1,...,r\}$. We derive from Proposition 4.3 and the inequality (4.26) the following control of the term $l_2\left(\bar{s}_m, \hat{s}_m\right)$ on the set $\Omega_m\left(\varepsilon\right)$

$$\frac{1-\varepsilon}{2\left(1+\varepsilon\right)^2} \sum_{i=1}^{r} \chi_m^2\left(i\right) \leq l_2\left(\bar{s}_m, \hat{s}_m\right) \leq \frac{1+\varepsilon}{2\left(1-\varepsilon\right)^2} \sum_{i=1}^{r} \chi_m^2\left(i\right). \tag{4.27}$$

Furthermore, the expectation of $\chi^2_m(i)$ is equal to

$$\begin{aligned}\mathbb{E}_s\left[\chi^2_m(i)\right] &= \sum_{J\in m}\frac{Var_s\left[\sum_{t\in J}\mathbb{1}_{\{Y_t=i\}}\right]}{\sum_{t\in J}s(t,i)} \\ &= \sum_{J\in m}\frac{\sum_{t\in J}s(t,i)\left[1-s(t,i)\right]}{\sum_{t\in J}s(t,i)}.\end{aligned}$$

Since $s(t,i)\leq 1\ \forall t,i$ and according to the hypothesis (2.11), we have

$$\frac{\sum_{t\in J}\rho\left[1-s(t,i)\right]}{|J|}\leq\frac{\sum_{t\in J}s(t,i)\left[1-s(t,i)\right]}{\sum_{t\in J}s(t,i)}\leq 1,$$

and since $r-1\geq r/2$ for $r\geq 2$, we have the following bounds

$$\rho\frac{r}{2}|m|\leq\sum_{i=1}^{r}\mathbb{E}_s\left[\chi^2_m(i)\right]\leq r|m|. \tag{4.28}$$

Moreover, we have

$$\left|\mathbb{E}_s\left[l_2(\bar{s}_m,\hat{s}_m)\mathbb{1}_{\Omega^c_m(\varepsilon)}\right]\right|\leq n\log\left(\frac{1}{\rho}\right)P\left(\Omega^c_m(\varepsilon)\right). \tag{4.29}$$

On the one hand, according to inequalities (4.27), (4.28) and (4.29), we get the following upper bound of the risk R_2

$$\begin{aligned}\mathbb{E}_s\left[l_2(s,\hat{s}_m)\right] &= l_2(s,\bar{s}_m)+\mathbb{E}_s\left[l_2(\bar{s}_m,\hat{s}_m)\mathbb{1}_{\Omega_m(\varepsilon)}\right]+\mathbb{E}_s\left[l_2(\bar{s}_m,\hat{s}_m)\mathbb{1}_{\Omega^c_m(\varepsilon)}\right] \\ &\leq l_2(s,\bar{s}_m)+\frac{1+\varepsilon}{2(1-\varepsilon)^2}r|m|+n\log\left(\frac{1}{\rho}\right)P\left(\Omega^c_m(\varepsilon)\right).\end{aligned}$$

On the other hand we get by the same way the following lower bound of the risk R_2

$$\begin{aligned}\mathbb{E}_s\left[l_2(s,\hat{s}_m)\right] \geq\ & l_2(s,\bar{s}_m)+\frac{1-\varepsilon}{4(1+\varepsilon)^2}\left[\rho r|m|-\frac{16rn}{\rho}P\left(\Omega^c_m(\varepsilon)\right)\right] \\ & -n\log\left(\frac{1}{\rho}\right)P\left(\Omega^c_m(\varepsilon)\right).\end{aligned}$$

We conclude by the control of $P\left(\Omega_m(\varepsilon)^c\right)$ given by the following Lemma.

Lemma 4.4. *For every $\varepsilon>0$, let $\Omega_m(\varepsilon)$ be given by (4.25). For every $\varepsilon>0$ and $a>0$, there exists some constant $C(\Gamma,\rho,a,r,\varepsilon)$ such that*

$$P\left(\Omega_m(\varepsilon)^c\right)\leq\frac{C(\Gamma,\rho,a,r,\varepsilon)}{n^a}.$$

Proof. We remark that

$$P\left(\Omega_m\left(\varepsilon\right)^c\right) \leq \sum_{i=1}^{r} \sum_{J \in m} P\left(|N_J\left(i\right) - \sum_{t \in J} s\left(t,i\right)| > \varepsilon \sum_{t \in J} s\left(t,i\right)\right).$$

By applying Theorem 3.1, we obtain

$$\begin{aligned} & P\left(|N_J\left(i\right) - \sum_{t \in J} s\left(t,i\right)| > \varepsilon \sum_{t \in J} s\left(t,i\right)\right) \\ \leq \; & 2\exp\left(-\frac{\varepsilon^2\left(\sum_{t \in J} s\left(t,i\right)\right)^2}{2\left(\sum_{t \in J} s\left(t,i\right) + \frac{\varepsilon}{3}\sum_{t \in J} s\left(t,i\right)\right)}\right) \\ \leq \; & 2\exp\left(-\frac{\varepsilon^2}{2\left(1+\frac{\varepsilon}{3}\right)} \sum_{t \in J} s\left(t,i\right)\right). \end{aligned}$$

Set $\varepsilon' = \frac{\varepsilon^2}{2\left(1+\frac{\varepsilon}{3}\right)}$, we have $\mid J \mid \geq \Gamma\left[\log\left(n\right)\right]^2$ and $s \geq \rho$, then

$$P\left(\Omega_m\left(\varepsilon\right)^c\right) \leq 2r|m| \exp\left(-\varepsilon' \rho \Gamma\left[\log\left(n\right)\right]^2\right).$$

The result follows now from the fact that $|m| \leq n$. □

5 Proofs of Theorems about risk bounds for the penalized estimators

By the definitions of $\left(\hat{m}\right)_c$ given by (2.12) and the minimum contrast estimator $\hat{s}_m$ given by (2.5), we obtain for $c \in \{1,2\}$ that

$$\gamma_c\left(\tilde{s}_c\right) + pen_c\left(\left(\hat{m}\right)_c\right) \leq \gamma_c\left(\hat{s}_m\right) + pen_c\left(m\right) \leq \gamma_c\left(\bar{s}_m\right) + pen_c\left(m\right) \;\; \forall m \in \mathcal{M}_n.$$

Therefore,

$$l_c\left(s,\tilde{s}_c\right) \leq l_c\left(s,\bar{s}_m\right) + \overline{\gamma_c}\left(\bar{s}_m\right) - \overline{\gamma_c}\left(\tilde{s}_c\right) - pen_c((\hat{m})_c) + pen_c(m), \quad (5.30)$$

where $\overline{\gamma_c}$ is defined for $u \in \mathcal{S}$ by

$$\overline{\gamma_c}(u) = \gamma_c(u) - \mathbb{E}_s\left[\gamma_c(u)\right].$$

In the general procedure, the next principal step consists in controlling $\overline{\gamma_c}\left(\bar{s}_m\right) - \overline{\gamma_c}\left(\hat{s}_{m'}\right)$ uniformly over $m' \in \mathcal{M}_n$.

5.1 Risk bound for the least-squares penalized estimator

For the sake of simplicity in this whole section, we note γ for γ_1, l for l_1 , pen for pen_1, $\hat{m}$ for $(\hat{m})_1$ and $\tilde{s}$ for $\tilde{s}_1$. Let $m, m' \in \mathcal{M}_n$ and $(u, v) \in \mathcal{S}_m \times \mathcal{S}_{m'}$. We have

$$\begin{aligned}\bar{\gamma}(u) - \bar{\gamma}(v) &= 2\sum_{t=1}^{n}\sum_{i=1}^{r}\left[\mathbb{1}_{\{Y_t=i\}} - s(t,i)\right]\left[v(t,i) - u(t,i)\right] \\ &= 2\sum_{J\in m;K\in m'}\sum_{i=1}^{r}\overline{N_{J\cap K}(i)}\left[v(K,i) - u(J,i)\right],\end{aligned}$$

where $\overline{N_J(i)}$ is given by (4.23). By Cauchy-Schwarz inequality, we obtain for $m' \in \mathcal{M}_n$

$$\bar{\gamma}(\bar{s}_m) - \bar{\gamma}(\hat{s}_{m'}) \leq 2\sqrt{\sum_{J\in m;K\in m'}\sum_{i=1}^{r}\frac{\overline{N_{J\cap K}(i)}^2}{|J\cap K|}} \times ||\bar{s}_m - \hat{s}_{m'}||.$$

Definition 5.1. *Given some partition $m \in \mathcal{M}_n$ and some $i \in \{1, ..., r\}$, one defines the statistics $\widetilde{\chi}^2_m(i)$ by*

$$\widetilde{\chi}^2_m(i) = \sum_{L\in m}\frac{\overline{N_L(i)}^2}{|\ L\ |}$$

where $\overline{N_L(i)}$ is given by (4.23).

Then,

$$\bar{\gamma}(\bar{s}_m) - \bar{\gamma}(\hat{s}_{m'}) \leq 2\sqrt{\sum_{i=1}^{r}\widetilde{\chi}^2_{m\cap m'}(i)} \times ||\bar{s}_m - \hat{s}_{m'}||, \tag{5.31}$$

where $m \cap m'$ is a partition constructed on the grid $\{1, ..., n\}$ and defined by $m \cap m' = (I \cap J)_{I\in m, J\in m'}$. To bound the contrast, we need some technical results about the concentration of the random variable $\sum_{i=1}^{r}\widetilde{\chi}^2_m(i)$ for a given partition m. The following proposition provides such a result.

Proposition 5.2. *Let m be a partition of $\{1, ..., n\}$. For any positive x, we have*

$$P\left(\sum_{i=1}^{r}\widetilde{\chi}^2_m(i) \geq \frac{|m|}{4}r + 2\sqrt{r^2|m|\ x} + rx\right) \leq r\exp(-x).$$

Proof. We would like to apply Bernstein Theorem 3.1, so we need to control the expectation of the variables $\widetilde{\chi}^2_m(i)$, and their moments of order $p \geq 2$, for $i \in \{1, 2, ..., r\}$. Fix $i \in \{1, 2, ..., r\}$, we have

$$\begin{aligned}
\mathbb{E}_s\left[\widetilde{\chi}^2_m(i)\right] &= \sum_{J \in m} \frac{1}{\mid J \mid} Var_s\left[N_J(i)\right] \\
&= \sum_{J \in m} \frac{1}{\mid J \mid} \sum_{t \in J} s(t,i)\left[1 - s(t,i)\right] \\
&\leq \frac{|m|}{4}.
\end{aligned}$$

For every $J \in m$ and every integer $p \geq 2$, we have

$$\begin{aligned}
\mathbb{E}_s\left[\left| \frac{\overline{N_J(i)}^2}{\mid J \mid} \right|^p\right] &= \frac{1}{\mid J \mid^p} \mathbb{E}_s\left[\mid \overline{N_J(i)}^2 \mid^p\right] \\
&= \frac{1}{\mid J \mid^p} \int_0^{+\infty} (2p)\, x^{2p-1} P\left(\mid \overline{N_J(i)} \mid \geq x\right) dx.
\end{aligned}$$

Hoeffding's inequality provides a subgaussian type inequality for $\overline{N_J(i)}$: for any positive x we have

$$P\left(\mid \overline{N_J(i)} \mid \geq x\right) \leq 2 \exp\left(-\frac{2x^2}{\mid J \mid}\right).$$

Therefore

$$\begin{aligned}
\mathbb{E}_s\left[\left| \frac{\overline{N_J(i)}^2}{\mid J \mid} \right|^p\right] &\leq \frac{4p}{\mid J \mid^p} \int_0^{+\infty} x^{2p-1} \exp\left(-\frac{2x^2}{\mid J \mid}\right) dx \\
&\leq \frac{4p}{4^p} \int_0^{+\infty} u^{2p-1} \exp\left(-\frac{u^2}{2}\right) du.
\end{aligned}$$

Moreover

$$\begin{aligned}
\int_0^{+\infty} u^{2p-1} \exp\left(-\frac{u^2}{2}\right) du &= 2^{p-1} \int_0^{+\infty} u^{p-1} \exp(-u)\, du \\
&= 2^{p-1} p!,
\end{aligned}$$

then

$$\begin{aligned}
\mathbb{E}_s\left[\left| \frac{\overline{N_J(i)}^2}{\mid J \mid} \right|^p\right] &\leq \frac{4p}{4^p} 2^{p-1} p! \\
&\leq p!.
\end{aligned}$$

It follows that for every $p \geq 2$

$$\sum_{J \in m} \mathbb{E}_s \left[\left| \frac{\overline{N_J(i)}^2}{\mid J \mid} \right|^p \right] \leq \frac{p!}{2} \, 2|m| \, .$$

By applying Bernstein inequality recalled in (3.20), for any positive x one has

$$P\left(\sum_{J \in m} \frac{\overline{N_J(i)}^2}{\mid J \mid} \geq \frac{|m|}{4} + 2\sqrt{\mid m \mid \, x} + x \right) \leq \exp(-x) \, ,$$

and by summing in $i \in \{1, 2, ..., r\}$ we obtain the result of the proposition. □

We prove now the first part of Theorem 2.3. Fix $m \in \mathcal{M}_n$, let $\xi > 0$ and for every $m' \in \mathcal{M}_n$ consider $x_{m'} = L_{m'}|m'| + \xi$. We introduce the event

$$\Omega_\xi = \left\{ \forall m' \in \mathcal{M}_n, \sum_{i=1}^{r} \widetilde{\chi}^2_{m \cap m'}(i) \leq \frac{r}{4}\left(|m| + |m'|\right) + 2\sqrt{r\left(|m| + |m'|\right) x_{m'}} + r x_{m'} \right\}.$$

We sum up the resulting inequality of Proposition 5.2 over all $m' \in \mathcal{M}_n$ to provide

$$P\left(\Omega_\xi^c\right) \leq r \sum_{m' \in \mathcal{M}_n} \exp(-x_{m'}) = r\Sigma \exp(-\xi) \, ,$$

since $|m \cap m'| \leq |m| + |m'|$. Now, we will use the following inequality which holds for any positive number θ and any numbers a and b :

$$2 \, a \, b \leq \theta \, a^2 + \theta^{-1} \, b^2. \tag{5.32}$$

Fix $\eta \in (0,1)$. Using twice the preceeding inequality, we derive from (5.30) and (5.31) since $||u - \bar{s}_m|| \leq ||s - u|| + ||s - \bar{s}_m||$, that on Ω_ξ

$$\begin{aligned} & \eta^2 ||s - \tilde{s}||^2 \qquad (5.33) \\ \leq \; & \left(1 - \eta + \eta^{-1}\right) ||s - \bar{s}_m||^2 + pen(m) - pen(\hat{m}) \\ & + \left(\frac{r}{1-\eta}\right) |\widehat{m}| \left[\frac{1}{4} + \frac{\eta}{4} + 2\sqrt{L_{\widehat{m}}} + \; 2L_{\widehat{m}}\right] \\ & + \left(\frac{r}{1-\eta}\right) |m| \left[\frac{1}{4} + 1 + \frac{\eta}{4}\right] + \left(\frac{r}{1-\eta}\right) \left[1 + \frac{8}{\eta}\right] \xi. \end{aligned}$$

Using again repeatedly inequality (5.32) one derives that

$$2\sqrt{\left(|m| + |m'|\right) x_{m'}} \leq 2\left|m'\right| \sqrt{L_{m'}} + \left|m'\right| L_{m'} + |m'| \frac{\eta}{4} + |m| \left[1 + \frac{\eta}{4}\right] + \frac{8}{\eta} \xi.$$

Combining inequality (5.30) with (5.33) and the previous one applied to $m' = \hat{m}$, we deduce on Ω_ξ that

$$\begin{aligned}\eta^2||s-\tilde{s}||^2 &\leq \left(1-\eta+\eta^{-1}\right)||s-\bar{s}_m||^2 + pen(m) - pen(\hat{m}) \\ &+ \left(\frac{r}{1-\eta}\right)|\widehat{m}|\left[\frac{1}{4}+\frac{\eta}{4}+2\sqrt{L_{\widehat{m}}}+\ 2L_{\widehat{m}}\right] \\ &+ \left(\frac{r}{1-\eta}\right)|m|\left[\frac{1}{4}+1+\frac{\eta}{4}\right] + \left(\frac{r}{1-\eta}\right)\left[1+\frac{8}{\eta}\right]\xi.\end{aligned}$$

Since

$$pen\,(m) \geq K\ r\ |m|\ \left(\frac{1}{4}+2\sqrt{L_m}+\ 2L_m\right),$$

then choosing η such that $K = (1+\eta)\,/\,(1-\eta)$, i.e. $\eta = (K-1)\,/\,(K+1)$ one has the following inequality on the set Ω_ξ

$$\begin{aligned}\left(\frac{K-1}{K+1}\right)^2||s-\tilde{s}||^2 &\leq \left(\frac{K^2+4K-1}{K^2-1}\right)||s-\bar{s}_m||^2 + pen(m) \\ &+ \left(\frac{6K+4}{8}\right)r|m| + \left(\frac{K+1}{2(K-1)}\right)(9K+7)\,r\xi.\end{aligned}$$

Since

$$\frac{6K+4}{8} \leq \frac{5}{4}K,\ \left(\frac{K+1}{2(K-1)}\right)(9K+7) \leq 8K\left(\frac{K+1}{K-1}\right),$$

and

$$Kr|m| \leq 4pen(m),$$

then we get on Ω_ξ

$$\begin{aligned}||s-\tilde{s}||^2 &\leq \left(\frac{K+1}{K-1}\right)^2\left(\frac{K^2+4K-1}{K^2-1}\right)||s-\bar{s}_m||^2 \\ &+ \left(\frac{K+1}{K-1}\right)^2\left[6pen(m)+8K\left(\frac{K+1}{K-1}\right)r\xi\right] \\ &\leq \frac{8\,(K+1)^3}{(K-1)^3}\left[||s-\bar{s}_m||^2 + pen\,(m) + Kr\xi\right].\end{aligned}$$

Integrating this inequality with respect to ξ and minimizing the bound over $m \in \mathcal{M}_n$ we achieve the proof paraphrasing the method of Birgé and Massart in [1].

5.2 Risk bound for the log-likelihood penalized estimator

As in the preceeding Section, we note γ for γ_2, l for l_2 , pen for pen_2, $\hat{m}$ for $(\hat{m})_2$ and $\tilde{s}$ for $\tilde{s}_2$. Contrary to the least-square procedure estimation, following the work of Castellan [5], we control the term $\bar{\gamma}(\bar{s}_m) - \bar{\gamma}(\hat{s}_{m'})$ by decomposing it as follows

$$\begin{aligned} \bar{\gamma}(\bar{s}_m) - \bar{\gamma}(\hat{s}_{m'}) &= \bar{\gamma}(\bar{s}_{m'}) - \bar{\gamma}(\hat{s}_{m'}) + \bar{\gamma}(s) - \bar{\gamma}(\bar{s}_{m'}) \\ &\quad + \bar{\gamma}(\bar{s}_m) - \bar{\gamma}(s). \end{aligned} \tag{5.34}$$

First we give in the following Section some results essential for the control of the risk of the penalized estimator $\tilde{s}$ in Theorem 2.3.

5.2.1 Exponential bounds

The first result gives a concentration inequality of the random variable $\sum_{i=1}^{r} \chi_m^2(i)$ where $\chi_m^2(i)$ is given in Definition 4.1. The idea is to control $\sum_{i=1}^{r} \chi_m^2(i)$ for a fixed partition $m \in \mathcal{M}_n$ by Bernstein Theorem on a event where $\widehat{s}_m(t,i)$ is close to $\overline{s}_m(t,i)$ for every t and i.

Proposition 5.3. *Fix a partition m defined in (1.1). For any positive real numbers x and ε*

$$\begin{aligned} & P\left(\sum_{i=1}^{r} \chi_m^2(i)\, \mathbb{1}_{\Omega_m(\varepsilon)} \geq r|m| + 8r\left(1+\frac{\varepsilon}{3}\right)\sqrt{x|m|} + 4r\left(1+\frac{\varepsilon}{3}\right)x\right) \\ \leq\ & r\exp(-x) \end{aligned}$$

where $\Omega_m(\varepsilon)$ is defined in (4.25).

Proof. We would like to apply the Bernstein concentration inequality recalled in Section 3.1 to

$$\chi_m^2(i) = \sum_{J \in m} Z(J,i),$$

for a given $i \in \{1, \ldots, r\}$, where

$$Z(J,i) = \frac{\overline{N_J(i)}^2}{\sum_{t \in J} s(t,i)},$$

and $\overline{N_J(i)} = \sum_{t\in J}\left[\mathbb{1}_{\{Y_t=i\}} - s(t,i)\right]$. We set $\overline{\mathbb{1}_{\{Y_t=i\}}} = \mathbb{1}_{\{Y_t=i\}} - s(t,i)$, and according to inequality (3.19), we have to control the moments of $Z(J,i)$. For every $p \geq 2$ we have,

$$\begin{aligned}
&\mathbb{E}_s\left[Z(J,i)^p\,\mathbb{1}_{\Omega_m(\varepsilon)}\right] \\
&= \frac{1}{\left[\sum_{t\in J} s(t,i)\right]^p}\int_0^{+\infty}(2p)\,x^{2p-1}P\left[\left\{|\sum_{t\in J}\overline{\mathbb{1}_{\{Y_t=i\}}}|\geq x\right\}\cap\Omega_m(\varepsilon)\right]dx \\
&\leq \frac{1}{\left[\sum_{t\in J} s(t,i)\right]^p}\int_0^{\varepsilon\sum_{t\in J}s(t,i)}(2p)\,x^{2p-1}P\left[|\sum_{t\in J}\overline{\mathbb{1}_{\{Y_t=i\}}}|\geq x\right]dx.
\end{aligned}$$

By applying inequality (3.21), one obtains for $0 < x \leq \varepsilon\sum_{t\in J} s(t,i)$

$$\begin{aligned}
P\left[|\sum_{t\in J}\overline{\mathbb{1}_{\{Y_t=i\}}}|\geq x\right] &\leq 2\exp\left(-\frac{x^2}{2\left(\frac{x}{3}+\sum_{t\in J}s(t,i)\right)}\right) \\
&\leq 2\exp\left(-\frac{x^2}{2\left(1+\frac{\varepsilon}{3}\right)\sum_{t\in J}s(t,i)}\right).
\end{aligned}$$

Therefore

$$\begin{aligned}
&\mathbb{E}_s\left[Z(J,i)^p\,\mathbb{1}_{\Omega_m(\varepsilon)}\right] \\
&\leq \frac{1}{\left[\sum_{t\in J} s(t,i)\right]^p}\int_0^{\varepsilon\sum_{t\in J}s(t,i)} 4px^{2p-1}\exp\left(-\frac{x^2}{2\left(1+\frac{\varepsilon}{3}\right)\sum_{t\in J}s(t,i)}\right)dx \\
&\leq 4p\left(1+\frac{\varepsilon}{3}\right)^p\int_0^{+\infty}u^{2p-1}\exp\left(-\frac{u^2}{2}\right)du \\
&\leq 4p\left(1+\frac{\varepsilon}{3}\right)^p\int_0^{+\infty}(2t)^{p-1}\exp(-t)\,dt \\
&\leq 2^{p+1}p\left(1+\frac{\varepsilon}{3}\right)^p p!
\end{aligned}$$

and

$$\sum_{J\in m}\mathbb{E}_s\left[Z(J,i)^p\,\mathbb{1}_{\Omega_m(\varepsilon)}\right] \leq 2^{p+1}p\left(1+\frac{\varepsilon}{3}\right)^p p!\times|m|.$$

Since $p \leq 2^{p-1}$,

$$\sum_{J\in m}\mathbb{E}_s\left[Z(J,i)^p\,\mathbb{1}_{\Omega_m(\varepsilon)}\right] \leq \frac{p!}{2}\times\left[2^5\left(1+\frac{\varepsilon}{3}\right)^2|m|\right]\times\left(4\left(1+\frac{\varepsilon}{3}\right)\right)^{p-2}.$$

Moreover $\mathbb{E}\left[\chi^2_m(i)\right] \leq |m|$ (see the proof of the inequality (4.28)) so we get for every positive x that

$$\begin{aligned} & P\left(\sum_{J\in m} Z(J,i)\, 1\!\!1_{\Omega_m(\varepsilon)} \geq |m| + 8\left(1+\frac{\varepsilon}{3}\right)\sqrt{x|m|} + 4\left(1+\frac{\varepsilon}{3}\right)x\right) \\ \leq \quad & \exp(-x). \end{aligned}$$

We conclude the proof of Proposition 5.3 by summing in $i \in \{1,...,r\}$. □

The second result given in this section is an exponential bound of $\bar{\gamma}(s) - \bar{\gamma}(u)$ for every $u \in \mathcal{S}$. Here we control the contrast as a function of Hellinger distance instead of Kullback-Leiber information which is possibly infinite.

Proposition 5.4. *For every $u \in \mathcal{S}$ and any positive x,*

$$P\left(\bar{\gamma}(s) - \bar{\gamma}(u) \geq nK(s,u) - 2nh^2(s,u) + 2xr\right) \leq r\exp(-x),$$

where

$$\bar{\gamma}(s) - \bar{\gamma}(u) = \sum_{i=1}^{r}\sum_{t=1}^{n} \overline{1\!\!1_{\{Y_t=i\}}} \log\left[\frac{u(t,i)}{s(t,i)}\right],$$

and h^2 is the Hellinger distance defined by (2.14).

Proof. Let a positive number a. By Markov inequality, we have for $i \in \{1,...,r\}$

$$\begin{aligned} & P\left(\sum_{t=1}^{n} \overline{1\!\!1_{\{Y_t=i\}}} \log\left[\frac{u(t,i)}{s(t,i)}\right] \geq a\right) \\ \leq \quad & \exp\left[-\frac{a}{2} + \log\left(\mathbb{E}_s\left[\exp\left[\frac{1}{2}\sum_{t=1}^{n}\left(1\!\!1_{\{Y_t=i\}} - s(t,i)\right)\log\left[\frac{u(t,i)}{s(t,i)}\right]\right]\right]\right)\right] \\ \leq \quad & \exp\left[-\frac{a}{2} + \frac{1}{2}\sum_{t=1}^{n} s(t,i)\log\left[\frac{s(t,i)}{u(t,i)}\right]\right] \\ & \prod_{t=1}^{n}\mathbb{E}_s\left[\exp\left[\frac{1}{2}1\!\!1_{\{Y_t=i\}}\log\left[\frac{u(t,i)}{s(t,i)}\right]\right]\right]. \end{aligned}$$

Let $t \in \{1,...n\}$ and $i \in \{1,...,r\}$. We have

$$\begin{aligned} \mathbb{E}_s\left[\exp\left[\frac{1}{2}1\!\!1_{\{Y_t=i\}}\log\left[\frac{u(t,i)}{s(t,i)}\right]\right]\right] &= \sqrt{\frac{u(t,i)}{s(t,i)}}s(t,i) + (1-s(t,i)) \\ &= 1 - s(t,i)\left[1 - \sqrt{\frac{u(t,i)}{s(t,i)}}\right]. \end{aligned}$$

Then

$$\begin{aligned}
& \log\left(\mathbb{E}_s\left[\exp\left[\frac{1}{2}\mathbb{1}_{\{Y_t=i\}}\log\left[\frac{u(t,i)}{s(t,i)}\right]\right]\right]\right) \\
\leq\ & -s(t,i)\left[1-\sqrt{\frac{u(t,i)}{s(t,i)}}\right] \\
\leq\ & \frac{1}{2}\left[u(t,i)-s(t,i)-\left(\sqrt{s(t,i)}-\sqrt{u(t,i)}\right)^2\right].
\end{aligned}$$

If we choose for x positive real number

$$\begin{aligned}
a &= 2x+\sum_{t=1}^{n} s(t,i)\log\left[\frac{s(t,i)}{u(t,i)}\right] \\
&\quad +\sum_{t=1}^{n}\left[u(t,i)-s(t,i)-\left(\sqrt{s(t,i)}-\sqrt{u(t,i)}\right)^2\right],
\end{aligned}$$

then

$$P\left(\sum_{t=1}^{n}\overline{\mathbb{1}_{\{Y_t=i\}}}\log\left[\frac{u(t,i)}{s(t,i)}\right]\geq a\right)\leq\exp(-x),$$

and we conclude the proof by summing for $i\in\{1,...,r\}$. □

5.2.2 Proof of Theorem for the likelihood procedure estimation

According to the decomposition (5.34), we control first the two terms : $\bar{\gamma}(\bar{s}_m)-\bar{\gamma}(s)$ and $\bar{\gamma}(\bar{s}_{m'})-\bar{\gamma}(\hat{s}_{m'})$.

- Control of the term $\bar{\gamma}(\bar{s}_m)-\bar{\gamma}(s)$. We have to control

  $$\mathbb{E}_s\left[(\bar{\gamma}(\bar{s}_m)-\bar{\gamma}(s))\,\mathbb{1}_{\Omega_m(\varepsilon)}\right]=-\mathbb{E}_s\left[(\bar{\gamma}(\bar{s}_m)-\bar{\gamma}(s))\,\mathbb{1}_{\Omega_m(\varepsilon)^c}\right],$$

  and we bound this expectation by

  $$\begin{aligned}
  \left|\mathbb{E}_s\left[(\bar{\gamma}(\bar{s}_m)-\bar{\gamma}(s))\,\mathbb{1}_{\Omega_m(\varepsilon)}\right]\right| &\leq \left|\mathbb{E}\left[(\bar{\gamma}(\bar{s}_m)-\bar{\gamma}(s))\,\mathbb{1}_{\Omega_m^c(\varepsilon)}\right]\right| \\
  &\leq nr\log\left(\frac{1}{\rho}\right)P(\Omega_m(\varepsilon)^c).
  \end{aligned}$$

  We use the result of Lemma 4.4 and we obtain a constant $C_1(\Gamma,\rho,r,\varepsilon)$ such that

  $$\mathbb{E}_s\left[(\bar{\gamma}(\bar{s}_m)-\bar{\gamma}(s))\,\mathbb{1}_{\Omega_m(\varepsilon)}\right]\leq C_1(\Gamma,\rho,r,\varepsilon). \tag{5.35}$$

- Control of the term $\bar{\gamma}\left(\bar{s}_{m'}\right) - \bar{\gamma}\left(\hat{s}_{m'}\right)$. We write for $m' \in \mathcal{M}_n$

$$\bar{\gamma}\left(\bar{s}_{m'}\right) - \bar{\gamma}\left(\hat{s}_{m'}\right) = \sum_{t=1}^{n}\sum_{i=1}^{r} \overline{\mathbb{1}_{\{Y_t=i\}}} \log\left[\frac{\hat{s}_{m'}(t,i)}{\bar{s}_{m'}(t,i)}\right].$$

By Cauchy-Schwarz inequality,

$$\begin{aligned} & \bar{\gamma}\left(\bar{s}_{m'}\right) - \bar{\gamma}\left(\hat{s}_{m'}\right) \qquad (5.36) \\ \leq & \sqrt{\sum_{i=1}^{r}\sum_{K\in m'} \frac{\left(\sum_{t\in K} \overline{\mathbb{1}_{\{Y_t=i\}}}\right)^2}{\sum_{t\in K} s(t,i)}} \times V_s\left(\hat{s}_{m'}, \bar{s}_{m'}\right) \\ \leq & \sqrt{\sum_{i=1}^{r} \chi^2_{m'}(i)} \times V_s\left(\hat{s}_{m'}, \bar{s}_{m'}\right). \end{aligned}$$

where V_s is defined by (4.24) and $\chi^2_{m'}(i)$ given in Definition 4.1. First we have to bound

$$\sqrt{\sum_{i=1}^{r} \chi^2_{m'}(i)}$$

by the concentration arguments proposed in Proposition 5.3 and then we have to use the relation between $V_s^2\left(\hat{s}_{m'}, \bar{s}_{m'}\right)$ and $l\left(\hat{s}_{m'}, \bar{s}_{m'}\right)$ exposed in Proposition 4.3. On $\Omega_{m_f}(\varepsilon)$ with m_f satisfying (2.17) we have for all m' constructed on the partition m_f

$$\left|\frac{\widehat{s}_{m'}(t,i)}{\bar{s}_{m'}(t,i)} - 1\right| \leq \varepsilon$$

for all $t \in \{1,...,n\}$, and $i \in \{1,...,r\}$. Therefore according to the inequality (5.36) and Proposition 4.3, we have on $\Omega_{m_f}(\varepsilon)$ that for all $m' \in \mathcal{M}_n$

$$\bar{\gamma}\left(\bar{s}_{m'}\right) - \bar{\gamma}\left(\hat{s}_{m'}\right) \leq \sqrt{\sum_{i=1}^{r} \chi^2_{m'}(i)}\ \sqrt{\frac{2}{1-\varepsilon} l\left(\bar{s}_{m'}, \hat{s}_{m'}\right)}.$$

Using inequality (5.32) with $\theta = (1+\varepsilon)/(1-\varepsilon)$, we obtain on $\Omega_{m_f}(\varepsilon)$ that for all $m' \in \mathcal{M}_n$

$$\bar{\gamma}\left(\bar{s}_{m'}\right) - \bar{\gamma}\left(\hat{s}_{m'}\right) \leq \frac{1}{2}\left(\frac{1+\varepsilon}{1-\varepsilon}\right)\sum_{i=1}^{r} \chi^2_{m'}(i) + \frac{1}{1+\varepsilon} l\left(\bar{s}_{m'}, \hat{s}_{m'}\right).$$

We now introduce the following event defined for positive ξ

$$\Omega_1(\xi) = \left\{ \begin{array}{c} \sum_{i=1}^{r} \chi^2_{m'}(i)\, 1\!\!1_{\Omega_{m_f}(\varepsilon)} \\ \leq r|m'| + 8r\left(1+\frac{\varepsilon}{3}\right)\sqrt{x_{m'}|m'|} + 4r\left(1+\frac{\varepsilon}{3}\right)x_{m'}, \forall m' \in \mathcal{M}_n \end{array} \right\},$$

where $x_{m'} = L_{m'}|m'| + \xi$ for all $m' \in \mathcal{M}_n$, with $L_{m'}$ chosen such that the condition (2.16) is satisfied and $\chi^2_{m'}(i)$ is given in Definition 4.1 for all $i \in \{1, \ldots, r\}$ and for all $m' \in \mathcal{M}_n$. We get on $\Omega_1(\xi)$ that for all $m' \in \mathcal{M}_n$

$$\begin{aligned} & \left[\bar{\gamma}(\bar{s}_{m'}) - \bar{\gamma}(\hat{s}_{m'})\right] 1\!\!1_{\Omega_{m_f}(\varepsilon)} \\ \leq\ & \frac{1}{2}\left(\frac{1+\varepsilon}{1-\varepsilon}\right)\left[r|m'| + 8r\left(1+\frac{\varepsilon}{3}\right)\sqrt{(|m'|L_{m'}+\xi)\,|m'|}\right] \\ & + 2r\left(\frac{1+\varepsilon}{1-\varepsilon}\right)\left(1+\frac{\varepsilon}{3}\right)(|m'|L_{m'}+\xi) \\ +\ & \frac{1}{1+\varepsilon} l(\bar{s}_{m'}, \hat{s}_{m'})\, 1\!\!1_{\Omega_{m_f}(\varepsilon)}. \end{aligned}$$

By the C_r inequality with $r = 2$ ($|x+y|^{1/2} \leq |x|^{1/2} + |y|^{1/2}$) and by using inequality (5.32) with $\theta = \frac{\varepsilon}{4}$, we obtain on $\Omega_1(\xi)$ that for all $m' \in \mathcal{M}_n$

$$\begin{aligned} & \left[\bar{\gamma}(\bar{s}_{m'}) - \bar{\gamma}(\hat{s}_{m'})\right] 1\!\!1_{\Omega_{m_f}(\varepsilon)} \qquad (5.37) \\ \leq\ & \frac{1}{2}\left(\frac{1+\varepsilon}{1-\varepsilon}\right) r|m'|\left[1 + \left(1+\frac{\varepsilon}{3}\right)\left(\varepsilon + 8\sqrt{L_{m'}} + 4L_{m'}\right)\right] \\ & + 2r\xi\left(\frac{1+\varepsilon}{1-\varepsilon}\right)\left(1+\frac{\varepsilon}{3}\right)\left(1+\frac{4}{\varepsilon}\right) + \frac{1}{1+\varepsilon} l(\bar{s}_{m'}, \hat{s}_{m'})\, 1\!\!1_{\Omega_{m_f}(\varepsilon)}. \end{aligned}$$

Let us finally control $l(s, \tilde{s})$ on the set $\Omega_{m_f}(\varepsilon)$ in defining the set

$$\Omega_2(\xi) = \left\{\bar{\gamma}(s) - \bar{\gamma}(\bar{s}_{m'}) \leq nK(s, \bar{s}_{m'}) - 2nh(s, \bar{s}_{m'}) + 2x_{m'}r, \forall m' \in \mathcal{M}_n\right\}$$

Combining inequality (5.34) with (5.37) applied for $m' = \hat{m}$, we obtain on $\Omega_1(\xi) \cap \Omega_2(\xi)$

$$\begin{aligned} l(s, \tilde{s})\, 1\!\!1_{\Omega_{m_f}(\varepsilon)} \leq\ & l(s, \bar{s}_m) + pen(m) - pen(\hat{m}) + R\, 1\!\!1_{\Omega_{m_f}(\varepsilon)} \\ & + r|\hat{m}|C(\varepsilon)\left[\frac{1}{2} + 4\sqrt{L_{\hat{m}}} + 4L_{\hat{m}}\right] \\ & + 2r\xi\left[1 + \left(\frac{1+\varepsilon}{1-\varepsilon}\right)\left(1+\frac{\varepsilon}{3}\right)\left(1+\frac{4}{\varepsilon}\right)\right] \\ & + \left[nK(s, \bar{s}_{\hat{m}}) - 2nh^2(s, \bar{s}_{\hat{m}}) + \frac{1}{1+\varepsilon} l(\bar{s}_{\hat{m}}, \tilde{s})\right] 1\!\!1_{\Omega_{m_f}(\varepsilon)}, \end{aligned}$$

where

$$R = \bar{\gamma}\left(\bar{s}_m\right) - \bar{\gamma}\left(s\right), \tag{5.38}$$

and

$$\begin{aligned} C(\varepsilon) &= \left\{ \left[\left(\frac{1+\varepsilon}{1-\varepsilon} \right) \left(1+\varepsilon\left(1+\varepsilon\right)\right) \right] \vee \left[\frac{(1+\varepsilon)^2}{1-\varepsilon} \right] \right\}, \\ &= \left(\frac{1+\varepsilon}{1-\varepsilon} \right) \left(1+\varepsilon\left(1+\varepsilon\right)\right) \\ &\leq \left(\frac{1+\varepsilon}{1-\varepsilon} \right)^3. \end{aligned}$$

Then we use respectively the following different relations:

- the relation between the loss function and the Kullback-Leibler information: $l\left(s,\tilde{s}\right) = nK\left(s,\tilde{s}\right)$,
- the following Pythagore's type identity

$$K(s,\tilde{s}) = K(s,\bar{s}_{\hat{m}}) + K(\bar{s}_{\hat{m}},\tilde{s}),$$

- the relation between the Kullback-Leibler information and the Hellinger distance given by (2.15),
- and the following inequality

$$h^2(s,\tilde{s}) \leq 2\left[h^2(s,\bar{s}_{\hat{m}}) + h^2(\bar{s}_{\hat{m}},\tilde{s})\right].$$

We obtain according to the form of the penalty given by (2.18) and since $\varepsilon/\left(1+\varepsilon\right) < 1$ on $\Omega_1\left(\xi\right) \cap \Omega_2\left(\xi\right)$

$$\begin{aligned} \frac{\varepsilon}{1+\varepsilon} n h^2(s,\tilde{s}) \mathbb{1}_{\Omega_{m_f}(\varepsilon)} &\leq nK\left(s,\bar{s}_m\right) + pen\left(m\right) + R\ \mathbb{1}_{\Omega_{m_f}(\varepsilon)} \\ &+ r|\hat{m}| \left[\frac{1}{2} + 4\sqrt{L_{\hat{m}}} + 4L_{\hat{m}} \right] \left[\left(\frac{1+\varepsilon}{1-\varepsilon} \right)^3 - \lambda \right] \\ &+ 2r\xi \left[1 + \left(\frac{1+\varepsilon}{1-\varepsilon} \right) \left(1 + \frac{\varepsilon}{3} \right) \left(1 + \frac{4}{\varepsilon} \right) \right]. \end{aligned}$$

We take ε such that $\lambda = \left(\left(1+\varepsilon\right)/\left(1-\varepsilon\right)\right)^3$, *i.e.* $\varepsilon = \left(\lambda^{1/3}-1\right)/\left(\lambda^{1/3}+1\right)$, and bound

$$2\left[1 + \left(\frac{1+\varepsilon}{1-\varepsilon} \right) \left(1 + \frac{\varepsilon}{3} \right) \left(1 + \frac{4}{\varepsilon} \right) \right] \leq 16 \left(\frac{1+\varepsilon}{\varepsilon\left(1-\varepsilon\right)} \right)^3.$$

We have on $\Omega_1(\xi) \cap \Omega_2(\xi)$

$$\begin{aligned} nh^2(s,\tilde{s})\mathbb{1}_{\Omega_{m_f}(\varepsilon)} &\leq \frac{2\lambda^{1/3}}{\lambda^{1/3}+1}\left[nK(s,\bar{s}_m)+pen(m)\right] \\ &+ \frac{2\lambda^{1/3}}{\lambda^{1/3}+1}\left[R\ \mathbb{1}_{\Omega_{m_f}(\varepsilon)} + 16r\xi K\left(\frac{K^{1/3}-1}{K^{1/3}+1}\right)\right]. \end{aligned}$$

Since we derive from Propositions 5.3 and 5.4 that

$$\begin{aligned} P\left(\Omega_1(\xi)^c\right) &\leq r\sum_{m'\in\mathcal{M}_n}\exp\left(-x_{m'}\right) \\ \text{and } P\left(\Omega_2(\xi)^c\right) &\leq r\sum_{m'\in\mathcal{M}_n}\exp\left(-x_{m'}\right), \end{aligned}$$

we deduce that

$$P\left(\Omega_1(\xi)^c \cup \Omega_2(\xi)^c\right) \leq 2r\sum_{m'\in\mathcal{M}_n}\exp\left(-x_{m'}\right) \leq 2r\Sigma\exp\left(-\xi\right),$$

thus $P\left(\Omega_1(\xi)\cap\Omega_2(\xi)\right) \geq 1-2r\Sigma\exp(-\xi)$. Now integrating this inequality with respect to ξ and the control of $\mathbb{E}_s[R\mathbb{1}_{\Omega_{m_f}(\varepsilon)}]$ given by (5.35) where R is defined by (5.38) allow to conclude that

$$\mathbb{E}_s\left[nh^2(s,\tilde{s})\,\mathbb{1}_{\Omega_{m_f}(\varepsilon)}\right] \leq \frac{2\lambda^{1/3}}{\lambda^{1/3}-1}\left[l(s,\bar{s}_m)+pen(m)\right] + C(\Sigma,r,K,\rho,\Gamma,\varepsilon).$$

Furthermore, with the control of $P\left(\Omega_{m_f}(\varepsilon)^c\right)$ given in Lemma 4.4, there exists a constant $C_2(\Gamma,\rho,r,\varepsilon)$ such that

$$\mathbb{E}_s\left[nh^2(s,\tilde{s})\,\mathbb{1}_{\Omega_{m_f}(\varepsilon)^c}\right] \leq C(\Gamma,\rho,r,\varepsilon),$$

since $h^2(s,\tilde{s}) \leq 1$. Hence, we conclude that

$$\mathbb{E}_s\left[nh^2(s,\tilde{s})\right] \leq \frac{2\lambda^{1/3}}{\lambda^{1/3}-1}\left[l(s,\bar{s}_m)+pen(m)\right] + C(\Sigma,r,K,\rho,\Gamma,\varepsilon),$$

and minimizing the bound over $m\in\mathcal{M}_n$ we complete the proof of Theorem 2.3.

6 Conclusion

In conclusion, in theoretical point of view the proof of the risk bound in the likelihood procedure needs some more technicals points. So in the case where the problem can be put into the least-square framework, the least-square procedure is preferable.

References

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

[2] Braun R. K. Braun, J. V. and H.-G. Müller, *Multiple changepoint fitting via quasilikelihood, with application to dna sequence segmentation.*

[3] J. V. Braun and H.-G. Müller, *Statistical methods for DNA sequence segmentation*, Biometrika **13** (1998), no. 2, 301–314. MR 2001e:62020

[4] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and regression trees*, Chapman & Hall, 1984.

[5] G. Castellan, *Modified Akaike's criterion for histogram density estimation*, C. R. Acad. Sci., Paris, Sér. I, Math. 330 **8** (2000), 729–732.

[6] G. Churchill, *Stochastic Models for Heterogeneous DNA Sequences*, Bulletin of Mathematical Biology **51** (1989), no. 1, 79–94.

[7] F. Muri, *Searching gene transfers on bacillus subtilis using hidden markov models*, Compstat'98 Proceedings in Computational Statistics (1998), 98–100.

[8] D.A. Henderson R.J. Boys and D.J. Wilkinson, *Detecting homogeneous segments in DNA sequences by using hidden Markov models*, J. Roy. Statist. Soc. Ser. C **48** (1999), no. 4, 489–503. MR 2000h:92020