Electronic Journal of Statistics
Vol. 11 (2017) 800–842
ISSN: 1935-7524
DOI: 10.1214/17-EJS1246

Model selection for the segmentation of multiparameter exponential family distributions

Alice Cleynen

Institut Montpelliérain Alexander Grothendieck, France
CNRS UMR 5149, France
e-mail: alice.cleynen@umontpellier.fr

and

Emilie Lebarbier

AgroParisTech, UMR 518 MIA, F-75005 Paris, France
INRA, UMR 518 MIA, F-75005 Paris, France
e-mail: emilie.lebarbier@agroparistech.fr

Abstract: We consider the segmentation problem of univariate distributions from the exponential family with multiple parameters. In segmentation, the choice of the number of segments remains a difficult issue due to the discrete nature of the change-points. In this general exponential family distribution framework, we propose a penalized log-likelihood estimator where the penalty is inspired by papers of L. Birgé and P. Massart. The resulting estimator is proved to satisfy some oracle inequalities. We then further study the particular case of categorical variables by comparing the values of the key constants when derived from the specification of our general approach and when obtained by working directly with the characteristics of this distribution. Finally, simulation studies are conducted to assess the performance of our criterion and to compare our approach to other existing methods, and an application on real data modeled using the categorical distribution is provided.

MSC 2010 subject classifications: Primary 62G05, 62G07; secondary 62P10.
Keywords and phrases: Exponential family, distribution estimation, change-point detection, model selection.

Received July 2016.

1. Introduction

Segmentation, the partition of a profile into segments of homogeneous distribution, is a very useful tool for the modelization and interpretation of complex data. For instance in biology, the output of an RNA-Seq experiment can be modeled through the segmentation of negative binomial random variables which are homogeneous along coding or non-coding regions [15]; the composition of nucleotides along the genome can be modeled through a categorical segmentation framework to identify transcription sites ([11] and references therein); the output of array-CGH experiments can be modeled trough the segmentation of Gaussian random variables along regions of identical copy-number ([28] and references therein), etc. In almost all segmentation frameworks for independent variables, if the number of segments is fixed *a priori*, identifying the optimal segmentation with respect to either the log-likelihood or the least-squares criteria can be very easily and efficiently performed through the use of algorithms such as dynamic programming [5] and its pruned versions [27, 35, 31].

A crucial step, which is the main difficulty in segmentation approaches, is therefore the choice of the number of segments K. To this end, a huge effort has been made in the last two decades to derive estimators of K and their properties. If, in this context, almost all methods for choosing the number of segments can be seen as penalized-likelihood approaches (Akaike Information Criterion, [1], Bayes Information Criterion [38], Integrated Completed Likelihood [36], etc), we and other authors (see for instance [29, 8, 39]) have previously emphasized how crucial the choice of the penalty function is in contexts such as segmentation where the size of the collection of models grows with the size of the data.

The difficulty in this choice is exacerbated by the fact that to obtain satisfying results (consistency of the estimator, oracle inequality, etc), the penalty usually depends strongly on the choice of the distribution, preventing general approaches to be applied in a straightforward manner. Therefore most of earlier works have focused on variables distributed from specified distributions, for instance Gaussian [29], or either Poisson or negative binomial distributions [17].

The aim of this paper is to provide a penalty function for the more general framework of exponential family distributions for univariate data, but with possibly many parameters. The model that will be considered throughout the paper is the following:

$$Y_t \sim \mathcal{G}(\boldsymbol{\theta}_t) = s(t),\ 1 \leq t \leq n,\ Y_t \text{ independent},$$

where $\mathcal{G}$ is a distribution from the exponential family with parameter $\boldsymbol{\theta}_t \in \mathbb{R}^d,\ d \geq 1$. In the context of segmentation models, the parameter $\boldsymbol{\theta}_t$ is supposed to be piece-wise constant along the time-line $1, \ldots, n$ with $K-1$ changes. The goal is therefore to identify a partition of $\{1, \ldots, n\}$ into K segments within which the observations can be modeled as following the same distribution while their distributions differ between segments.

The model selection approach developed in the following sections is based on the pioneer work of [6, 4] introducing non-asymptotic model selection procedures. This penalized contrast procedure consists in selecting a model amongst a collection such that its performance is as close as possible to that of the best but unreachable model in terms of risk. In this sense, this work is quite similar to our previous results obtained for Poisson and negative binomial distributions [17]. It is in fact partly motivated by the numerous similarities we had observed between those two distributions that turn out to be closely related to properties of the log-partition function of the exponential family.

In the next section, we will recall some general properties of the exponential family, introduce our collection of models and our penalized-likelihood framework. After relating our work to previous papers, we will state our main result in Section 3. In Section 4 the main statement is proved, while relinquishing all intermediary results to the Appendix. In Section 5, we show under some additional constraint on the sufficient statistics that our penalized estimator verifies both a statistical oracle inequality in terms of Kullback-Leibler risk and a trajectorial oracle inequality in terms of Kullback-Leibler divergence. In Section 6 we study the particular case of categorical variables to assess the precision lost in dealing with general exponential family instead of directly bounding the particular distribution. Finally, in Section 7, we first illustrate the performance of our approach on a simulation study based on the exponential distribution. Then compare our approach to some of the segmentation approaches available in the literature first on a smaller Poisson simulation study and then on a Student simulation study to assess the robustness of our approach to the exponential family distribution assumption. Finally, we propose an application to DNA sequence distribution on a real data-set which we model using a piece-wise constant categorical distribution.

2. Framework and related work

In our framework we will consider the minimal canonical form of the exponential family distribution which we will write as

$$s(t) = g(Y_t) \exp\left[\boldsymbol{\theta}_t.\mathbf{T}_t - A(\boldsymbol{\theta}_t)\right],$$

where the . symbol denotes the canonical scalar product of $\mathbb{R}^d$, A is the log-partition function of $\mathcal{G}$ and $\mathbf{T}_t = T(Y_t)$ ($\in \mathbb{R}^d$) is the minimal sufficient statistic associated with variable Y_t. s will denote the joint distribution of the $\{Y_t\}_{1\leq t\leq n}$, which, since the Y_ts are independent, is simply the product of the $s(t)$.

For instance, in the case of random variables distributed from the Poisson distribution, $Y_t \sim \mathcal{P}(\lambda_t)$, one has $s(t) = g(Y_t)\exp\left[\boldsymbol{\theta}_t.\mathbf{T}_t - A(\boldsymbol{\theta}_t)\right]$ with $g(x) = \frac{1}{x!}$, $\boldsymbol{\theta}_t = \log\lambda_t$, $\mathbf{T}_t = Y_t$ and $A(\boldsymbol{\theta}_t) = \exp\boldsymbol{\theta}_t$.

2.1. Some properties of the exponential family

Exponential family distributions, and in particular the log-partition function, have been well studied in the past years. In a pioneer work [13], Brown has described the fundamental properties of exponential family distributions, such as parametrization using sufficient statistics, differentiability of the log-partition function and its relation to moments, etc. More recently, [37] has demonstrated the strong links between graphical models and exponential family, [26] has studied the sub-exponential growth of the cumulants of an exponential family distribution and studied convergence rates of regularization algorithm under sparsity assumptions while [30] has studied consistency properties of the lasso procedure to estimate parameters of an exponential family distribution under some convexity and sparsity assumptions. These latter results lie on some concentration inequalities of a quantity that can be interpreted in our context as the centered sufficient statistics, and on some (restricted) strong convexity assumption on the Fisher information matrix.

In our framework, we will assume that our distributions are non-degenerate, *i.e.* that there exists a subset of $\mathbb{R}^d$ such that the log-partition function A is $\mathcal{C}^\infty$. Moreover, we will restrict this set to a convex compact subset $\boldsymbol{\nu}$ on which we can ensure that the first moments of the sufficient statistics are bounded. Among the key features of minimal exponential family is the relationship between the derivatives of A and the moments of the sufficient statistics.

- The first moment is given by $\mathbf{E}[\mathbf{T}_t] = \nabla A(\boldsymbol{\theta}_t)$ and will be further denoted $\mathbf{E}_t$. Moreover, using minimal representation of the exponential family ensures that the gradient mapping $\nabla A : \boldsymbol{\nu} \to \nabla A(\boldsymbol{\nu})$ is a bijection (see for instance [37]).
- The second moment is given by $Cov[\mathbf{T}_t] = \nabla^2 A(\boldsymbol{\theta}_t)$. In the case of non-degenerate distributions, this matrix is symmetric and definite positive. This implies that for any compact set $\mathbb{K}$; there exists a non-negative constant $m_{\mathbb{K}}$ such that $\nabla^2 A$ is lower bounded by $m_{\mathbb{K}}$, *i.e.* A is $m_{\mathbb{K}}$-strongly convex on $\mathbb{K}$. In particular, there exists $m_\nu > 0$ such that A is m_ν-strongly convex on $\boldsymbol{\nu}$. Finally, for notation simplifications, we will write, for $1 \leq i \leq d$, $V_t^{(i)} = Var[T_t^{(i)}]$.

The following definition introduces sub-gamma joint distributions:

Definition 2.1. *A joint distribution s is said to be sub-gamma with variance parameter $v \geq 1$ and scale parameter c, written* $\mathbf{SG}(v, c)$*, if* $\forall 1 \leq t \leq n$, $\forall 1 \leq i \leq d$

$$\log \mathbf{E}\left[e^{z(T_t^{(i)} - E_t^{(i)})}\right] \vee \log \mathbf{E}\left[e^{-z(T_t^{(i)} - E_t^{(i)})}\right] \quad \leq \quad \frac{z^2}{2}\frac{vV_t^{(i)}}{1 - cz}, \quad \forall\ 0 < |z| < \frac{1}{c}.$$

Note that this definition is slightly different from the traditional sub-gamma definition since the variance term has been decomposed into the true variance of the sufficient statistic, $V_t^{(i)}$, and an adjustment parameter v. This necessarily implies that $v \geq 1$.

Thus for an $\mathbf{SG}(v, c)$ distribution, since by assumption for all $1 \leq i \leq d$ the $T_t^{(i)}$ are independent, a direct consequence follows from Chernoff's inequality: for any J interval of $\{1, \dots, n\}$,

$$\mathbf{P}\left[\left|\sum_{t \in J}\left(T_t^{(i)} - E_t^{(i)}\right)\right| \geq \sqrt{2vxV_J^{(i)}} + cx\right] \leq 2e^{-x}$$

where $V_J^{(i)} = \sum_{t \in J} V_t^{(i)}$, and thus, with $\kappa = \max\{v, c\}$,

$$\mathbf{P}\left[\left|\sum_{t \in J}\left(T_t^{(i)} - E_t^{(i)}\right)\right| \geq x\right] \leq 2e^{-\frac{x^2}{2\kappa(V_J^{(i)} + x)}}. \tag{1}$$

Recall that the cumulants of $T_t^{(i)}$, denoted c_k, are such that

$$\log \mathbf{E}\left[e^{z(T_t^{(i)}-E_t^{(i)})}\right] = \sum_{k\geq 2} c_k \frac{z^k}{k!}.$$

Distributions from the exponential family which cumulants' growth is exponentially bounded (*i.e.* for which there exists a positive constant α such that $\forall k, c_k \leq \alpha^k$) can be shown to be $\mathbf{SG}(v,c)$ for appropriate parameters v and c depending on the cumulants.

2.2. Penalized maximum-likelihood estimator

In this change-point setting, we will want to consider partitions m of the set $\{1,\ldots,n\}$ on which our models will be piece-wise constant. More precisely, for a given partition m and with J denoting a generic segment of m, we define the collection of models associated to m as:

$$\mathcal{S}_m = \{s_m \mid \forall J \in m, \forall t \in J, s_m(t) = \mathcal{G}(\boldsymbol{\theta}_J)\}.$$

We will consider the log-likelihood empirical contrast γ_n and the associated Kullback-Leibler divergence $K(s,u) = \mathbf{E}[\gamma_n(u) - \gamma_n(s)]$ between distributions s and u so that if $s(t) = \mathcal{G}(\boldsymbol{\theta}_t)$ and $u(t) = \mathcal{G}(\boldsymbol{p}_t)$, we have

$$\begin{aligned} \gamma_n(s) &= \sum_{t=1}^{n} \left[A(\boldsymbol{\theta}_t) - \boldsymbol{\theta}_t.\mathbf{T}_t\right], \quad \text{and} \\ K(s,u) &= \sum_{t=1}^{n} \Big[\nabla A(\boldsymbol{\theta}_t).(\boldsymbol{\theta}_t - \boldsymbol{p}_t) - (A(\boldsymbol{\theta}_t) - A(\boldsymbol{p}_t))\Big]. \end{aligned}$$

The minimal contrast estimator $\hat{s}_m$ of s on the collection $\mathcal{S}_m$ is therefore $\hat{s}_m = \arg\min_{u\in\mathcal{S}_m} \gamma_n(u)$ and is given by

$$\hat{s}_m(t) \quad = \quad g(Y_t)\exp\left[[\nabla A]^{-1}(\bar{\mathbf{T}}_J).\mathbf{T}_t - A\left([\nabla A]^{-1}(\bar{\mathbf{T}}_J)\right)\right],$$

where $\mathbf{T}_J = \sum_{t\in J}\mathbf{T}_t$ is the sum of the sufficient statistics on interval J, and $\bar{\mathbf{T}}_J = \mathbf{T}_J/|J|$ its mean where $|J|$ denotes the length of the interval J, i.e. the number of points that belong to J; the bijective mapping of the gradient of A ensuring the existence and uniqueness of $[\nabla A]^{-1}(\bar{\mathbf{T}}_J)$.

Similarly, the projection $\bar{s}_m$ of s in terms of Kullback-Leibler divergence on $\mathcal{S}_m$ is $\bar{s}_m \; = \; \arg\min_{u\in\mathcal{S}_m} K(s,u)$ and is given by

$$\bar{s}_m(t) \quad = \quad g(Y_t)\exp\left[[\nabla A]^{-1}(\bar{\mathbf{E}}_J).\mathbf{T}_t - A\left[[\nabla A]^{-1}(\bar{\mathbf{E}}_J)\right]\right],$$

where $\mathbf{E}_J = \sum_{t\in J}\mathbf{E}_t$ and $\bar{\mathbf{E}}_J = \mathbf{E}_J/|J|$. Details on the computation of $\bar{s}_m$ and $\hat{s}_m$ are given in Appendix A.

Note that the true distribution s will never be assumed to belong to any collection of models $\mathcal{S}_m$.

In penalized-likelihood settings, it is classical to propose estimators that achieve some *oracle inequality.* Since minimizing the Kullback-Leibler risk or even the Kullback-Leibler divergence would require knowing the true distribution s (indeed neither the risk oracle $m_R(s) = \arg\min_m \mathbf{E}[K(s, \hat{s}_m)]$ nor the divergence oracle $m_D(s) = \arg\min_m [K(s, \hat{s}_m)]$ can be reached), the idea is to choose a penalty function $pen(m)$ such that the penalized estimator $\hat{s}_{\hat{m}}$ - where $\hat{m} = \arg\min \gamma_n(\hat{s}_m) + pen(m)$ - satisfies either

(i) a *statistical* oracle inequality:

$$\mathbf{E}[K(s, \hat{s}_{\hat{m}})] \leq C_1 \mathbf{E}[K(s, \hat{s}_{m_R(s)})] + C_2 \tag{2}$$

(ii) or a *trajectorial* oracle inequality: with high probability

$$K(s, \hat{s}_{\hat{m}}) \leq C_1 K(s, \hat{s}_{m_D(s)}) + C_2; \tag{3}$$

where in both cases, C_2 is negligible compared to either C_1 $\mathbf{E}[K(s, \hat{s}_{m_R(s)})]$ or to $C_1 K(s, \hat{s}_{m_D(s)})$. Typically, for trajectorial inequalities, one wishes to show that for any $x > 0$, with probability greater than $1 - e^{-x}$

$$K(s, \hat{s}_{\hat{m}}) \leq C_1 K(s, \hat{s}_{m_D(s)}) + C_2 x.$$

One can note that the trajectorial oracle inequality is slightly stronger than the statistical oracle inequality since a simple integration of the former leads to

$$\mathbf{E}[K(s, \hat{s}_{\hat{m}})] \leq C_1 \mathbf{E}\left[\inf_m K(s, \hat{s}_m)\right] + C_2,$$

which is very close to

$$\mathbf{E}[K(s, \hat{s}_{\hat{m}})] \leq C_1 \inf_m \mathbf{E}\left[K(s, \hat{s}_m)\right] + C_2.$$

To achieve such results, here as in previous works (see for example [32]), we will introduce an event of large probability, $\Omega_{m_f}(\varepsilon)$, where the fluctuation of each centered marginal is bounded. On this event, we will derive tight controls of the Kullback-Leibler divergence and risk of some models which will lead to the shape of our penalty function. In section 3, specific further controls are obtained to achieve an oracle inequality in terms of Hellinger risk on the whole space: the small probability of $\Omega_{m_f}(\varepsilon)^C$ compensates the coarse bounds obtained on this event. In section 5, we restrict the space to $\Omega_{m_f}(\varepsilon)$ and we present two oracle inequalities of types (2) and (3). Of note, since most of our controls result from the previous section, our constants C_1 and C_2 are unlikely to be optimal.

The choice of ε is therefore crucial in insuring that both C_1 and C_2 are as small as possible, while having the negligibility property of C_2 compared to $C_1 K(s, \hat{s}_{m(s)})$ or $C_1 \mathbf{E}[K(s, \hat{s}_{m_R(s)})]$. In practice, this choice is data-driven and is efficiently performed through the use of the slope heuristic [3]. In this paper, we therefore consider a generic but fixed ε, and aim at obtaining the shape of the penalty function.

2.3. Related work

As stated above, a vast literature has focused on deriving appropriate penalty shapes for specific distributions, such as Gaussian, Poisson or Negative Binomial distributions. Still, procedures that can adapt to a more general framework are not numerous. The most classic approach consists in forgetting entirely the underlying distribution and working in a non-parametric setting. The majority of those approaches rely on the use of Hidden Markov Models, which typically assume a predefined number of states (which can then be revisited to yield an unknown number of segments) and therefore do not always adapt to our context [24, 22]. Moreover, the use of non-parametric HMMs implies a large computational complexity which prevents their application in many settings.

In [33], the authors propose a non-parametric approach which partitions the profile sequentially through the maximization of a criterion based on a weighted L^2 norm of the characteristic functions. The change-points thus identified are then challenged through permutation tests. This method yields interesting results in practice but suffers from the computational complexity of permutation tests and from the non-optimality of the sequential partitioning.

Still in a non-parametric framework, Arlot *et. al.* propose to transform the segmentation problem of univariate, multivariate or complex data into a penalized least-squares segmentation problem on some transformed data through the choice of a kernel [2]. This transformation is then used to optimize a penalized criterion where the empirical contrast γ_n is a kernel least-squares criterion that operates in a reproducible kernel Hilbert space. The penalization, addressing the problem of the choice of the number of segments, is derived such as to obtain an oracle inequality on the resulting estimator, and the authors obtain the same penalty shape as ours. This approach is extremely powerful as it can deal with a wide range of data types without requiring any assumption on the underlying distribution, however, an appropriate kernel must be chosen.

In a parametric framework, two approaches aim at selecting the number of segments based on the ICL criterion. In a Bayesian setting, [36] computes exactly, among other criteria, the ICL for any distribution of the exponential family provided the prior on the segmentation verifies some factorability assumption. In a frequentist setting, [18] computes its conditional counterpart through the use of some forward-backward algorithm that can be performed efficiently in the exponential family distributions. While none aim at guaranteeing some specific property of the resulting estimator, these approaches tend to select the number of segments with the lowest uncertainty. However, their use is restricted by the complexity of the algorithms: $\mathcal{O}(Kn^2)$ for the former, $\mathcal{O}(K^2n)$ for the latter.

In a recent work, [21] proposed an approach dedicated to one-parameter exponential family which can be interpreted as a mixture between a likelihood-ratio testing approach and a penalized likelihood model selection approach. In an asymptotic framework (except for the Gaussian case) the authors obtain an exponential bound for the probability of underestimating K (note that this requires that the true distribution truly is piece-wise constant) and therefore propose a procedure which maximizes the probability of correctly estimating K.

3. Main result

We first introduce a minimal partition m_f and we consider a collection of partitions said to be constructed on this m_f as given in the following definition.

Definition 3.1. *Let m_f be a partition of $\{1, \dots, n\}$. Then $\mathcal{M}_n$ is a collection of partitions of $\{1, \dots, n\}$ constructed on m_f if m_f is a refinement of every m in $\mathcal{M}_n$; i.e. if any segment of any element of $\mathcal{M}_n$ is the union of (consecutive) segments of m_f.*

In the sequel, we will consider a collection of partitions $\mathcal{M}_n$ built on a minimal partition m_f that verifies $\forall J \in m_f, |J| \geq \Gamma \log(n)^2$ for Γ a positive constant. Intuitively, this assumption ensures that each segment will contain a sufficient number of points to correctly estimate its distribution. Indeed, in segmentation frameworks, we are given a unique observation (of length n) of our model with which we need to estimate all parameters. What allows such a success is that observations belonging to the same segment share the same distribution and can thus be seen as *i.i.d* random variables from a marginal distribution of s, and the estimation can be performed independently of all other segments. A sufficient number of observations per segment is thus needed to correctly estimate the corresponding parameters. From a technical point of view, this assumption is required for ensuring that the event $\Omega_{m_f}(\varepsilon)$ introduced in Section 2.2 and defined as

$$\Omega_{m_f}(\varepsilon) = \bigcap_{1 \leq i \leq d} \bigcap_{J \in m_f} \left\{ \left| T_J^{(i)} - E_J^{(i)} \right| \leq \varepsilon \; V_J^{(i)} \right\} \tag{4}$$

for $\varepsilon > 0$ is indeed of large probability.

Indeed, by using equation (1), we have

$$\mathbf{P}\Big(\Omega_{m_f}(\varepsilon)^C\Big) \leq \sum_{i=1}^{d} \sum_{J \in m_f} P\Big[\Big| T_J^{(i)} - E_J^{(i)} \Big| \geq \varepsilon \; V_J^{(i)} \Big] \leq \sum_{i=1}^{d} \sum_{J \in m_f} 2e^{-\frac{\varepsilon^2 V_J^{(i)}}{2\kappa(1+\varepsilon)}}.$$

The condition $|J| \geq \Gamma(\log(n))^2$ implies that the probability of the event $\Omega_{m_f}(\varepsilon)^C$ decreases with n faster than any power of n, and therefore for n large enough there exists a constant such that

$$\mathbf{P}\Big(\Omega_{m_f}(\varepsilon)^C\Big) \quad \leq \quad \frac{C(a, d, \varepsilon, \Gamma, \kappa, V^{min})}{n^a},$$

with $a > 1$.

The negligible probability of $\Omega_{m_f}(\varepsilon)^C$ will ensure that the constant C_2 (see inequality (2)) is small compared to $C_1 \mathbf{E}[K(s, \hat{s}_{m_R(s)})]$. It will also enforce the validity of our trajectorial and statistical Kullback-Leibler oracle inequalities (see Section 5) since they are restricted to the event $\Omega_{m_f}(\varepsilon)$.

The following theorem states the main result for distributions from the exponential family:

Theorem 3.2. *Let $s = \prod_t s(t)$ be a distribution from the exponential family such that its parameters $\boldsymbol{\theta}_t$ belong to a convex compact set $\boldsymbol{\nu}$ of $\mathbb{R}^d$ with diameter*

$\boldsymbol{\nu}$. Assume that there exist two positive constants v and c such that s is $\mathbf{SG}(v,c)$ and let κ denote $\max\{v,c\}$. Let $\mathcal{M}_n$ be a collection of partitions constructed on a partition m_f such that there exists $\Gamma > 0$ satisfying $\forall J \in m_f, |J| \geq \Gamma \log(n)^2$, and let $(L_m)_{m\in\mathcal{M}_n}$ be some family of positive weights satisfying

$$\Sigma = \sum_{m\in\mathcal{M}_n} \exp(-L_m|m|) < +\infty. \tag{5}$$

Let $\varepsilon > 0$ and let β_ε be a positive constant depending on ε, the compact $\boldsymbol{\nu}$ and the distribution $\mathcal{G}$. If for every $m \in \mathcal{M}_n$

$$pen(m) \geq \beta_\varepsilon d|m| \left(1 + 4\sqrt{L_m}\right)^2, \tag{6}$$

then

$$\begin{aligned}\mathbf{E}\left[h^2(s,\hat{s}_{\hat{m}})\right] \quad \leq \quad & C_{\beta_\varepsilon} \inf_{m\in\mathcal{M}_n} \{K(s,\bar{s}_m) + pen(m)\} \\ & + C(d,\nu,\kappa,\Gamma,\beta_\varepsilon,\Sigma,V^{max},V^{min}),\end{aligned} \tag{7}$$

where h^2 denotes the Hellinger distance, $V^{min} = \min_{1\leq i\leq d}\min_{1\leq t\leq n}\{V_t^{(i)}\}$ and $V^{max} = \max_{1\leq i\leq d}\max_{1\leq t\leq n}\{V_t^{(i)}\}$ are bounds on the variance of the sufficient statistics.

Section 4 is dedicated to the proof of this result.

Remark 3.3. *The penalty constant β_ε will be shown in the proof of Theorem 3.2 to be*

$$\beta_\varepsilon = \frac{V^{max}\kappa}{2m_\varepsilon}(1+\varepsilon)^3,$$

where m_ε is a lower bound on the eigenvalues of $\nabla^2 A$ on $K(\varepsilon)$, the compact corresponding to the pre-image of $\Omega_{m_f}(\varepsilon)$ by ∇A. Intuitively, m_ε can be understood as a bound on the smallest variance of the sufficient statistics, so that β_ε is close to a ratio of largest variance over smallest variance, which is linked both to the expression of the d sufficient statistics and to the compact set $\boldsymbol{\nu}$.

It is therefore not surprising that, when working directly with specific distributions such as the Gaussian [29] or Poisson [17] distributions, this ratio disappears. Indeed, in those cases, the variance is either fixed or related to the expectation of the statistic, and simplifications in the computations prior to controlling the fluctuations of the chi-square statistic can lead to the derivation of tighter bounds.

In order to obtain oracle inequalities in the sequel, the risk associated to the estimator $\hat{s}_m$ is needed. The following proposition gives a bound on the Kullback-Leibler risk associated to $\hat{s}_m$.

Proposition 3.4. *Under the assumptions of Theorem 3.2, for all m in $\mathcal{M}_n$, we have*

$$K(s,\bar{s}_m) + C(\varepsilon)d|m| - \frac{C(a,\Gamma,\varepsilon,d,\mathcal{G})}{n^{(a-1)/2}} \leq \mathbf{E}[K(s,\hat{s}_m)],$$

where $a > 1$ and $C(\varepsilon)$ is a constant that depends on ε and the distribution $\mathcal{G}$.

The dependency of $C(\varepsilon)$ and $C(a, \Gamma, \varepsilon, d, \mathcal{G})$ on $\mathcal{G}$ is made explicit in the proof of the proposition which is given in Appendix B.2.

Choice of the weights $(L_m)_m$ in our change-point setting. The penalty function (6) depends on the collection $\mathcal{M}_n$ through the choice of the weights $(L_m)_{m\in\mathcal{M}_n}$. These weights L_m are usually chosen to depend on m only through the number of segments $|m|$, i.e. $L_m = L_{|m|}$ (see for example [32]). Moreover, in our change-point setting, since in practice there is no reason that the partitions are regulars, we explore the collection of partitions in an exhaustive way. More precisely, we consider the collection $\mathcal{M}_n$ of all partitions constructed on the regular partition m_f with N segments such that $N = \lfloor n/(\Gamma \log n^2)\rfloor$, $\lfloor x \rfloor$ denoting the integer part of x. In this case, the number of partitions having K segments is bounded by $\binom{N}{K}$, resulting in $L_m = 1.1 + \log\left(\frac{N}{|m|}\right)$ (see [29] for a justification of this choice). This leads to a penalty function of the form:

$$pen(m) = \beta_\varepsilon d|m| \left(1 + 4\sqrt{1.1 + \log\left(\frac{N}{|m|}\right)}\right)^2, \tag{8}$$

and the inequality (7) of Theorem 3.2 therefore becomes:

$$\begin{aligned}\mathbf{E}\left[h^2(s, \hat{s}_{\hat{m}})\right] \leq C_{\beta_\varepsilon} \inf_{m\in\mathcal{M}_n} &\left\{K(s, \bar{s}_m) + d\beta_\varepsilon|m| \left(1 + 4\sqrt{1.1 + \log\left(\frac{N}{|m|}\right)}\right)^2\right\} \\ &+ C(d, \nu, \kappa, \Gamma, \beta_\varepsilon, \Sigma, V^{max}, V^{min}).\end{aligned} \tag{9}$$

Plugging the risk of $\hat{s}_m$ (see Proposition 3.4) into inequality (9) leads to the oracle inequality given in Corollary 3.5.

Corollary 3.5. *Let $s = \prod_t s(t)$ be the joint distribution of n independent random variables distributed from $\mathcal{G}$ in the exponential family and such that their parameters $\boldsymbol{\theta}_t$ belong to a convex compact $\boldsymbol{\nu}$ of $\mathbb{R}^d$. Let $\mathcal{M}_n$ be the collection of all possible partitions constructed on the regular partition m_f with N segments such that $N = \lfloor n/(\Gamma \log n^2)\rfloor$, and assume $\exists$ v and c such that s is $\mathbf{SG}(v, c)$ and let κ stand for the maximum of $\{v, c\}$.*

There exists some constant C such that

$$\begin{aligned}\mathbf{E}\left[h^2(s, \hat{s}_{\hat{m}})\right] \quad \leq \quad & C\log(N) \inf_{m\in\mathcal{M}_n} \left\{\mathbf{E}[K(s, \hat{s}_m)]\right\} \\ & + C(d, \nu, \kappa, \Gamma, \beta_\varepsilon, \Sigma, V^{max}, V^{min}).\end{aligned}$$

Remark 3.6. *The oracle inequality obtained in this framework compares the Hellinger risk of the penalized estimator to the optimal Kullback-Leibler risk. This is classic in density estimation and segmentation frameworks (see for instance [32, 14, 17]) as unless one is inclined to specify additional constraints on the distribution, the Kullback-Leibler divergence is possibly infinite - and hence difficult to control. In Section 5, adding a constraint on the sufficient statistics* *(which is equivalent to restraining the whole space to some event which we show to have large probability), we propose an oracle inequality in terms of Kullback-Leibler risk of the penalized estimator.*

Remark 3.7. *The oracle inequality states that the performance of the penalized estimator is close to that of the best estimator up to a $\log N$ factor. This logarithm term is unavoidable and even necessary, as explained for example in [32] (page 238).*

Remark 3.8. *In practice when exploring the collection $\mathcal{M}_n$, we do not restrain to partitions with a minimal segment length. The search is performed in an exhaustive manner in the sense that all the possible partitions of $\{1, \dots, n\}$ are considered. The procedure is therefore applied (and implemented) using the penalty (8) with $N = n$.*

4. Proof of Theorem 3.2

As is classically done in model selection frameworks [7], starting from equation

$$\gamma_n(\hat{s}_{\hat{m}}) + pen(\hat{m}) \ \leq \ \gamma_n(\hat{s}_m) + pen(m) \ \leq \ \gamma_n(\bar{s}_m) + pen(m),$$

and introducing the centered loss $\overline{\gamma_n}(u) = \gamma_n(u) - \mathbf{E}[\gamma_n(u)]$, we write

$$K(s, \hat{s}_{\hat{m}}) \quad \leq \quad K(s, \bar{s}_m) + \overline{\gamma_n}(\bar{s}_m) - \overline{\gamma_n}(\hat{s}_{\hat{m}}) - pen(\hat{m}) + pen(m). \tag{10}$$

As in [14, 17], we subsequently decompose $\overline{\gamma_n}(\bar{s}_m) - \overline{\gamma_n}(\hat{s}_{m'})$ for any $m' \in \mathcal{M}_n$ in

$$\begin{aligned} &\overline{\gamma_n}(\bar{s}_m) - \overline{\gamma_n}(\hat{s}_{m'}) \\ &= (\overline{\gamma_n}(\bar{s}_{m'}) - \overline{\gamma_n}(\hat{s}_{m'})) + (\overline{\gamma_n}(s) - \overline{\gamma_n}(\bar{s}_{m'})) + (\overline{\gamma_n}(\bar{s}_m) - \overline{\gamma_n}(s)), \end{aligned} \tag{11}$$

and control each term separately, with the main term now defined on the same partition. To this effect, we define the compact subset $K(\varepsilon)$ of $\mathbb{R}^d$ as the pre-image by ∇A of the domains induced by Ω_{m_f}, that is

$$K(\varepsilon) = \left\{ \mathbf{z} \in \mathbb{R}^d \,\middle|\, \nabla A(\mathbf{z}) \in \bigcup_{m \in \mathcal{M}} \bigcup_{J \in m} \mathcal{B}(\bar{E}_J, \varepsilon \min_i \bar{V}_J^{(i)}) \right\},$$

where $\mathcal{B}(\mathbf{x}, r)$ denotes the closed ball centered in $\mathbf{x}$ with radius r of $\mathbb{R}^d$. Since we consider the union of a finite number of compacts, homeomorphic properties of ∇A ensure that $K(\varepsilon)$ is a compact set of $\mathbb{R}^d$. Since $\nabla^2 A$ is definite positive, there exists $m_\varepsilon > 0$ such that A is m_ε-strongly convex on the compact set $K(\varepsilon)$. In the following m_ε is chosen as a lower bound on the smallest eigenvalue of $\nabla^2 A$ on $K(\varepsilon)$. Moreover, we denote M_ε an upper bound of the eigenvalues of $\nabla^2 A$ on $K(\varepsilon)$.

Subsections 4.1, 4.2 and 4.3 give the control of the three terms of the decomposition (11), for which proof we refer the reader to Appendix B. The first two terms have to be controlled uniformly with respect to $m' \in \mathcal{M}_n$. The first one is the most delicate to handle since it requires the control of an empirical process which is not bounded. From this control will appear the shape of the penalty function. For the second term, as explained in [32], we have no guarantee that the ratios $s/\bar{s}_{m'}$ remain bounded for all m' and we will consequently introduce the Hellinger distance.

4.1. Control of term $\overline{\gamma_n}(\bar{s}_{m'}) - \overline{\gamma_n}(\hat{s}_{m'})$

Controlling this term is typically done through the study of a chi-square statistic:

$$\chi^2_m \;=\; \sum_{i=1}^{d} \chi^2(m,i) \;=\; \sum_{i=1}^{d} \sum_{J\in m} \frac{\left(T_J^{(i)} - E_J^{(i)}\right)^2}{V_J^{(i)}}. \tag{12}$$

The following proposition gives an exponential concentration bound for χ^2_m on the restricted event $\Omega_{m_f}(\varepsilon)$. The proof is given in the appendix B.1.1.

Proposition 4.1. *Let $Y_1, \dots, Y_n$ be independent random variables with joint distribution s (from the exponential family) and verifying $\mathbf{SG}(v,c)$, with $\kappa = \max\{v,c\}$. Let m be a partition of $\mathcal{M}_n$ with $|m|$ segments and χ^2_m be the statistic given by (12). For any positive x, we have*

$$\mathbf{P}\left[\chi^2_m \mathbf{1}_{\Omega_{m_f}(\varepsilon)} \geq d\ \kappa\left(|m| + 8(1+\varepsilon)\sqrt{x|m|} + 4(1+\varepsilon)x\right)\right] \quad\leq\quad d\ e^{-x}.$$

Let us now consider a positive constant ξ and introduce

$$\Omega_1(\xi) = \bigcap_{m'\in\mathcal{M}_n} \{\chi^2_{m'} \mathbf{1}_{\Omega_{m_f}(\varepsilon)} \leq d\kappa[|m'| + 8(1+\varepsilon)\sqrt{(L_{m'}|m'|+\xi)|m'|} \\ + 4(1+\varepsilon)(L_{m'}|m'|+\xi)]\}.$$

Then we can control $\overline{\gamma_n}(\bar{s}_{m'}) - \overline{\gamma_n}(\hat{s}_{m'})$ with the following proposition,

Proposition 4.2.

$$\begin{aligned}
&(\overline{\gamma_n}(\bar{s}_{m'}) - \overline{\gamma_n}(\hat{s}_{m'}))\,\mathbf{1}_{\Omega_{m_f}(\varepsilon)\cap\Omega_1(\xi)} \leq \frac{1}{1+\varepsilon} K(\bar{s}_{m'}, \hat{s}_{m'}) \\
&\quad + \frac{dV^{max}\kappa(1+\varepsilon)}{2m_\varepsilon} \\
&\quad\quad \times \left[|m'| + 8(1+\varepsilon)\sqrt{(L_{m'}|m'|+\xi)|m'|} + 4(1+\varepsilon)(L_{m'}|m'|+\xi)\right].
\end{aligned}$$

The proof is given in the appendix B.3.

4.2. Control of term $\overline{\gamma_n}(\bar{s}_m) - \overline{\gamma_n}(s)$

The expectation of the second term can be bounded using the following proposition proved in appendix B.4:

Proposition 4.3. *Let s be a distribution from the exponential family verifying* $\mathbf{SG}(v,c)$ *for some positive constants v and c, let $\kappa = \max\{v,c\}$, m a partition of $\mathcal{M}_n$ and $\bar{s}_m$ the projection of s on $\mathcal{S}_m$ and $\Omega_{m_f}(\varepsilon)$ defined by equation (4). Then*

$$\left|\mathbf{E}\left[(\overline{\gamma_n}(\bar{s}_m) - \overline{\gamma_n}(s))\mathbf{1}_{\Omega_{m_f}(\varepsilon)}\right]\right| \quad \leq \quad \frac{C(a,d,\varepsilon,\Gamma,\kappa,\nu,V^{min},V^{max})}{n^{(a-1)/2}},$$

where ν is the diameter of the compact $\boldsymbol{\nu}$*.*

4.3. Control of term $\overline{\gamma_n}(s) - \overline{\gamma_n}(\bar{s}_{m'})$

We control the last term using proposition 4.4 for which a proof can be found in [17].

Proposition 4.4. *Let s and u be two joint distributions on the same space (not necessarily from the exponential family). Then for any positive x,*

$$\mathbf{P}\left[\overline{\gamma_n}(s) - \overline{\gamma_n}(u) \geq K(s,u) - 2h^2(s,u) + 2dx\right] \quad \leq \quad e^{-dx} \leq de^{-x}.$$

4.4. Proof of the theorem

We define

$$\Omega_2(\xi) = \bigcap_{m'\in\mathcal{M}_n}\left\{\overline{\gamma_n}(s) - \overline{\gamma_n}(\bar{s}_{m'}) \leq K(s,\bar{s}_{m'}) - 2h^2(s,\bar{s}_{m'}) + 2d(L_{m'}|m'| + \xi)\right\}, \tag{13}$$

and introduce the following event: $\Omega(\varepsilon,\xi) = \Omega_{m_f}(\varepsilon) \cap \Omega_1(\xi) \cap \Omega_2(\xi)$. Recall that we have, from equation (11),

$$\begin{aligned}&(\overline{\gamma_n}(\bar{s}_m) - \overline{\gamma_n}(\hat{s}_{\hat{m}}))\mathbf{1}_{\Omega(\varepsilon,\xi)}\\ &\qquad = (\overline{\gamma_n}(s) - \overline{\gamma_n}(\bar{s}_{\hat{m}}))\mathbf{1}_{\Omega(\varepsilon,\xi)} + (\overline{\gamma_n}(\bar{s}_m) - \overline{\gamma_n}(s))\mathbf{1}_{\Omega(\varepsilon,\xi)}\\ &\qquad\quad + (\overline{\gamma_n}(\bar{s}_{\hat{m}}) - \overline{\gamma_n}(\hat{s}_{\hat{m}}))\mathbf{1}_{\Omega(\varepsilon,\xi)}.\end{aligned}$$

Then since $\Omega(\varepsilon,\xi)$ encompasses all events required from the previous propositions, we can bound the first and last of those terms using propositions 4.2 and 4.4. Further denoting $R = \overline{\gamma_n}(\bar{s}_m) - \overline{\gamma_n}(s)$, we obtain

$$\begin{aligned}&(\overline{\gamma_n}(\bar{s}_m) - \overline{\gamma_n}(\hat{s}_{\hat{m}}))\mathbf{1}_{\Omega(\varepsilon,\xi)}\\ &\quad \leq \left[K(s,\bar{s}_{\hat{m}}) - 2h^2(s,\bar{s}_{\hat{m}})\right]\mathbf{1}_{\Omega(\varepsilon,\xi)} + R\mathbf{1}_{\Omega(\varepsilon,\xi)} + \frac{1}{1+\varepsilon}K(\bar{s}_{\hat{m}},\hat{s}_{\hat{m}})\mathbf{1}_{\Omega(\varepsilon,\xi)}\\ &\qquad + C(\varepsilon)d\left[|\hat{m}| + 8(1+\varepsilon)\sqrt{(L_{\hat{m}}|\hat{m}| + \xi)|\hat{m}|} + 4(1+\varepsilon)(L_{\hat{m}}|\hat{m}| + \xi)\right]\\ &\qquad + 2dL_{\hat{m}}|\hat{m}| + 2d\xi.\end{aligned}$$

where $C(\varepsilon) = \dfrac{V^{max}\kappa(1+\varepsilon)}{2m_\varepsilon}$. We plug the previous equation into equation (10) and obtain:

$$\frac{\varepsilon}{1+\varepsilon}h^2(s,\hat{s}_{\hat{m}})\mathbf{1}_{\Omega(\varepsilon,\xi)} \quad \leq \quad K(s,\bar{s}_m)\mathbf{1}_{\Omega(\varepsilon,\xi)} + R\mathbf{1}_{\Omega(\varepsilon,\xi)} - pen(\hat{m}) + pen(m)$$

$$+d|\hat{m}|C_2(\varepsilon)\left(1+4\sqrt{L_{\hat{m}}}\right)^2$$
$$+2\xi d\left[1+(1+\varepsilon)C(\varepsilon)\left(\frac{8}{\varepsilon}+2\right)\right],$$

where $C_2(\varepsilon)=\dfrac{V^{max}\kappa}{2m_\varepsilon}(1+\varepsilon)^3$.

Then since by assumption, $pen(\hat{m})\geq\beta_\varepsilon d|\hat{m}|\left(1+4\sqrt{L_{\hat{m}}}\right)^2$, choosing $\beta_\varepsilon=C_2(\varepsilon)$ yields:

$$h^2(s,\hat{s}_{\hat{m}})\mathbf{1}_{\Omega(\varepsilon,\xi)}\leq C_{\beta_\varepsilon}\left[K(s,\bar{s}_m)\mathbf{1}_{\Omega(\varepsilon,\xi)}+R\mathbf{1}_{\Omega(\varepsilon,\xi)}+pen(m)\right]+d\xi C(\beta_\varepsilon),$$

with $C_{\beta_\varepsilon}=\dfrac{1+C_2^{-1}(\beta_\varepsilon)}{C_2^{-1}(\beta_\varepsilon)}$.

From propositions 4.1 and 4.4 (with $x=L_{m'}|m'|+\xi$ and $u=\bar{s}_{m'}$) come both

$$\mathbf{P}\left(\Omega_1(\xi)^C\right)\leq d\sum_{m'\in\mathcal{M}_n}e^{-(L_{m'}|m'|+\xi)}\quad\text{and}\quad\mathbf{P}\left(\Omega_2(\xi)^C\right)\leq d\sum_{m'\in\mathcal{M}_n}e^{-(L_{m'}|m'|+\xi)},$$

so that using hypothesis (5),

$$\mathbf{P}\left(\Omega_1(\xi)^C\cup\Omega_2(\xi)^C\right)\leq 2d\sum_{m'\in\mathcal{M}_n}e^{-(L_{m'}|m'|+\xi)}\leq 2d\Sigma e^{-\xi},$$

and thus $\mathbf{P}\left(\Omega_1(\xi)\cap\Omega_2(\xi)\right)\geq 1-2d\Sigma e^{-\xi}$. Integrating over ξ and using proposition 4.3 leads to

$$\mathbf{E}\left[h^2(s,\hat{s}_{\hat{m}})\mathbf{1}_{\Omega_{m_f}(\varepsilon)}\right]$$
$$\leq C_{\beta_\varepsilon}\left[K(s,\bar{s}_m)+\frac{C(a,d,\varepsilon,\Gamma,\kappa,\nu,V^{min},V^{max})}{n^{(a-1)/2}}+pen(m)\right]+2d\Sigma C(\beta_\varepsilon).$$

And since $\mathbf{E}\left[h^2(s,\hat{s}_{\hat{m}})\mathbf{1}_{\Omega_{m_f}(\varepsilon)^C}\right]\leq\dfrac{C(a,d,\beta_\varepsilon,\Gamma,\kappa,V^{min})}{n^a}$, we have

$$\mathbf{E}\left[h^2(s,\hat{s}_{\hat{m}})\right]\leq C_{\beta_\varepsilon}\left[K(s,\bar{s}_m)+pen(m)\right]+C'(d,\nu,\kappa,\Gamma,\beta_\varepsilon,\Sigma,V^{max},V^{min}).$$

The proof is concluded by minimizing over $m\in\mathcal{M}_n$.

5. Kullback-Leibler results on $\Omega_{m_f}(\varepsilon)$

The oracle inequality in terms of risk obtained in Section 3 (see Corollary 3.5) is expressed in terms of two different losses (Hellinger and Kullback-Leibler). As can be seen in the proof of the corresponding theorem, once the considered risk has been controlled on $\Omega_{m_f}(\varepsilon)$, it has to be bounded by some constant on $\Omega_{m_f}(\varepsilon)^C$ in order to conclude to the oracle inequality. In the above, this is easily done as the Hellinger distance is always bounded by 1. In this section, we show that if we restrict the space to $\Omega_{m_f}(\varepsilon)$, it is possible to exhibit risk bounds in terms of Kullback-Leibler.

Then we show that the controls derived for the statistical oracle inequality are sufficient to achieve a trajectorial inequality. It is however likely that at the price of additional technical results one could achieve a finer inequality with optimal bounds.

5.1. Risk bounds

The following theorem gives a risk bound of the penalized estimator in terms of Kullback-Leibler on $\Omega_{m_f}(\varepsilon)$.

Theorem 5.1. *Let $s = \prod_t s(t)$ be a distribution from the exponential family such that its parameters $\boldsymbol{\theta}_t$ belong to a convex compact set $\boldsymbol{\nu}$ of $\mathbb{R}^d$ with diameter ν. Assume that there exist two positive constants v and c such that s is $\mathbf{SG}(v,c)$ and let κ denote $\max\{v,c\}$. Let $\mathcal{M}_n$ be a collection of partitions constructed on a partition m_f such that there exists $\Gamma > 0$ satisfying $\forall J \in m_f, |J| \geq \Gamma \log(n)^2$, and let $(L_m)_{m\in\mathcal{M}_n}$ be some family of positive weights satisfying*

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

Let $\varepsilon > 0$ and let β'_ε be a positive constant depending on ε, the compact $\boldsymbol{\nu}$ and the distribution $\mathcal{G}$. If for every $m \in \mathcal{M}_n$

$$pen(m) \geq \beta'_\varepsilon d|m| \left(1 + 4\sqrt{L_m}\right)^2,$$

then

$$\begin{aligned}\mathbf{E}\left[K(s,\hat{s}_{\hat{m}})\mathbf{1}_{\Omega_{m_f}(\varepsilon)}\right] &\leq C'_{\beta_\varepsilon} \inf_{m\in\mathcal{M}_n} \left[K(s,\bar{s}_m) + pen(m)\right] \\ &\quad + C'(d,\nu,m_\nu,m_\varepsilon,\varepsilon,\Gamma,\kappa,V^{min},V^{max},\Sigma).\end{aligned}$$

Corollary 5.2. *Under the assumptions of Corollary 3.5, there exist some constants C_1 and C_2 such that*

$$\mathbf{E}\left[K(s,\hat{s}_{\hat{m}})\mathbf{1}_{\Omega_{m_f}(\varepsilon)}\right] \leq C_1 \log(N) \inf_{m\in\mathcal{M}_n} \{\mathbf{E}[K(s,\hat{s}_m)]\} + C_2$$

This theorem is proved in the exact same manner as Theorem 3.2, with the exception of the control of the term $\overline{\gamma_n}(s) - \overline{\gamma_n}(\bar{s}_{m'})$ (corresponding to proposition 4.3) which can be performed more finely on $\Omega_{m_f}(\varepsilon)$ via the following proposition.

Proposition 5.3. *Let $x > 0$, then with probability greater than $1 - de^{-x}$*

$$(\overline{\gamma_n}(s) - \overline{\gamma_n}(\bar{s}_{m'}))\mathbf{1}_{\Omega_{m_f}(\varepsilon)} \leq \frac{1}{1+\varepsilon} K(s,\bar{s}_{m'}) + dV^{max}(1+\varepsilon)x\left(\frac{1}{m_\nu} + \frac{\nu}{3}\right). \quad (14)$$

The proof is given in the appendix B.5. Now, defining a new set

$$\Omega_2(\varepsilon,\xi) = \bigcap_{m'\in\mathcal{M}_n} \{(\overline{\gamma_n}(s) - \overline{\gamma_n}(\bar{s}_{m'}))\, \mathbf{1}_{\Omega_{m_f}(\varepsilon)} \le \frac{1}{1+\varepsilon} K(s,\bar{s}_{m'}) \\ + dV^{max}(1+\varepsilon)(L_{m'}|m'|+\xi)\left(\frac{1}{m_\nu}+\frac{\nu}{3}\right)\},$$

we get $\mathbf{P}\left(\Omega_2(\varepsilon,\xi)^C\right) \ge d\Sigma e^{-\xi}$ using proposition 5.3. Replacing this new set in $\Omega(\varepsilon,\xi) = \Omega_{m_f}(\varepsilon) \cap \Omega_1(\xi) \cap \Omega_2(\varepsilon,\xi)$, we obtain

$$\begin{aligned}
&(\overline{\gamma_n}(\bar{s}_m) - \overline{\gamma_n}(\hat{s}_{\hat{m}}))\mathbf{1}_{\Omega(\varepsilon,\xi)} \\
&\quad \le \frac{1}{1+\varepsilon}(K(s,\bar{s}_{\hat{m}}) + K(\bar{s}_{\hat{m}},\hat{s}_{\hat{m}}))\mathbf{1}_{\Omega(\varepsilon,\xi)} + R\mathbf{1}_{\Omega(\varepsilon,\xi)} \\
&\qquad +\ dV^{max}(1+\varepsilon)(L_{\hat{m}}|\hat{m}|+\xi)\left[\frac{1}{m_\nu}+\frac{\nu}{3}\right] \\
&\qquad +\ dC(\varepsilon)\left[|\hat{m}| + 8(1+\varepsilon)\sqrt{(L_{\hat{m}}|\hat{m}|+\xi)|\hat{m}|} + 4(1+\varepsilon)(L_{\hat{m}}|\hat{m}|+\xi)\right] \\
&\quad \le \frac{1}{1+\varepsilon}(K(s,\bar{s}_{\hat{m}}) + K(\bar{s}_{\hat{m}},\hat{s}_{\hat{m}}))\mathbf{1}_{\Omega(\varepsilon,\xi)} + R\mathbf{1}_{\Omega(\varepsilon,\xi)} \\
&\qquad + dC(\varepsilon)|\hat{m}|\left[1 + (1+\varepsilon)\left(8\sqrt{L_{\hat{m}}} + \varepsilon + 4L_{\hat{m}}\right) + L_{\hat{m}} 2m_\varepsilon\left(\frac{1}{m_\nu}+\frac{\nu}{3}\right)\right] \\
&\qquad + d(1+\varepsilon)\left(4C(\varepsilon)\left(\frac{4}{\varepsilon}+1\right) + V^{max}\left(\frac{1}{m_\nu}+\frac{\nu}{3}\right)\right)\xi,
\end{aligned}$$

with $C(\varepsilon) = \dfrac{V^{max}\kappa(1+\varepsilon)}{2m_\varepsilon}$ and since $\kappa \ge 1$. Then we get

$$\begin{aligned}
&\frac{\varepsilon}{1+\varepsilon}K(s,\hat{s}_{\hat{m}})\mathbf{1}_{\Omega(\varepsilon,\xi)} \\
&\quad \le K(s,\bar{s}_m) + R\mathbf{1}_{\Omega(\varepsilon,\xi)} - pen(\hat{m}) + pen(m) + d|\hat{m}|C_{max}(\varepsilon)\left(1+4\sqrt{L_{\hat{m}}}\right)^2 \\
&\qquad + \xi d(1+\varepsilon)\left[V^{max}\left(\frac{1}{m_\nu}+\frac{\nu}{3}\right) + 4C(\varepsilon)\left(\frac{4}{\varepsilon}+1\right)\right],
\end{aligned}$$

where $C_{max}(\varepsilon) = V^{max}(1+\varepsilon)^3\kappa\left(\frac{1}{2m_\varepsilon} \vee \left(\frac{1}{m_\nu}+\frac{\nu}{3}\right)\right)$. Then since by assumption, $pen(\hat{m}) \ge \beta'_\varepsilon d|\hat{m}|\left(1+4\sqrt{L_{\hat{m}}}\right)^2$, choosing $\beta'_\varepsilon = C_{max}(\varepsilon)$ yields:

$$K(s,\hat{s}_{\hat{m}})\mathbf{1}_{\Omega(\varepsilon,\xi)} \quad \le \quad C'_{\beta_\varepsilon}\left[K(s,\bar{s}_m) + R\mathbf{1}_{\Omega(\varepsilon,\xi)} + pen(m)\right] + d\xi C'(\beta_\varepsilon),$$

with $C'_{\beta_\varepsilon} = \dfrac{1+C^{-1}_{max}(\beta'_\varepsilon)}{C^{-1}_{max}(\beta'_\varepsilon)}$, and finally

$$\begin{aligned}
&\mathbf{E}\left[K(s,\hat{s}_{\hat{m}})\mathbf{1}_{\Omega_{m_f}(\varepsilon)}\right] \\
&\quad \le C'_{\beta_\varepsilon}\left[K(s,\bar{s}_m) + \frac{C(a,d,\nu,m_\nu,m_\varepsilon,\varepsilon,\Gamma,\kappa,V^{min},V^{max})}{n^{(a-1)/2}} + pen(m)\right]
\end{aligned}$$

$$
\begin{aligned}
&+ 2d\Sigma C'(\beta_\varepsilon),\\
&\leq C'_{\beta_\varepsilon}\left[K(s,\bar{s}_m) + pen(m)\right] + C'(d,\nu,m_\nu,m_\varepsilon,\varepsilon,\Gamma,\kappa,V^{min},V^{max},\Sigma)
\end{aligned}
$$

which we minimize over $m \in \mathcal{M}_n$.

5.2. A trajectorial oracle inequality

When considering the loss approach, the definition of the oracle corresponds to the best estimator within the collection we consider in terms of Kullback-Leibler loss, *i.e.* the estimator associated to $m_D = \arg\min_m[K(s,\hat{s}_m)]$. The following theorem gives an oracle inequality in this sense.

Theorem 5.4. *Under the assumptions of theorem 5.1, let $\varepsilon > 0$ and let β''_ε be a positive constant depending on ε, the compact $\boldsymbol{\nu}$ and the distribution $\mathcal{G}$. If for every $m \in \mathcal{M}_n$*

$$pen(m) \geq \beta''_\varepsilon d|m|\left(1 + 4\sqrt{L_m}\right)^2,$$

then for all $\xi > 0$, there exists an event of probability at least $1 - 2d\Sigma e^{-\xi}$ on which

$$
\begin{aligned}
K(s,\hat{s}_{\hat{m}})\mathbf{1}_{\Omega_{m_f}(\varepsilon)} \quad \leq \quad & C''_{\beta_\varepsilon}\inf_{m\in\mathcal{M}_n}\left\{K(s,\hat{s}_m)\mathbf{1}_{\Omega_{m_f}(\varepsilon)} + pen(m)\right\}\\
&+C''(d,\nu,\kappa,\Gamma,\beta_\varepsilon,\Sigma,V^{max},V^{min})\xi. \qquad (15)
\end{aligned}
$$

To prove this result, we start with a slightly different decomposition of $K(s,\hat{s}_{\hat{m}})$ than equation (10):

$$
\begin{aligned}
K(s,\hat{s}_{\hat{m}}) \quad \leq \quad & K(s,\hat{s}_m) + (\overline{\gamma_n}(\hat{s}_m) - \overline{\gamma_n}(\bar{s}_m)) + (\overline{\gamma_n}(\bar{s}_m) - \overline{\gamma_n}(\hat{s}_{\hat{m}}))\\
&-pen(\hat{m}) + pen(m),
\end{aligned}
$$

where, for any $m' \in \mathcal{M}_n$, we still use the decomposition (11):

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

The two first terms of this latter decomposition are controlled in the same manner as for Theorem 5.1 (using propositions 4.1 and 5.3 respectively). Though the two other terms do not require a uniform control, the former results encompass their control for a fixed partition m. This results in

$$
\begin{aligned}
&(\overline{\gamma_n}(\hat{s}_m) - \overline{\gamma_n}(\hat{s}_{\hat{m}}))\mathbf{1}_{\Omega(\varepsilon,\xi)}\\
&\quad \leq \frac{1}{1+\varepsilon}(K(s,\bar{s}_{\hat{m}}) + K(\bar{s}_{\hat{m}},\hat{s}_{\hat{m}}))\mathbf{1}_{\Omega(\varepsilon,\xi)}\\
&\qquad + \frac{1}{1+\varepsilon}(K(s,\bar{s}_m) + K(\bar{s}_m,\hat{s}_m))\mathbf{1}_{\Omega(\varepsilon,\xi)}
\end{aligned}
$$

$$
\begin{aligned}
&+\ dV^{max}(1+\varepsilon)(L_{\hat{m}}|\hat{m}|+\xi)\left[\frac{1}{m_\nu}+\frac{\nu}{3}\right]\\
&+\ dV^{max}(1+\varepsilon)(L_m|m|+\xi)\left[\frac{1}{m_\nu}+\frac{\nu}{3}\right]\\
&+\ dC(\varepsilon)\left[|m|+8(1+\varepsilon)\sqrt{(L_m|m|+\xi)|m|}+4(1+\varepsilon)(L_m|m|+\xi)\right]\\
&+\ dC(\varepsilon)\left[|\hat{m}|+8(1+\varepsilon)\sqrt{(L_{\hat{m}}|\hat{m}|+\xi)|\hat{m}|}+4(1+\varepsilon)(L_{\hat{m}}|\hat{m}|+\xi)\right].
\end{aligned}
$$

Then choosing $\beta''_\varepsilon = C_{max}(\varepsilon) = \beta'_\varepsilon$, we get, with probability greater than $1 - 2\Sigma de^{-\xi}$,

$$
\begin{aligned}
&K(s,\hat{s}_{\hat{m}})\mathbf{1}_{\Omega_{m_f}(\varepsilon)}\\
&\leq \frac{2+\varepsilon}{\varepsilon}K(s,\hat{s}_m)\mathbf{1}_{\Omega_{m_f}(\varepsilon)}+2\frac{1+\varepsilon}{\varepsilon}pen(m)\\
&\qquad +2\xi d\frac{(1+\varepsilon)^2}{\varepsilon}\left[V^{max}\left(\frac{1}{m_\nu}+\frac{\nu}{3}\right)+4C(\varepsilon)\left(\frac{4}{\varepsilon}+1\right)\right].
\end{aligned}
$$

With $C''_{\beta_\varepsilon}=2C'_{\beta_\varepsilon}=2\dfrac{1+C^{-1}_{max}(\beta'_\varepsilon)}{C^{-1}_{max}(\beta'_\varepsilon)}$, we obtain

$$
K(s,\hat{s}_{\hat{m}})\mathbf{1}_{\Omega_{m_f}(\varepsilon)} \quad\leq\quad 2C'_{\beta_\varepsilon}\left[K(s,\hat{s}_m)\mathbf{1}_{\Omega_{m_f}(\varepsilon)}+pen(m)\right]+2d\xi C'(\beta_\varepsilon).
$$

6. A developed example: Categorical distributions

In this section we study the particular case of categorical variables by first illustrating how our general approach can be used when working with a specific distribution, and then showing how the main results could be derived if working directly with the characteristics of the categorical distribution instead of dealing with the log-partition function from the exponential family. We focus only on our Hellinger oracle inequality which is defined on the whole space. We conclude this section by comparing the constants obtained in each case.

6.1. Categorical distribution: The general approach

Here we suppose that Y can take values between 1 and $d+1$, with $d \geq 2$ and we write $\{p^{(i)}; 1 \leq i \leq d\}$ the probability that Y belongs to categories 1 through d (so that $p^{(d+1)} = 1 - \sum_{i=1}^d p^{(i)}$).

In the canonical form, the parameters $\boldsymbol{\theta}$ are given by $\theta^{(i)} = \log \frac{p^{(i)}}{1-\sum_{i=1}^d p^{(i)}}$ (for $1 \leq i \leq d$) and we have

- $T^{(i)} = \mathbf{1}_{\{Y=i\}}$
- $A(\boldsymbol{\theta}) = \log(1+\sum_{i=1}^d e^{\theta^{(i)}})$

- $E^{(i)} = \dfrac{e^{\theta^{(i)}}}{1+\sum_{j=1}^{d} e^{\theta^{(j)}}}$ and

$$\nabla^2 A(\boldsymbol{\theta}) = \begin{pmatrix} E^{(1)}(1-E^{(1)}) & -E^{(1)}E^{(2)} & \dots & -E^{(1)}E^{(d)} \\ -E^{(1)}E^{(2)} & E^{(2)}(1-E^{(2)}) & \dots & -E^{(2)}E^{(d)} \\ \vdots & & & \\ -E^{(1)}E^{(d)} & \dots & & E^{(d)}(1-E^{(d)}) \end{pmatrix}$$

The inverse mapping of the gradient of the log-partition function is given by $[[\nabla A]^{-1}(\boldsymbol{\beta})]^{(i)} = \log \dfrac{\beta^{(i)}}{1-\sum_{j=1}^{d} \beta^{(j)}}$.

The empirical contrast can therefore be expressed as

$$\gamma_n(s) = \sum_{t=1}^{n} \left[\log\left(1+\sum_{i=1}^{d} e^{\theta_t^{(i)}}\right) - \sum_{i=1}^{d} \mathbf{1}_{\{Y_t=i\}} \theta_t^{(i)} \right],$$

which translates into usual notations into

$$\gamma_n(s) = -\sum_{t=1}^{n} \sum_{i=1}^{d+1} \mathbf{1}_{\{Y_t=i\}} p_t^{(i)}.$$

6.1.1. Sub-gamma property of the categorical distribution

Let us assume that the probabilities of each category (including category $d+1$) are bounded away from 0 and 1, so that there exists $a>0$ and $b<1$ such that $a \leq p^{(i)} \leq b$ for all i. For sake of simplicity, we take $a=\rho$ and $b=1-\rho$ for $\rho>0$ small. This translates into inequality

$$-\log\frac{1-\rho}{\rho} < \theta^{(i)} < \log\frac{1-\rho}{\rho}$$

for the natural parameters. Our compact set $\boldsymbol{\nu}$ is therefore defined by $\big[-\log\frac{1-\rho}{\rho}; \log\frac{1-\rho}{\rho}\big]^d$, and $\nu = 2\log\frac{1-\rho}{\rho}$.

Moreover, $V^{(i)} = p^{(i)}(1-p^{(i)}) \leq (1-\rho)^2$, and in the sequel we choose $V^{max} = (1-\rho)^2$.

The Laplace transform of sufficient statistic $T^{(i)}$ is

$$\begin{aligned} \log \mathbf{E}\left[e^{z\left(T^{(i)}-E^{(i)}\right)}\right] &= \log\left[\frac{1+\sum_j e^{\theta^{(j)}} + e^{\theta^{(i)}}(e^z-1)}{1+\sum_j e^{\theta^{(j)}}}\right] \\ &= \log\left[1+E^{(i)}(e^z-1)\right] = \sum_{k\geq 2} c_k^{(i)} \frac{z^k}{k!}, \end{aligned}$$

which is analytic in z provided $z \leq \log\left(1+\frac{1}{E^{(i)}}\right)$. Setting $R = \log 2$ and $z \in [-R; R]$, one can show (see for instance [25]) that the cumulants associated with $T^{(i)}$ (the $\{c_k^{(i)}\}_k$) can be obtained through the recurrence property $c_1^{(i)} = p^{(i)}$ and for $k \geq 2$,

$$c_{k+1}^{(i)} = p^{(i)}(1-p^{(i)})\frac{\partial c_k^{(i)}}{\partial p^{(i)}},$$

where we have switched back to usual proportion notations for sake of readability.

By induction, we can show that

$$\left|\frac{c_k(T^{(i)})}{c_2(T^{(i)})^{k/2}}\right| \leq \frac{1}{2}k!\left(\frac{2}{\sqrt{V^{(i)}}}\right)^{k-2},$$

and finally that the categorical distribution is $\mathbf{SG}(1,2)$, so that we can take $\kappa = 2$.

6.1.2. Computation of m_ν and m_ε

In what follows, we will consider a fixed ε such that $0 < \varepsilon < \rho$.

Consider a given vector of proportions $\pi = (\pi^{(1)}, \ldots, \pi^{(d)})$, and the variance-covariance matrix $A(\pi)$ associated to a categorical distribution of parameter π. Gershgorin theorem states that all eigenvalues of $A(\pi)$ lie within at least one of the Gershgorin discs, *i.e.* they are all greater than

$$\begin{aligned}\Pi(\pi) &= \min_i \left\{\pi^{(i)}(1-\pi^{(i)}) - \sum_{j\neq i}\left|-\pi^{(i)}\pi^{(j)}\right|\right\}\\ &= \min_i \left\{\pi^{(i)}\left(1-\pi^{(i)}\right) - \pi^{(i)}\left(1-\pi^{(i)}-\pi^{(d+1)}\right)\right\}\\ &= \min_i \left\{\pi^{(i)}\pi^{(d+1)}\right\}\end{aligned}$$

From this, we directly obtain that $m_\nu = \rho^2$.

To compute m_ε, we consider $z \in K(\varepsilon)$. There exists E and V such that $\nabla A(z) \in \mathcal{B}(E, \varepsilon V)$, that is there exists E and V such that $\forall i$, $\frac{e^{z^{(i)}}}{1+\sum e^{z^{(j)}}}$ belongs to $[E^{(i)} - \varepsilon V; E^{(i)} + \varepsilon V]$. We therefore have:

$$\frac{e^{z^{(i)}}}{\left[1+\sum e^{z^{(j)}}\right]} \leq (E^{max} + \varepsilon V^{max}) \leq (1-\rho)(1+\varepsilon) \quad \text{and}$$

$$(\rho(1+\varepsilon)-\varepsilon) \leq (E^{min} - \varepsilon V^{max}) \leq \frac{e^{z^{(i)}}}{\left[1+\sum e^{z^{(j)}}\right]}$$

Additional technical computations show that m_ε is lower bounded by $(\rho-\varepsilon)^2$.

6.2. Direct results for the categorical variables

This section is dedicated to the derivation of direct controls when studying the categorical distribution. As before, Y will take values in $1, \dots, d+1$ with $p^{(i)}$ the probability of category i, which we once more bound by $\rho \leq p^{(i)} \leq 1-\rho$ for all i. Once again, it is possible to reduce the number of parameters to d, however keeping all $d+1$ parameters leads to more tractable quantities to control and to smaller resulting constants. For sake of readability and to avoid confusions, we will denote $r = d+1$.

The distribution $s(t)$ of Y_t can be decomposed in r terms by denoting $s(t,i) = \mathbf{P}\left(Y_t = i\right) = p_t^{(i)} = \mathbf{E}\left[\mathbf{1}_{\{Y_t=i\}}\right]$. In this specific case, quantities such as the contrast or the Kullback-Leibler risk can be identified:

- the model S_m associated to a given partition m is:

$$\mathcal{S}_m = \left\{ \begin{array}{c} u : \{1, ..., n\} \times \{1, ..., r\} \rightarrow [0,1] \text{ such that} \\ \forall J \in m, \forall i \in \{1, 2, ..., r\}, u\left(t, i\right) = u\left(J, i\right) \quad \forall\, t \in J \end{array} \right\}.$$

- the empirical contrast is the negative log-likelihood defined by:

$$\gamma_n\left(s\right) = -\sum_{t=1}^{n} \sum_{i=1}^{r} \mathbf{1}_{\{Y_t=i\}} \log\left[s\left(t,i\right)\right].$$

- associated to this contrast, the Kullback-Leibler information between s and u is: $K\left(s,u\right) = \sum_{t=1}^{n} \sum_{i=1}^{r} s\left(t,i\right) \log\left[\frac{s(t,i)}{u(t,i)}\right]$.
- the minimum contrast estimator of s is

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

  with $N_J\left(i\right) = \sum_{t \in J} \mathbf{1}_{\{Y_t=i\}}$, and the projection

$$\bar{s}_m\left(t,i\right) \quad = \quad \arg\min_{u \in \mathcal{S}_m} K(s,u) = \frac{\sum_{t \in J} s\left(t,i\right)}{\mid J \mid}.$$

See for instance [20] for more details on notations and computations.

The following theorem gives the direct version of Theorem 3.2.

Theorem 6.1. *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 $i \in \{1, 2, ..., r\}$, $P\left(Y_t = i\right) = s\left(t,i\right)$. Let $\left(L_m\right)_{m \in \mathcal{M}_n}$ be some family of positive weights and define Σ as*

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

Assume that

- *there exists some positive absolute constant ρ such that $\rho \leq s(t,i) \leq 1-\rho$, for all t and i,*

- *$\mathcal{M}_n$ is a collection of partitions constructed on a partition m_f such that $|J| \geq \Gamma\left[\log(n)\right]^2 \forall J \in m_f$ where Γ is a positive absolute constant.*

Let $\lambda_\varepsilon > 1/2$. If for every $m \in \mathcal{M}_n$

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

then

$$\mathbf{E}\left[h^2\left(s, \hat{s}_{\hat{m}}\right)\right] \leq C_{\lambda_\varepsilon} \inf_{m \in \mathcal{M}_n} \left\{K\left(s, \overline{s}_m\right) + pen\,(m)\right\} + C\left(\Sigma, r, \lambda_\varepsilon, \rho, \Gamma\right),$$

with $C_{\lambda_\varepsilon} = \frac{2(2\lambda_\varepsilon)^{1/3}}{(2\lambda_\varepsilon)^{1/3} - 1}$.

The proof of this result is postponed to the appendix C.

6.3. Comparison of the constants

The main constant of interest is C_{β_ε} as it compares the risk of our estimator to that of the oracle.

In the general approach, this constant is expressed as $C_{\beta_\varepsilon} = \frac{1+C_2^{-1}(\varepsilon)}{C_2^{-1}(\varepsilon)}$ where $C_2(\varepsilon) = \beta_\varepsilon = \dfrac{V^{max}\kappa}{2m_\varepsilon}(1+\varepsilon)^3$. Using the results of Section 6.1, we have

$$C_2(\varepsilon) = (1+\varepsilon)^3 \frac{(1-\rho)^2}{(\rho-\varepsilon)^2} \leq \left(\frac{1+\varepsilon}{\rho-\varepsilon}\right)^3$$

and finally

$$C_{\beta_\varepsilon} \leq \frac{(1+\rho)(\beta_\varepsilon)^{1/3}}{\rho(\beta_\varepsilon)^{1/3} - 1} \quad \text{with } \beta_\varepsilon > \frac{1}{\rho^3}.$$

In the direct approach, the constant of interest is $C_{\lambda_\varepsilon} = \frac{2(2\lambda_\varepsilon)^{1/3}}{(2\lambda_\varepsilon)^{1/3}-1}$ with $\lambda_\varepsilon = \frac{1}{2}\frac{(1+\varepsilon)^3}{(1-\varepsilon)^3} > \frac{1}{2}$.

The constants β_ε and λ_ε are the constants from the penalty function ($pen(m) > \beta_\varepsilon(r-1)|m|(1+4\sqrt{L_m})^2$ in the general approach and $pen(m) > \lambda_\varepsilon r|m|(1+4\sqrt{L_m})^2$ in the direct approach). One can note that β_ε is greater than λ_ε, and even, $\beta_\varepsilon(r-1)$ is greater than $\lambda_\varepsilon\ r$, meaning that the constraint on the penalty function in the direct approach is finer than in the general approach.

More importantly, as ρ goes to 0, β_ε explodes while the constraint on λ_ε is unaffected. This dependency on ρ in the general framework comes from the $\frac{V^{max}}{m_\varepsilon}$ ratio, which intervenes when comparing the Kullback divergence and our V^2 term between the penalized estimator $\hat{s}_m$ and the projection of the distribution $\overline{s}_m$. Because the distribution is not specified, no direct simplification occurs and the use of Taylor development leads to a constant $\frac{m_\varepsilon}{2V^{max}}$ When using the constraints on the proportions, the term $\frac{V^{max}}{m_\varepsilon}$ will be of the order of $\frac{V^{max}}{V^{min}} \simeq \frac{(1-\rho)^2}{\rho^2}$ which explodes when ρ goes to 0. In the direct approach, simplifications of the log-partition function allow to work directionally, and therefore avoid the global ratio of variances: the smallest eigenvalue of the covariance matrix in the general case is now directly the variance of the sufficient statistic, and the ratio simplifies.

Concerning the constants of interest, we can notice that the general shape of C_{β_ε} and C_{λ_ε} are the same. In fact, if we introduce $\gamma_\varepsilon = \frac{1}{2}\rho^3\beta_\varepsilon$ so that $\gamma_\varepsilon > \frac{1}{2}$, we get

$$C_{\gamma_\varepsilon} \quad \leq \quad \frac{1+\rho}{\rho}\frac{(2\gamma_\varepsilon)^{1/3}}{(2\gamma_\varepsilon)^{1/3}-1}.$$

so that $C_{\gamma_\varepsilon} \simeq \frac{1+\rho}{\rho}\frac{1}{2}C_{\lambda_\varepsilon}$ and this constant gets significantly bigger as the constraints on p loosens (*i.e.* ρ decreases).

Note that in practice, both penalty constants β_ε and λ_ε are tuned from the data. More precisely, as is classical in segmentation or density estimation settings, the slope heuristic [3] can be used to this purpose. Since the empirical contrast is the same regardless of the computations of the controls, providing the shape $|m|(1+4\sqrt{L_m})^2$ to the slope heuristic will result in the choice of the same penalized estimator - due to the same choice of number of segments K. Consequently, we will obtained the same oracle constant C_1, smaller than both C_{β_ε} and C_{λ_ε} since the later are only upper bounds on the true oracle constant.

7. Simulation study and application to DNA sequences

In this section, we apply our estimator in three scenarios. The first is a simulation study where the observations are sampled from the exponential distribution with piece-wise constant rate parameter. In this case the distribution is continuous and the sufficient statistic is uni-dimensional (and is the variable itself). The second is a small comparison study with some of the related work presented in Section 2.3. Here we first simulate smaller profiles from the Poisson distribution (in order to accommodate as many available softwares as possible) and then study their behavior to model misspecification by simulating from a Student distribution while using a Gaussian segmentation model. Finally, we consider an application to a real data-set on the analysis of DNA sequences in terms of base-composition. Segmentation models are used to identify homogeneous regions which can be related to structural and functional biological regions. In this case we model the observations with a categorical distribution with piece-wise constant proportion parameters as described in the previous section.

7.1. Simulation study with exponential distribution

Simulation design. We simulated datasets of length $n = 10^5$ for which the number of segments K was drawn from a Poisson distribution with mean

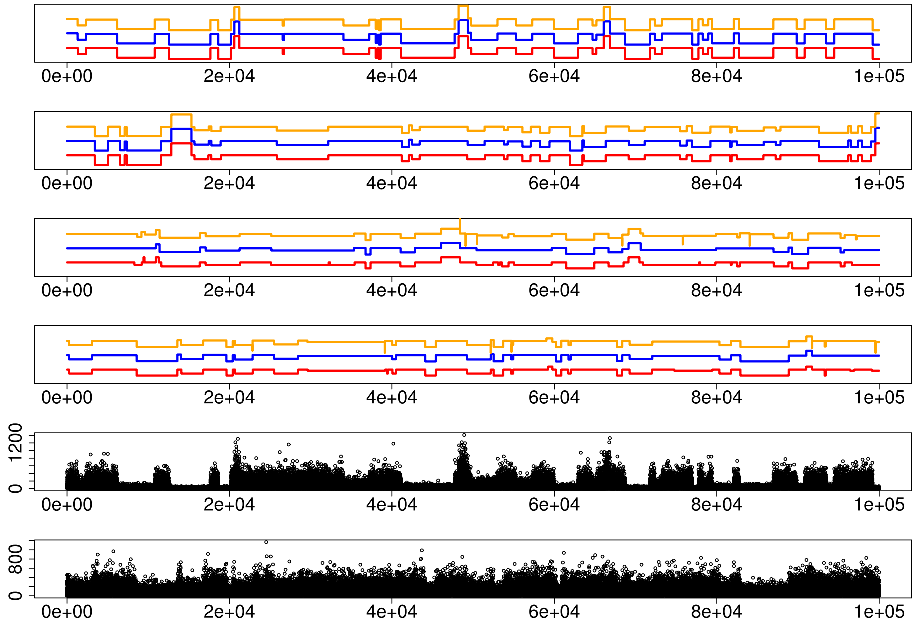

Fig 1. ***Simulated datasets and segmentation.*** *The four first figures show examples of simulation design for each of the four groups of rate values; the last two figures show examples of dataset to segment in the first and last group of values. Red lines (bottom) indicate the true distribution (the inverse of the rate, $1/\lambda$, is represented) while blue lines (middle) indicate the estimated distribution and the orange lines (top) indicate the optimal segmentation (w.r.t. likelihood loss) for the true number of segments.*

$\bar{K} = 50$, and the $K-1$ change-points were sampled uniformly on $\{2,\dots,n-1\}$ subject to the constraint that segments had to be of length at least 100. We considered 4 sets of values for the rate parameters of the exponential distribution. In all scenarios, odd segments had a rate of 0.01 while the rate on even segments was chosen randomly with probability $0.4, 0.2, 0.3$ and 0.1 among the values $(0.05, 0.1, 0.02, 0.005)$ for datasets 1 through 100, $(0.02, 0.05, 0.015, 0.005)$ for datasets 101 through 200, $(0.015, 0.02, 0.0125, 0.007)$ for datasets 201 through 300, and $(0.015, 0.02, 0.011, 0.008)$ for datasets 301 through 400. For each combination of parameters, we simulated 100 datasets, which are similar to those shown in Figure 1.

Quality criteria. The performance of our procedure was assessed *via* different criteria: if $\hat{s}$ denotes the considered estimator,

- The difference between the true number of segments and the estimated, $\Delta = K - \hat{K}$;
- The two components of the Hausdorff distance (as in [23]) in order to assess the quality of the change-point locations. More precisely, we consider the two quantities $\mathcal{E}(m_{\hat{s}}||m_s)$ and $\mathcal{E}(m_s||m_{\hat{s}})$, between the partitions

associated with the true and estimated distributions s and $\hat{s}$, where

$$\mathcal{E}(m_A||m_B) = \sup_{b\in m_B} \inf_{a\in m_A} |a-b|.$$

The first quantity $\mathcal{E}(m_{\hat{s}}||m_s)$ assesses how our estimated segmentation is able to recover the true change-points. Intuitively, segmentations with large number of segments are likely to yield a small value of $\mathcal{E}(m_{\hat{s}}||m_s)$. On the contrary, the second quantity $\mathcal{E}(m_s||m_{\hat{s}})$ judges how relevant the proposed change-points are compared to the true partition: a segmentation with too many segments will most likely have change-points far from the true ones (unless all change-points cluster around true breakpoints). The Hausdorff distance can then be recovered as $\sup\{\mathcal{E}(m_{\hat{s}}||m_s), \mathcal{E}(m_s||m_{\hat{s}})\}$.

- The Kullback divergence between true and estimated distributions, computed as $\sum_t \left(\frac{\hat{\lambda_t}}{\lambda_t} - \log\frac{\hat{\lambda_t}}{\lambda_t} - 1\right)$ for the Exponential distribution; and
- A pseudo-Hellinger distance between true and estimated distributions, defined as $\sum_t \left(1 - 2\frac{\sqrt{\lambda_t\hat{\lambda}_t}}{\lambda_t+\hat{\lambda}_t}\right)$.

Remark 7.1. *Note that the Hellinger distance between two exponential distributions is defined as* $1 - \prod_t \left(2\frac{\sqrt{\lambda_t\hat{\lambda}_t}}{\lambda_t+\hat{\lambda}_t}\right)$. *However, with profiles of length* 10^5*, computing this product yields values very close to the numerical precision of computers and therefore the Hellinger distance is almost always equal to one. We therefore chose to represent the pseudo-Hellinger distance, taking values between* 0 *and* n*, for better assessment of the performance of our estimator.*

In this framework, since the penalty depends on the partition through its dimension, the segmentation was performed using the pruned dynamic programming algorithm [35, 16] from which we obtained the optimal segmentation $\hat{m}_k$ for every k up to $Kmax = 200$, and where the optimal segmentation in k segments is defined as $\arg\min_{m\in\mathcal{M}_n^k} \gamma_n(\hat{s}_m)$ ($\mathcal{M}_n^k$ being the set of all partitions of $\{1,\dots,n\}$ that have k segments). The number of segments was then obtained using the penalty function proposed in (8) with $N = n$ and where the constant is calibrated using the slope heuristic (see [3]). The resulting estimated distribution can then be written as $\hat{s}_{\hat{m}_{\hat{K}}}$.

Since our datasets were simulated from a true segmentation model, to allow a fair assessment of our estimator, the Hausdorff, Kullback and Hellinger criteria are also computed for the optimal segmentation for the true number of segments, which we will denote $\hat{m}_K$, and the resulting estimated distribution $\hat{s}_{\hat{m}_K}$. These estimators differ in Figures 1 and 2 by the following colors: blue for our estimator $\hat{s}_{\hat{m}_{\hat{K}}}$ and orange for $\hat{s}_{\hat{m}_K}$.

Moreover, when comparing these estimators in terms of Kullback-Leibler divergence (in Figure 2 c), we also consider the trajectorial oracle model, that is the empirical estimator associated to $\hat{m}_{\tilde{K}}$ where $\tilde{K} = \arg\min_{k\in\{1,\dots,K_{max}\}} K(s, \hat{s}_{\hat{m}_k})$. This estimator is represented by the light blue color in Figure 2 c.

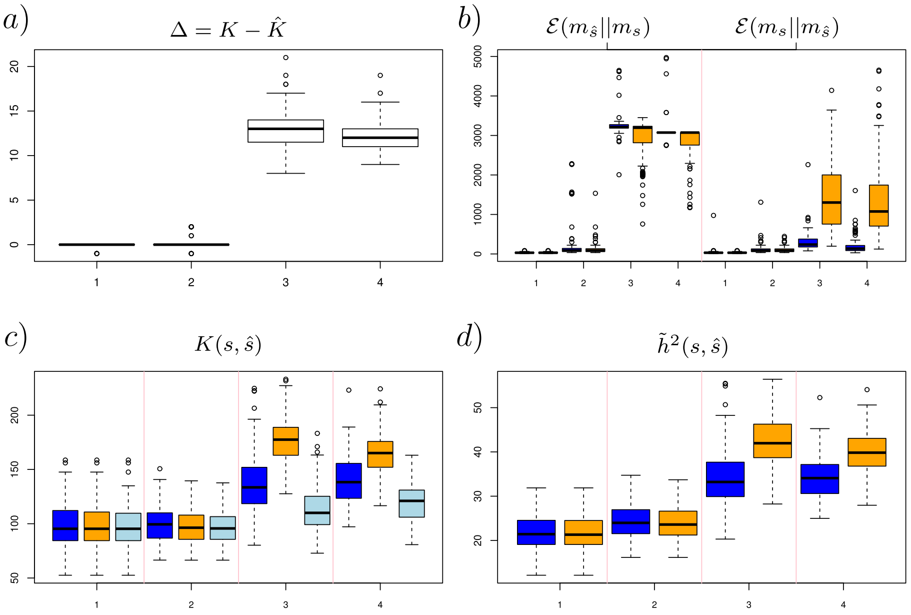

FIG 2. ***Evaluation of the performance of our proposed method.*** *a) difference between the true number of segments and the estimated number. b) boxplot of $\mathcal{E}(m_{\hat{s}}||m_s)$ (left) and $\mathcal{E}(m_s||m_{\hat{s}})$ (right) over the hundred simulations in each framework. In each case, left boxplots (blue) assess our estimator, right boxplots (orange) assess the estimator corresponding to the true number of segments. c) boxplot of the Kullback-Leibler divergence to the true distribution for the estimated distribution (blue-left), the optimal segmentation in K segments (orange-middle) in each simulation framework, and the optimal segmentation (light blue-right). d) same as c) but for the pseudo-Hellinger distance.*

Results. As is shown in Figure 2 a, when the detection problem is easy (sets 1 and 2), our method tends to recover the true number of segments, and therefore its performances are the same as that of $\hat{s}_{\hat{m}_K}$ (see Figures 2 b, c and d). However, when the scenario becomes more difficult to segment (sets 3 and 4), our method has a tendency to underestimate the number of segments. This behavior is classical and desired in studies of model selection for segmentation in order to avoid false detection. This phenomenon is further illustrated in Figure 2 b as our estimator yields high values of $\mathcal{E}(m_{\hat{s}}||m_s)$ due to the missed segments (left-hand-side of the figure) but low values of $\mathcal{E}(m_s||m_{\hat{s}})$ as the segments we propose tend to correspond to true segments (right-hand-side of the figure). On the contrary, the segmentation $\hat{m}_K$ has lower values of $\mathcal{E}(m_{\hat{s}}||m_s)$ as it has more change-points thus lower distances to the missed ones, but higher values of $\mathcal{E}(m_s||m_{\hat{s}})$ as some of its segments are spurious. On a particular example, presented in Figure 1 by comparing the red line (true segmentation) to the blue one (corresponding to our estimator), and more so on the bottom sub-figure illustrating a simulation from the fourth group, we observe that our method tends to fail to recover small segments with rate very close to surrounding ones, whereas the optimal segmentation for the true number of change-points (as represented by the orange line) still fails to recover those small segments, thus proposing additional ones, typically very short (average length=9) with very different rate values. Finally, Figures 2 c and d show that in terms of Kullback-Leibler divergence and pseudo-Hellinger distance, our estimator performs at least as well as the one with the true number of segments, and significantly better when the profiles become more difficult to segment. Moreover, as is shown in Figure 2 c, even in the most difficult scenarios, our estimator has a performance very close to that of the trajectorial oracle.

7.2. Comparison study and robustness to the model

Among the related works presented in Section 2.3, only 3 have an available R package, namely `ecp` for the permutation test approach of [33] implemented within the `edivisive` function, `EBS` [19] for the Bayesian ICL approach of [36] and `stepR` for the likelihood ratio approach of [21] implemented within the `smuceR` function.

Because of the high complexity of the non-parametric approaches and the little choice in exponential family distribution implemented in the various packages, we chose to compare the different approaches on small simulation studies (profile length of 1000 datapoints) in two cases: the first with data simulated from the exponential family, the Poisson distribution, and the second with data simulated from a Student distribution.

Poisson simulation. Here again the parameters of the simulations were drawn randomly as follows: the number of segments K was drawn from a Poisson distribution with mean $\bar{K} = 10$, and the $K-1$ change-points were sampled uniformly on $\{2, \dots, n-1\}$ subject to the constraint that segments had to be of length at least 20. We considered 4 sets of values for the Poisson rate parameters. In all scenarios, odd segments had a rate of 5 while the rate on even segments was chosen randomly with probability 0.4, $0.2, 0.3$ and 0.1 among the values $(3, 1, 10, 7, 5)$ for datasets 1 through 100, $(2, 1, 8, 6.5, 5)$ for datasets 101 through 200, $(2, 1.5, 6.5, 5.5, 5)$ for datasets 201 through 300 and $(4.5, 3.5, 6.5, 5.5, 5)$ for datasets 301 through 400.

The performance of the competitive segmentation methods were assessed *via*:

- The difference between the true number of segments and the estimated, $\Delta = K - \hat{K}$;
- The adjusted rand-index criterion
- $\mathcal{E}(m_{\hat{s}}||m_s)$ and $\mathcal{E}(m_s||m_{\hat{s}})$, between the estimators and the true segmentation; and
- The runtime of the algorithms.

All the results are given in Figure 3. All methods tend to underestimate the number of segments, especially when the profiles become harder to segment (see Figure 3 a). Our method performs slightly better in the easiest scenarios, while EBS performs significantly worse, almost unaffected by the different scenarios.

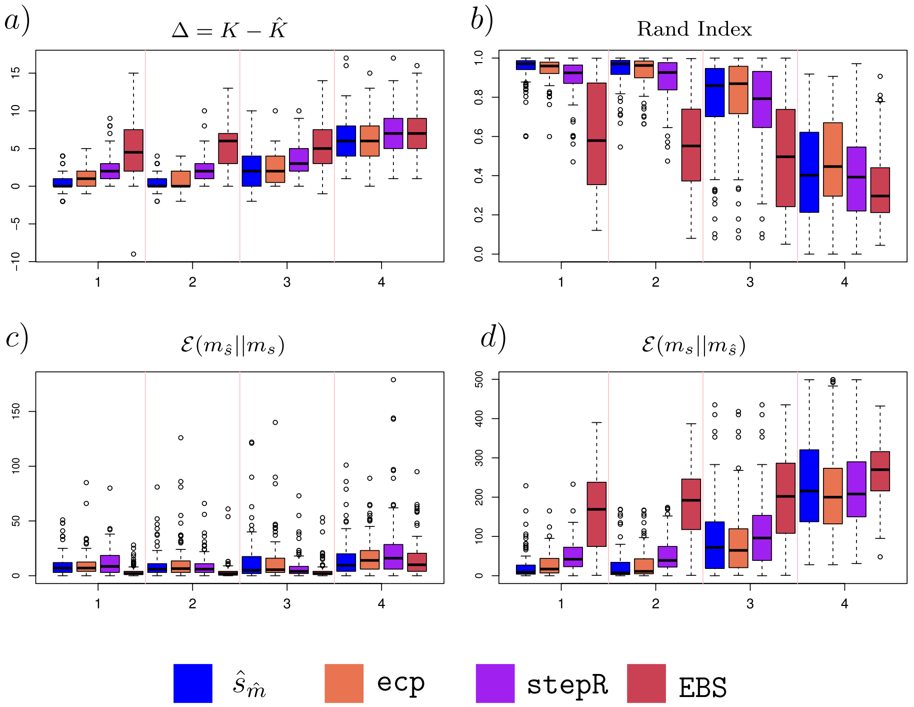

FIG 3. ***Comparison of our method with three others.*** *a) difference between the true number of segments and the estimated. b) boxplot of the adjusted rand-index criterion (1 is perfect agreement). c) and d) boxplot of $\mathcal{E}(m_{\hat{s}}||m_s)$ and $\mathcal{E}(m_s||m_{\hat{s}})$ respectively over the hundred simulations in each framework.*

Interestingly, the second best algorithm is the non-parametric approach of [33]. This behavior is translated in the Rand Index performance (see Figure 3 b), with our penalized estimator and the non-parametric approach exhibiting similar performances, tightly followed by the likelihood ratio approach of [21], while the Bayesian approach of [36] is the worst. Note however that the `EBS` algorithm does not provide an optimal segmentation but the posterior distribution of the change-points, and the Hausdorff and Rand Index criteria for this approach were computed using the MAP of each change-point location.

In terms of Hausdorff distance, all methods perform very well in terms of $\mathcal{E}(m_{\hat{s}}||m_s)$ criterion, as the change-points that are identified by the methods correspond to or are very close to true change-points. On the contrary, once again their performance in terms of $\mathcal{E}(m_s||m_{\hat{s}})$ is poorer as they tend to miss some of the change-points - as they underestimate the number of segments (see Figure 3 c and d). It is interesting to note that `EBS` has the best $\mathcal{E}(m_{\hat{s}}||m_s)$ performance. This is related to the fact that the ICL tends to select the number of segments for which the confidence in the segmentation is the highest, therefore those change-points that are recovered by the algorithm are recovered with very

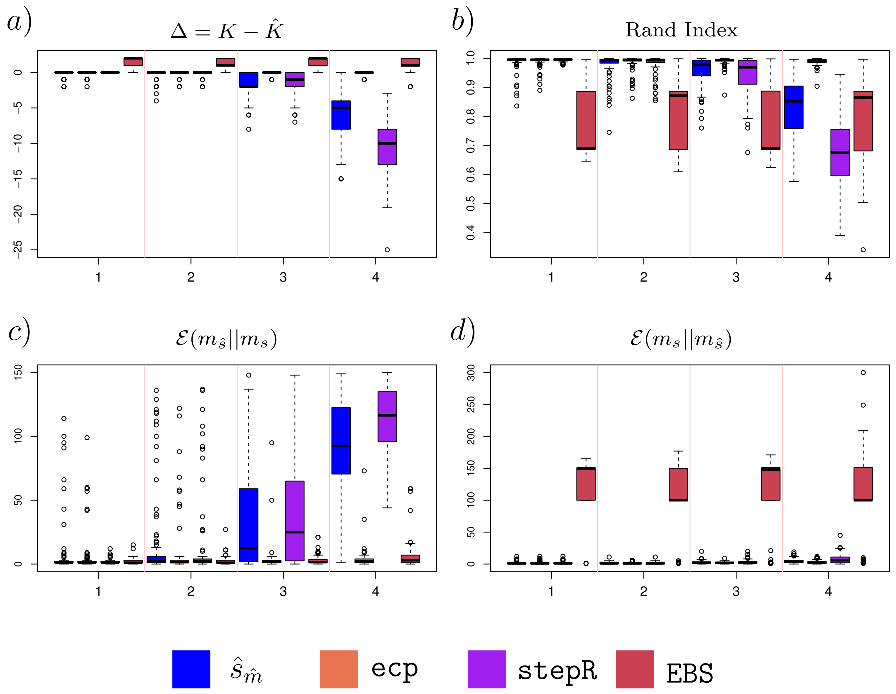

FIG 4. ***Comparison of our method with three others.*** *a) difference between the true number of segments and the estimated. b) boxplot of the adjusted rand-index criterion (1 is perfect agreement). c) and d) boxplot of $\mathcal{E}(m_{\hat{s}}||m_s)$ and $\mathcal{E}(m_s||m_{\hat{s}})$ respectively over the hundred simulations in each framework.*

high confidence, and they tend to correspond to true change-points. Once more, on those criteria, our penalized estimator exhibits the best performance tightly followed by the non-parametric approach `ecp`. Finally, in terms of runtime, our method out-beats all others with an average running time of 0.4 seconds per profile, followed by `stepR` with an average of 1.31 seconds, `ecp` with an average of 22.08 seconds and `EBS` with an average of 24 seconds.

Student simulation: This simulation study aims at assessing the robustness of our approach to the exponential family model assumption. To this effect, we simulated 1000 points datasets with both fixed change-points, $m = (300, 550, 700, 900, 1000)$ and fixed means on each segments: 0 on odd segments and 1 on even segments. We considered four noise scenarios, all from the Student distribution with different degrees of freedom $\nu = \{50, 10, 6, 3\}$ ($\nu = 50$ being the closest Gaussian case) and simulated 100 datasets for each scenario.

The three parametric approaches, our penalized estimator, stepR and EBS, were used with a Gaussian model assumption. The performance of the methods were assessed with the same criteria as for the Poisson distribution.

Results are given in Figure 4. We observe that the non-parametric approach `ecp` consistently out-beats all other methods, with excellent performances for all criteria. When the simulation scheme is close to a Gaussian simulation, our penalized estimator performs as well as `ecp`, however when the degree of freedom of the Student distribution decreases, the number of segments is overestimated, resulting in performances decrease in terms of Rand Index. It is interesting to note that in this case the results in terms of $\mathcal{E}(m_s||m_{\hat{s}})$ are quite stable and reasonable, as we do recover the true change-points, but the results in terms of $\mathcal{E}(m_{\hat{s}}||m_s)$ deteriorate as we identify spurious segments. This analysis remains true for the stepR approach, with performances deteriorating faster than for our penalized estimator. On the contrary, `EBS` still underestimates the number of segments, consistently identifying the first two change-points but missing the smaller segments, therefore its tendency in terms of segmentation quality criteria remain identical to that of the Poisson simulation.

Finally, in this scenario `ecp` is clearly worse than the other methods in terms of runtime performances: our method still out-beats all others with an average running time of 0.3 seconds per profile, followed by `stepR` with an average of 0.35 seconds, `EBS` with an average of 2.88 seconds and `ecp` with an average of 32.7 seconds. These differences in performance compared to the Poisson simulation (in particular, `EBS` drastically improves with a runtime about 10 times faster) are explained by the usage of the Gaussian loss that results in an optimization cost function which can be computed as a least-squares cost and for which the computational runtime is significantly reduced.

7.3. Application to a DNA sequence

The objective of this application is to find regions of a DNA sequence which are homogeneous in terms of base composition, that is which present a stability in the frequencies of the four nucleotide letters. These regions are thought to correspond to areas of the genome which are biologically significant. To this end, we apply our procedure modeling the data with categorical variables with $d = 3$ (see Section 6 for the model).

Here we consider the bacteriophage Lambda genome with length $n = 48502$ base pairs which is a parasite of the intestinal bacterium *Escherichia coli*. This genome has been used for the comparison of segmentation methods (see [11] and references therein) such as HMM ([9], [34]) or penalized quasi-likelihood [10]. The data and its annotation are publicly available from the National Center for Biotechnology Information (NCBI) pages at the url address http://www.ncbi.nlm.nih.gov/.

From a computational point of view, the large size of the Lambda genome hampers the direct use of the classical dynamic programming (DP) algorithm. Here we propose a hybrid algorithm that consists in first selecting a small number of relevant change-points using the CART algorithm [12], and then using DP on this set of candidate change-points. As in the simulation study, the penalty constant is calibrated using the slope heuristic. Four change-points (*i.e.* five segments) are selected by our criterion at positions 22546, 27829, 38004 and 46528. The associated regions are characterized by different base composition

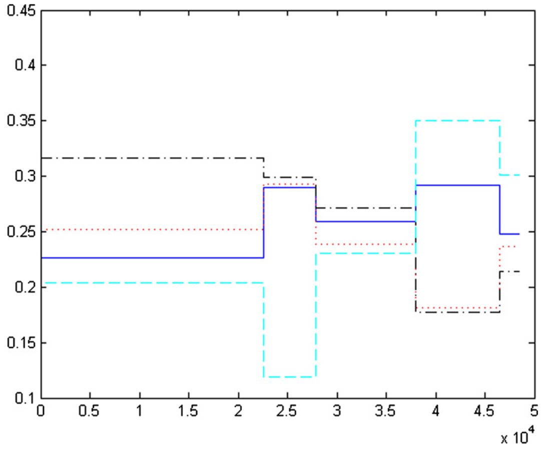

Fig 5. *Estimated probabilities on each segment of the selected segmentation: blue and '-' for the base adenine A, red and ':' for the base cytosine C, black and '-.' for the base guanine G and cyan and '–' for the base thymine T.*

as shown in Figure 5 which represents the estimated probabilities of each base for the obtained segmentation.

These change-points are very close (and even on some occasion the precise same) as the one obtained in [10], which concluded to 3 more change-points. This reference also supposes bases to be independent and uses a penalized contrast procedure to perform the segmentation, and is in this sense the closest approach to ours. The segments we identify reflect changes in transcription direction. Indeed, this direction is forward up to base 20855, it then switches to reverse from base 22686 to 37940, switches back to forward between 38041 to 46427 and finally reverse again from 46459 to the end. Note that a refinement has been obtained when assuming a dependence relationship between bases (see [9] and [34]).

8. Conclusion

We have proposed a general approach to the selection of the number of segments in the segmentation framework where the data can be modeled using a distribution from the exponential family. As expected, the log-partition function and its many properties are instrumental in the computation of the bounds and the derivation of the oracle inequality. While the main result compares the Hellinger risk of our penalized estimator to the Kullback-Leibler risk of the oracle, we show that on an event of large probability we can obtain both a similar statistical oracle inequality in terms of Kullback-Leibler risk and a trajectorial oracle inequality in terms of Kullback-Leibler divergence.

Our work therefore opens the door to a vast range of applications which can directly be tackled with our penalty function. Indeed, while the constants do depend on the choice of the distribution and on the compact set to which the parameters belong, they are in practice directly calibrated from the data using the slope heuristic. Here, using the particular case of categorical variables as an example, we have shown that the loss in tightness of the main constant is not a drastic issue by comparing the results obtained from the general approach to that from the direct one. Moreover, we believe that this work should easily be extended to multivariate distributions from the exponential family.

Moreover, we have shown in many examples through simulation and application studies (long profiles from the exponential distribution and shorter profiles from the Poisson distribution and Student distribution modeled through a Gaussian loss in the simulation studies, and categorical distribution in the application to DNA sequences) that our approach is a powerful method to detect significant changes in the distribution of the data, which can often be related to phenomenon of interest. When the model is well-specified, it out-beats all other similar approaches both in terms of segmentation quality and runtime, despite a tendency to underestimate the number of segments in order to avoid false detection. As the model becomes more ill-specified, for instance in the Student simulation study with small parameter, its performance deteriorates, with a tendency to over-segment the data. While still consistently recovering the true change-points, it proposes additional smaller segments to decrease the variance of the estimator on the segments. In such cases, at the cost of a higher computational complexity, one might prefer a non-parametric approach that does not suffer from those drawbacks.

Appendix A: Computations of $\hat{s}_m$ and $\bar{s}_m$

Let $m \in \mathcal{M}_n$ and $u \in \mathcal{S}_m$ such that for $J \in m$ and for $t \in J$ we have $u(t) = g(Y_t)\exp\left[\boldsymbol{p}_J.\mathbf{T}_t - A(\boldsymbol{p}_J)\right]$. Then

$$\gamma_n(u) = \sum_{J\in m}\sum_{t\in J}\Big[A(\boldsymbol{p}_J) - \boldsymbol{p}_J.\mathbf{T}_t\Big] = \sum_{J\in m}\gamma_n(u,J).$$

Since $\forall J \in m,\ \frac{d\gamma_n(u,J)}{d\boldsymbol{p}_J} = \sum_{t\in J}\left[\nabla A(\boldsymbol{p}_J) - \mathbf{T}_t\right] = |J|\nabla A(\boldsymbol{p}_J) - \mathbf{T}_J$, $\gamma_n(u,J)$ is minimal for $\boldsymbol{p}_J = [\nabla A]^{-1}(\bar{\mathbf{T}}_J)$, and finally

$$\hat{s}_m(t) \quad = \quad g(Y_t)\exp\left[[\nabla A]^{-1}(\bar{\mathbf{T}}_J).\mathbf{T}_t - A\left([\nabla A]^{-1}(\bar{\mathbf{T}}_J)\right)\right].$$

Similarly,

$$\begin{aligned} K(s,u) &= \sum_{J\in m}\sum_{t\in J}\Big[\nabla A(\boldsymbol{\theta}_t).(\boldsymbol{\theta}_t - \boldsymbol{p}_J) - (A(\boldsymbol{\theta}_t) - A(\boldsymbol{p}_J))\Big] \\ &= \sum_{J\in m}\sum_{t\in J}\Big[\ \mathbf{E}_t.(\boldsymbol{\theta}_t - \boldsymbol{p}_J) - (A(\boldsymbol{\theta}_t) - A(\boldsymbol{p}_J))\Big] = \sum_{J\in m}K_J(s,u). \end{aligned}$$

Since $\forall J \in m,\ \frac{dK_J(s,u)}{d\boldsymbol{p}_J} = -\mathbf{E}_J + |J|\nabla A(\boldsymbol{p}_J)$, $K_J(s,u)$ is minimal for $\boldsymbol{p}_J = [\nabla A]^{-1}(\bar{\mathbf{E}}_J)$, and finally

$$\bar{s}_m(t) \quad = \quad g(Y_t)\exp\left[[\nabla A]^{-1}(\bar{\mathbf{E}}_J).\mathbf{T}_t - A\left([\nabla A]^{-1}(\bar{\mathbf{E}}_J)\right)\right].$$

Appendix B: Results to prove the main Theorem 3.2 and Theorem 5.1

B.1. Intermediate results

Recall the definition of χ^2_m

$$\chi^2_m \;=\; \sum_{i=1}^{d} \chi^2(m,i) \;=\; \sum_{i=1}^{d} \sum_{J\in m} \frac{(T_J^{(i)} - E_J^{(i)})^2}{V_J^{(i)}},$$

and define, for $u(t) = \mathcal{G}(\boldsymbol{p}_t)$ and $u'(t) = \mathcal{G}(\boldsymbol{p}'_t)$,

$$V^2(u,u') = \sum_{i=1}^{d} \sum_{t} V_t^{(i)} \left[p_t^{(i)} - {p'}_t^{(i)}\right]^2 = \sum_{i=1}^{d} V_i^2(u,u'). \tag{16}$$

The two following subsections give the proof of the control of χ^2_m and links between χ^2_m, $V^2(\bar{s}_m, \hat{s}_m)$ and $K(\bar{s}_m, \hat{s}_m)$ respectively.

B.1.1. Proof of Proposition 4.1

We have

$$\mathbf{E}[\chi^2(m,i)] = \sum_{J\in m} \frac{1}{V_J^{(i)}} \mathbf{E}[(T_J^{(i)} - E_J^{(i)})^2] = |m|.$$

We introduce the variables $Z_J(i)$ such that

$$\chi^2(m,i) = \sum_{J\in m} Z_J(i) = \sum_{J\in m} \frac{(T_J^{(i)} - E_J^{(i)})^2}{V_J^{(i)}},$$

and control their moments using

$$\begin{aligned}
\mathbf{E}\left[Z_J(i)^p \mathbf{1}_{\Omega_{m_f}(\varepsilon)}\right] &\leq \left(\frac{1}{V_J^{(i)}}\right)^p \int_0^{+\infty} x^p \; dP\left[\left\{(T_J^{(i)} - E_J^{(i)})^2 \geq x\right\} \cap \Omega_{m_f}(\varepsilon)\right] dx \\
&\leq \left(\frac{1}{V_J^{(i)}}\right)^p \int_0^{\varepsilon V_J^{(i)}} 2p\; x^{2p-1} P\left[|T_J^{(i)} - E_J^{(i)}| \geq x\right] dx \\
&\leq \left(\frac{1}{V_J^{(i)}}\right)^p \int_0^{\varepsilon V_J^{(i)}} 4p\; x^{2p-1} e^{-\frac{x^2}{2\kappa V_J^{(i)}(1+\varepsilon)}} dx \\
&\leq 4p\kappa^p\,(1+\varepsilon)^p \int_0^{+\infty} u^{2p-1} e^{-\frac{u^2}{2}} du \\
&\leq 4p\kappa^p\,(1+\varepsilon)^p \int_0^{+\infty} (2t)^{p-1}\, e^{-t} dt \\
&\leq 2^{p+1} p\kappa^p\,(1+\varepsilon)^p\, p!
\end{aligned}$$

We can then conclude by applying Bernstein's inequality (see for instance [32] with $v = 2^5 \left(\kappa(1+\varepsilon)\right)^2 |m|$ and $c = 4\left(\kappa(1+\varepsilon)\right)$):

$$\mathbf{P}\left[\chi^2(m,i)\mathbf{1}_{\Omega_{m_f}(\varepsilon)} \geq \kappa\left(|m| + 8(1+\varepsilon)\sqrt{x|m|} + 4(1+\varepsilon)x\right)\right] \quad \leq \quad e^{-x}.$$

While for a given i the $\{Z_J(i)\}_{J\in m}$ are independent variables, in general for a given J the variables $\{Z_J(i)\}_{1\leq i\leq d}$ are not. We conclude the proof using lemma B.1:

Lemma B.1. *Let $X_1,\ldots,X_n$ be real random variables and let $a_1,\ldots,a_n \in \mathbb{R}^n$ and $b_1,\ldots,b_n \in [0,1]^n$ such that $\forall\ 1 \leq i \leq n,\ \mathbf{P}(X_i \geq a_i) \leq b_i$. Then*

$$\mathbf{P}\left(\sum_{i=1}^n X_i \geq \sum_{i=1}^n a_i\right) \leq \sum_{i=1}^n b_i.$$

B.1.2. Link between χ^2_m, $V^2(\bar{s}_m,\hat{s}_m)$ and $K(\bar{s}_m,\hat{s}_m)$

By definition, we have

$$\begin{aligned} K(\bar{s}_m,\hat{s}_m) = \sum_{J\in m} |J| [A\left([\nabla A]^{-1}(\bar{\mathbf{T}}_J)\right) - A\left([\nabla A]^{-1}(\bar{\mathbf{E}}_J)\right) \\ - \bar{\mathbf{E}}_J.\left([\nabla A]^{-1}(\bar{\mathbf{T}}_J) - [\nabla A]^{-1}(\bar{\mathbf{E}}_J)\right)], \end{aligned}$$

and note that

$$V^2(\bar{s}_m,\hat{s}_m) = \sum_{i=1}^d \sum_{J\in m} V_J^{(i)} \left[[\nabla A]^{-1}(\bar{\mathbf{E}}_J)^{(i)} - [\nabla A]^{-1}(\bar{\mathbf{T}}_J)^{(i)}\right]^2. \tag{17}$$

Using Taylor development, and since A is m_ε-strongly convex on $K(\varepsilon)$, we have

$$\begin{aligned} & K(\bar{s}_m,\hat{s}_m)\mathbf{1}_{\Omega_{m_f}(\varepsilon)} \\ \geq \quad & \sum_{J\in m} |J| \frac{m_\varepsilon}{2} ||[\nabla A]^{-1}(\bar{\mathbf{T}}_J) - [\nabla A]^{-1}(\bar{\mathbf{E}}_J)||^2 \mathbf{1}_{\Omega_{m_f}(\varepsilon)} \\ \geq \quad & \sum_{i=1}^d \sum_{J\in m} |J| \frac{m_\varepsilon}{2V_J^{(i)}} V_J^{(i)} \left[[\nabla A]^{-1}(\bar{\mathbf{T}}_J)^{(i)} - [\nabla A]^{-1}(\bar{\mathbf{E}}_J)^{(i)}\right]^2 \mathbf{1}_{\Omega_{m_f}(\varepsilon)}. \end{aligned}$$

Hence,

$$K(\bar{s}_m,\hat{s}_m)\mathbf{1}_{\Omega_{m_f}(\varepsilon)} \geq \frac{m_\varepsilon}{2V^{max}} V^2(\bar{s}_m,\hat{s}_m)\ \mathbf{1}_{\Omega_{m_f}(\varepsilon)}. \tag{18}$$

Moreover, using the mean value inequality, we get

$$||[\nabla A]^{-1}(\bar{\mathbf{T}}_J) - [\nabla A]^{-1}(\bar{\mathbf{E}}_J)||^2 \geq \frac{1}{M_\varepsilon^2} ||\bar{\mathbf{T}}_J - \bar{\mathbf{E}}_J||^2,$$

and therefore

$$\begin{aligned} K(\bar{s}_m,\hat{s}_m)\mathbf{1}_{\Omega_{m_f}(\varepsilon)} &\geq \sum_{J\in m}|J|\frac{m_\varepsilon}{2M_\varepsilon^2}||\bar{\mathbf{T}}_J-\bar{\mathbf{E}}_J||^2\mathbf{1}_{\Omega_{m_f}(\varepsilon)} \\ &\geq \frac{V^{min}m_\varepsilon}{2M_\varepsilon^2}\chi_m^2\mathbf{1}_{\Omega_{m_f}(\varepsilon)}. \qquad (19) \end{aligned}$$

B.2. Proof of proposition 3.4

From (19), we have

$$\frac{V^{min}m_\varepsilon}{2M_\varepsilon^2}\left(\mathbf{E}[\chi_m^2]-\mathbf{E}[\chi_m^2\mathbf{1}_{\Omega_{m_f}(\varepsilon)^C}]\right)\leq \mathbf{E}[K(\bar{s}_m,\hat{s}_m)\mathbf{1}_{\Omega_{m_f}(\varepsilon)}].$$

Here we use again Cauchy-Schwarz inequality to control $\mathbf{E}[\chi_m^2\mathbf{1}_{\Omega_{m_f}(\varepsilon)^C}]$:

$$\begin{aligned} &\mathbf{E}[\chi_m^2\mathbf{1}_{\Omega_{m_f}(\varepsilon)^C}] \\ &\quad= \mathbf{E}\left[\sum_{i=1}^d\sum_{J\in m}\frac{(T_J^{(i)}-E_J^{(i)})^2}{V_J^{(i)}}\mathbf{1}_{\Omega_{m_f}(\varepsilon)^C}\right] \\ &\quad\leq \sum_{i=1}^d\left[\mathbf{E}\left(\sum_{J\in m}\frac{(T_J^{(i)}-E_J^{(i)})^2}{V_J^{(i)}}\right)^2\right]^{\frac{1}{2}}\mathbf{P}\Big(\Omega_{m_f}(\varepsilon)^C\Big)^{1/2} \\ &\quad\leq \sum_{i=1}^d\left[\mathbf{E}\left(\sum_{J\in m}\frac{(T_J^{(i)}-E_J^{(i)})^4}{{V_J^{(i)}}^2}+\sum_{J\in m}\sum_{J'\in m}\frac{(T_J^{(i)}-E_J^{(i)})^2}{V_J^{(i)}}\frac{(T_{J'}^{(i)}-E_{J'}^{(i)})^2}{V_{J'}^{(i)}}\right)\right]^{\frac{1}{2}} \\ &\qquad\times\mathbf{P}\Big(\Omega_{m_f}(\varepsilon)^C\Big)^{1/2} \\ &\quad\leq d\left[|m|\left(\frac{\sigma_4}{\Gamma^2{V^{min}}^2(\log n)^4}+|m|\right)\right]^{\frac{1}{2}}\mathbf{P}\Big(\Omega_{m_f}(\varepsilon)^C\Big)^{1/2}. \end{aligned}$$

And since by assumption $|J|>\Gamma(\log n)^2$, we have $|m|<\sqrt{n}/\Gamma$ as soon as $n>4$, and finally,

$$\mathbf{E}[\chi_m^2\mathbf{1}_{\Omega_{m_f}(\varepsilon)^C}] \quad\leq\quad d|m|\frac{C(a,\Gamma,V^{min},\varepsilon,\kappa,\sigma_4)}{n^{a/2}}\leq\frac{C(a,\Gamma,V^{min},\varepsilon,\kappa,d,\sigma_4)}{n^{(a-1)/2}}.$$

Now using $\mathbf{E}[\chi_m^2]=d|m|$,

$$\frac{V^{min}m_\varepsilon}{2M_\varepsilon^2}\ d|m|-\frac{C(a,\Gamma,V^{min},\varepsilon,\kappa,d,\sigma_4,m_\varepsilon,M_\varepsilon)}{n^{(a-1)/2}}\leq\mathbf{E}[K(\bar{s}_m,\hat{s}_m)\mathbf{1}_{\Omega_{m_f}(\varepsilon)}].$$

Finally, introducing $C(\varepsilon)=\dfrac{V^{min}m_\varepsilon}{2M_\varepsilon^2}$, and since $\mathbf{E}\left[K(\bar{s}_m,\hat{s}_m)\mathbf{1}_{\Omega_{m_f}(\varepsilon)^C}\right]\geq 0$, we have

$$K(s,\bar{s}_m)+C(\varepsilon)\ d|m|-\frac{C(a,\Gamma,V^{min},\varepsilon,\kappa,d,\sigma_4,m_\varepsilon,M_\varepsilon)}{n^{(a-1)/2}}\leq\mathbf{E}[K(s,\hat{s}_m)].$$

B.3. Proof of proposition 4.2

We recall that

$$\begin{aligned}\overline{\gamma_n}(\bar{s}_{m'}) &= \sum_{J\in m'}\left[-[\nabla A]^{-1}(\bar{\mathbf{E}}_J).(\mathbf{T}_J-\mathbf{E}_J)\right],\\ \overline{\gamma_n}(\hat{s}_{m'}) &= \sum_{J\in m'}\left[-\nabla A^{-1}(\bar{\mathbf{T}}_J).(\mathbf{T}_J-\mathbf{E}_J)\right].\end{aligned}$$

We therefore have

$$\begin{aligned}\overline{\gamma_n}(\bar{s}_{m'})-\overline{\gamma_n}(\hat{s}_{m'}) &= \sum_{J\in m'}(\mathbf{T}_J-\mathbf{E}_J).\left([\nabla A]^{-1}(\bar{\mathbf{T}}_J)-[\nabla A]^{-1}(\bar{\mathbf{E}}_J)\right)\\ &= \sum_{i=1}^{d}\sum_{J\in m'}(T_J^{(i)}-E_J^{(i)})\left(\nabla A^{-1}(\bar{\mathbf{T}}_J)^{(i)}-[\nabla A]^{-1}(\bar{\mathbf{E}}_J)^{(i)}\right).\end{aligned}$$

Using Cauchy-Schwarz inequality,

$$\begin{aligned}&\overline{\gamma_n}(\bar{s}_{m'})-\overline{\gamma_n}(\hat{s}_{m'})\\ &\quad=\sum_{i=1}^{d}\sum_{J\in m'}\left[\frac{(T_J^{(i)}-E_J^{(i)})}{\sqrt{V_J^{(i)}}}\sqrt{V_J^{(i)}}\left([\nabla A]^{-1}(\bar{\mathbf{T}}_J)^{(i)}-[\nabla A]^{-1}(\bar{\mathbf{E}}_J)^{(i)}\right)\right]\\ &\quad\le\sqrt{\sum_{i=1}^{d}\sum_{J\in m'}\frac{(T_J^{(i)}-E_J^{(i)})^2}{V_J^{(i)}}}\\ &\qquad\times\sqrt{\sum_{i=1}^{d}\sum_{J\in m'}V_J^{(i)}\left[[\nabla A]^{-1}(\bar{\mathbf{T}}_J)^{(i)}-[\nabla A]^{-1}(\bar{\mathbf{E}}_J)^{(i)}\right]^2}\\ &\quad\le\sqrt{\chi^2_{m'}}\sqrt{V^2(\bar{s}_{m'},\hat{s}_{m'})},\end{aligned}$$

where χ^2_m and $V^2(\bar{s}_m,\hat{s}_m)$ have been defined in equation (12) and (17) respectively. Then from equation (18) and $2ab\le xa^2+x^{-1}b^2$, we get on $\Omega_{m_f}(\varepsilon)$

$$\overline{\gamma_n}(\bar{s}_{m'})-\overline{\gamma_n}(\hat{s}_{m'}) \quad\le\quad \frac{1+\varepsilon}{2m_\varepsilon}V^{max}\chi^2_{m'}+\frac{1}{1+\varepsilon}K(\bar{s}_{m'},\hat{s}_{m'}).$$

Finally, using proposition 4.1,

$$\begin{aligned}&(\overline{\gamma_n}(\bar{s}_{m'})-\overline{\gamma_n}(\hat{s}_{m'}))\mathbf{1}_{\Omega_{m_f}(\varepsilon)\cap\Omega_1(\xi)}\\ &\qquad\le\frac{d\kappa(1+\varepsilon)}{2m_\varepsilon}V^{max}\left[|m'|+8(1+\varepsilon)\sqrt{(L_{m'}|m'|+\xi)|m'|}\right.\\ &\qquad\qquad\left.+\,4(1+\varepsilon)(L_{m'}|m'|+\xi)\right]+\frac{1}{1+\varepsilon}K(\bar{s}_{m'},\hat{s}_{m'}).\end{aligned}$$

B.4. Proof of proposition 4.3

We recall that

$$\overline{\gamma_n}(\bar{s}_m) - \overline{\gamma_n}(s) = \sum_{J\in m}\sum_{t\in J}(\boldsymbol{\theta}_t - [\nabla A]^{-1}(\bar{\mathbf{E}}_J)).(\mathbf{T}_t - \mathbf{E}_t).$$

Then

$$\begin{aligned}
&|\mathbb{E}[(\overline{\gamma_n}(\bar{s}_m) - \overline{\gamma_n}(s))\mathbf{1}_{\Omega_{m_f}(\varepsilon)}| \\
&\qquad \leq \mathbb{E}\left[|\overline{\gamma_n}(\bar{s}_m) - \overline{\gamma_n}(s)|\,\mathbf{1}_{\Omega_{m_f}(\varepsilon)^C}\right] \\
&\qquad \leq \mathbb{E}\left[\left|\sum_{J\in m}\sum_{t\in J}\sum_{i=1}^{d}(\theta_t^{(i)} - [\nabla A]^{-1}(\bar{\mathbf{E}}_J)^{(i)})(T_t^{(i)} - E_t^{(i)})\right|\mathbf{1}_{\Omega_{m_f}(\varepsilon)^C}\right] \\
&\qquad \leq \nu\sum_{i=1}^{d}\left[\mathbb{E}\left(\sum_{t=1}^{n}(T_t^{(i)} - E_t^{(i)})\right)^2\right]^{1/2}\mathbf{P}\Big(\Omega_{m_f}(\varepsilon)^C\Big)^{1/2} \\
&\qquad \leq \nu\sum_{i=1}^{d}\left[\sum_{t=1}^{n}V_t^{(i)}\right]^{1/2}\mathbf{P}\Big(\Omega_{m_f}(\varepsilon)^C\Big)^{1/2} \\
&\qquad \leq d\nu\frac{\sqrt{V^{max}C(a,\varepsilon,\Gamma,\kappa,V^{min})}}{n^{(a-1)/2}}.
\end{aligned}$$

B.5. Proof of proposition 5.3

We recall that

$$\overline{\gamma_n}(s) - \overline{\gamma_n}(\bar{s}_{m'}) = \sum_{J\in m'}\sum_{t\in J}(\boldsymbol{\theta}_t - [\nabla A]^{-1}(\bar{\mathbf{E}}_J)).(\mathbf{E}_t - \mathbf{T}_t) = \sum_{J\in m'}\sum_{t\in J}\sum_{i=1}^{d}X_t^{(i)},$$

with $X_t^{(i)} = (\boldsymbol{\theta}_t^{(i)} - [\nabla A]^{-1}(\bar{\mathbf{E}}_J)^{(i)})(E_t^{(i)} - T_t^{(i)})$. Then on $\Omega_{m_f}(\varepsilon)$,

$$(\overline{\gamma_n}(s) - \overline{\gamma_n}(\bar{s}_{m'}))\,\mathbf{1}_{\Omega_{m_f}(\varepsilon)} = \sum_{J\in m'}\sum_{t\in J}\sum_{i=1}^{d}W_t^{(i)},$$

with $W_t^{(i)} = X_t^{(i)}\mathbf{1}_{\Omega_{m_f}(\varepsilon)}$. Noticing that we have,

- on $\Omega_{m_f}(\varepsilon)$, $|E_t^{(i)} - T_t^{(i)}| \leq \varepsilon V_t^{(i)} \leq \varepsilon V^{max}$
- and on $\boldsymbol{\nu}$, $|\boldsymbol{\theta}_t^{(i)} - [\nabla A]^{-1}(\bar{\mathbf{E}}_J)^{(i)}| \leq \nu$

we have $|W_t^{(i)}| \leq \nu\varepsilon V^{max}$. Moreover

$$\sum_{J,t}\mathbb{E}\left[(W_t^{(i)})^2\right] \quad = \quad \sum_{J,t}\mathbb{E}\left[(\boldsymbol{\theta}_t^{(i)} - [\nabla A]^{-1}(\bar{\mathbf{E}}_J)^{(i)})^2(E_t^{(i)} - T_t^{(i)})^2\mathbf{1}_{\Omega_{m_f}(\varepsilon)}\right]$$

$$\begin{aligned}
&\leq \sum_{J,t} (\boldsymbol{\theta}_t^{(i)} - [\nabla A]^{-1}(\bar{\mathbf{E}}_J)^{(i)})^2 \mathbb{E}\left[(E_t^{(i)} - T_t^{(i)})^2\right] \\
&\leq \sum_{J,t} (\boldsymbol{\theta}_t^{(i)} - [\nabla A]^{-1}(\bar{\mathbf{E}}_J)^{(i)})^2 V_t^{(i)} = V_i{}^2(s, \bar{s}_{m'}),
\end{aligned}$$

where $V_i{}^2(s,u)$ is defined in equation (16). Applying Bernstein, we have $\forall i$

$$\mathbf{P}\left[\sum_{J\in m'}\sum_{t\in J} W_t^{(i)} \mathbf{1}_{\Omega_{m_f}(\varepsilon)} \geq V_i(s,\bar{s}_{m'})\sqrt{2x} + \frac{\nu\varepsilon V^{max} x}{3}\right] \leq e^{-x}$$

Then using lemma B.1,

$$\mathbf{P}\left[\left(\overline{\gamma_n}(s) - \overline{\gamma_n}(\bar{s}_{m'})\right)\mathbf{1}_{\Omega_{m_f}(\varepsilon)} \geq \sum_i V_i(s,\bar{s}_{m'})\sqrt{2x} + \frac{dx\nu\varepsilon V^{max}}{3}\right] \leq de^{-x}. \tag{20}$$

Then we conclude the proof using respectively:

- using Cauchy-Schwarz inequality, we get

$$\sum_i V_i(s,\bar{s}_{m'})\sqrt{2x_{m'}} \leq \sqrt{V^2(s,\bar{s}_{m'})}\sqrt{2dx_{m'}};$$

- since A is m_ν-strongly convex on $\boldsymbol{\nu}$, we have $K(s,\bar{s}_m) \geq \frac{m_\nu}{2V^{max}} V^2(s,\bar{s}_m)$;
- and finally setting $x_{m'} = L_{m'}|m'| + \xi$ and using $2ab \leq xa^2 + x^{-1}b^2$ with $x = \frac{1}{1+\varepsilon}$.

Appendix C: Proof of Theorem 6.1

We obtain the result of Theorem 6.1 by following the same lines of the proof of Theorem 3.2. The main constant of interest is C_{λ_ε} which comes from the control of the term $\overline{\gamma_n}(\bar{s}_{m'}) - \overline{\gamma_n}(\hat{s}_{m'})$ on a particular set $\Omega_{m_f}(\varepsilon)$ (the controls of the two other terms $\overline{\gamma_n}(\bar{s}_m) - \overline{\gamma_n}(s)$ and $\overline{\gamma_n}(s) - \overline{\gamma_n}(\bar{s}_{m'})$ appeared in the negligible constant of the risk). Here again this control is obtained using two results: the control of a chi-square statistic χ_m^2 and the control $K(\bar{s}_{m'}, \hat{s}_{m'})$ through a quantity $V^2(\bar{s}_{m'}, \hat{s}_{m'})$, denoted here $V_{m'}^2$, on $\Omega_{m_f}(\varepsilon)$. This latter set is taken here as

$$\Omega_{m_f}(\varepsilon) \quad = \quad \bigcap_{J\in m}\bigcap_{i=1}^{r}\left\{\left|N_J(i) - \sum_{t\in J} s(t,i)\right| \leq \varepsilon \sum_{t\in J} s(t,i)(1-s(t,i))\right\}.$$

This quantity is defined in the same manner than in our general approach (see (4)), as we can recover $E_J^{(i)} = \sum_{t\in J} s(t,i)$, $V_J^{(i)} = \sum_{t\in J} s(t,i)(1-s(t,i))$ and $T_J^{(i)} = N_J(i)$. The main difference is that the dimension differs since $r = d+1$ terms are considered. Using the control of $N_J(i)$, that can be obtained through a direct application of the bounded version of Bernstein since $\overline{\mathbf{1}_{\{Y_t=i\}}} = \mathbf{1}_{\{Y_t=i\}} - s(t,i)$ is bounded by 1,

$$P\left[\left|N_J(i) - \sum_{t\in J} s(t,i)\right| \geq x\right] \leq 2\exp\left(-\frac{x^2}{2\left[\frac{x}{3} + \sum_{t\in J} s\left(t,i\right)\left(1-s(t,i)\right)\right]}\right), \tag{21}$$

we have, with $a > 1$,

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

C.1. Control of the term $\overline{\gamma_n}\left(\bar{s}_{m'}\right) - \overline{\gamma_n}\left(\hat{s}_{m'}\right)$

We introduce

$$\chi_m^2 = \sum_{i=1}^r \chi^2(m,i) = \sum_{i=1}^r \sum_{J\in m} Z_J(i),$$

with $Z_J(i) = \frac{\overline{N_J(i)}^2}{\sum_{t\in J} s(t,i)}$ and $\overline{N_J\left(i\right)} = \sum_{t\in J}\left(\mathbf{1}_{\{Y_t=i\}} - s(t,i)\right)$. We control the moment of $Z_J(i)$ (as in Section 4) using (21). Then since $\mathbb{E}\left[\chi_m^2\right] = r|m|$, by using Bernstein's inequality again, we conclude to

$$P\left(\sum_{i=1}^r \chi^2(m,i)\mathbf{1}_{\Omega_{m_f}(\varepsilon)} \geq r|m| + 8r\left(1+\frac{\varepsilon}{3}\right)\sqrt{x|m|} + 4r\left(1+\frac{\varepsilon}{3}\right)x\right)$$
$$\leq r\exp\left(-x\right).$$

We let denote the set

$$\Omega_1(\xi) = \bigcap_{m'\in\mathcal{M}_n}\{\chi_{m'}^2\mathbf{1}_{\Omega_{m_f}(\varepsilon)} \leq r[|m'| + 8\left(1+\frac{\varepsilon}{3}\right)\sqrt{(L_{m'}|m'|+\xi)|m'|}$$
$$+ 4\left(1+\frac{\varepsilon}{3}\right)(L_{m'}|m'|+\xi)]\},$$

then we can control $\overline{\gamma_n}(\bar{s}_{m'}) - \overline{\gamma_n}(\hat{s}_{m'})$ with the following proposition:

Proposition C.1.

$$(\overline{\gamma_n}(\bar{s}_{m'}) - \overline{\gamma_n}(\hat{s}_{m'}))\mathbf{1}_{\Omega_{m_f}(\varepsilon)\cap\Omega_1(\xi)}$$
$$\leq \frac{1}{2}\left(\frac{1+\varepsilon}{1-\varepsilon}\right) r\left[|m'| + 8\left(1+\frac{\varepsilon}{3}\right)\sqrt{(L_{m'}|m'|+\xi)|m'|}\right.$$
$$\left. + \ 4\left(1+\frac{\varepsilon}{3}\right)(L_{m'}|m'|+\xi)\right] + \frac{1}{1+\varepsilon}K(\bar{s}_{m'},\hat{s}_{m'}).$$

Proof: Using Cauchy-Schwarz inequality, we have

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

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

with

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

Applying Lemma 2.3 of [14], we get on $\Omega_{m_f}(\varepsilon)$,

$$\begin{aligned} \frac{1-\varepsilon}{2} V^2_{m'} &\leq K(\bar{s}_{m'}, \hat{s}_{m'}) \leq \frac{1+\varepsilon}{2} V^2_{m'}, \quad \text{and} \\ \overline{\gamma_n}(\bar{s}_{m'}) - \overline{\gamma_n}(\hat{s}_{m'}) &\leq \frac{1}{2}\left(\frac{1+\varepsilon}{1-\varepsilon}\right) \chi^2_m + \frac{1}{1+\varepsilon} K(\bar{s}_{m'}, \hat{s}_{m'}). \end{aligned}$$

C.2. Control of the terms $\overline{\gamma_n}(\bar{s}_m) - \overline{\gamma_n}(s)$ and $\overline{\gamma_n}(s) - \overline{\gamma_n}(\bar{s}_{m'})$

The term $\overline{\gamma_n}(s) - \overline{\gamma_n}(\bar{s}_{m'})$ is controlled using the same proposition as in our general case (see proposition 4.4) with dimension r instead of d and the term $\overline{\gamma_n}(\bar{s}_m) - \overline{\gamma_n}(s)$ is controlled by bounding its expectation as follows:

$$\begin{aligned} \left| \mathbb{E}\left[\left(\overline{\gamma_n}(\bar{s}_m) - \overline{\gamma_n}(s) \right) \mathbf{1}_{\Omega_{m_f}(\varepsilon)} \right] \right| &\leq \left| \mathbb{E}\left[\left(\overline{\gamma_n}(\bar{s}_m) - \overline{\gamma_n}(s) \right) \mathbf{1}_{\Omega_{m_f}(\varepsilon)^C} \right] \right| \\ &\leq nr \log\left(\frac{1}{\rho}\right) P\left(\Omega_{m_f}(\varepsilon)^C \right) \\ &\leq \frac{C(\Gamma, \rho, a, r, \varepsilon)}{n^{a-1}} \end{aligned}$$

C.3. Proof of Theorem 6.1

We consider sets of large probability, $\Omega_1(\xi)$ (defined just above) and $\Omega_2(\xi)$ (that is the same as in our general case (see (13)) but with r instead of d) and gather the previous results on the control of each terms of the main decomposition as for the proof of Theorem 3.2 given in Section 4.4.

Acknowledgments

The authors would like to thank Elodie Nédélec for her help in the categorical study, and Mahendra Mariadassou and Hoel Queffelec for helpful discussions on exponential families and their geometric properties.

References

[1] Akaike, H. (1973). Information theory and extension of the maximum likelihood principle. *Second international symposium on information theory*, 267–281. MR0483125

[2] Arlot, S., Celisse, A., and Harchaoui, Z. (2012). A kernel multiple change-point algorithm via model selection. *arXiv preprint arXiv: 1202.3878*.

[3] Arlot, S. and Massart, P. (2009). Data-driven calibration of penalties for least-squares regression. *The Journal of Machine Learning Research 10*, 245–279.

[4] Barron, A., Birgé, L., and Massart, P. (1999). Risk bounds for model selection via penalization. *Probability Theory Related Fields* **113**, 3, 301–413. MR1679028

[5] Bellman, R. (1961). On the approximation of curves by line segments using dynamic programming. *Commun. ACM* **4**, 6, 284. http://portal.acm.org/citation.cfm?id=366611.

[6] Birgé, L. and Massart, P. (1997). From model selection to adaptive estimation. In *Festschrift for Lucien Le Cam*. Springer, New York, 55–87.

[7] Birgé, L. and Massart, P. (2001). Gaussian model selection. *Journal of the European Mathematical Society* **3**, 3, 203–268. MR1848946

[8] Birgé, L. and Massart, P. (2007). Minimal penalties for Gaussian model selection. *Probability Theory Related Fields* **138**, 1–2, 33–73. MR2288064

[9] Boys, R. J. and Henderson, D. A. (2004). A bayseian approach to DNA sequence segmentation. *Biometrics* **60**, 2, 573–588.

[10] Braun, J. V., Braun, R., and Müller, H.-G. (2000). Multiple change-point fitting via quasilikelihood, with application to dna sequence segmentation. *Biometrika* **87**, 2, 301–314. MR1782480

[11] Braun, J. V. and Müller, H.-G. (1998). Statistical methods for DNA sequence segmentation. *Biometrika* **13**, 2, 301–314.

[12] Breiman, Friedman, Olshen, and Stone. (1984). Classification and regression trees. *Wadsworth and Brooks*. MR0726392

[13] Brown, L. D. (1986). Fundamentals of statistical exponential families with applications in statistical decision theory. *Lecture Notes-monograph series*, i–279.

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

[15] Cleynen, A., Dudoit, S., and Robin, S. (2014). Comparing segmentation methods for genome annotation based on rna-seq data. *Journal of Agricultural, Biological, and Environmental Statistics* **19**, 1, 101–118. MR3257904

[16] Cleynen, A., Koskas, M., Lebarbier, E., Rigaill, G., and Robin, S. (2014). Segmentor3isback: an R package for the fast and exact segmentation of seq-data. *Algorithms for Molecular Biology 9*, 6.

[17] Cleynen, A. and Lebarbier, E. (2014). Segmentation of the Poisson and negative binomial rate models: a penalized estimator. *ESAIM: Probability and Statistics*. MR3334013

[18] Cleynen, A., Luong, T. M., Rigaill, G., and Nuel, G. (2014). Fast estimation of the integrated completed likelihood criterion for change-point detection problems with applications to next-generation sequencing data. *Signal Processing 98*, 233–242.

[19] Cleynen, A. and Robin, S. (2016). Comparing change-point location in independent series. *Statistics and Computing* **26**, 1–2, 263–276. MR3439372

[20] Durot, C., Lebarbier, E., and Tocquet, A. (2009). Estimating the joint distribution of independent categorical variables via model selection. *Bernoulli* **15**, 2, 475–507.

[21] Frick, K., Munk, A., and Sieling, H. (2014). Multiscale change point inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **76**, 3, 495–580. MR3210728

[22] Gassiat, E., Cleynen, A., and Robin, S. (2016). Inference in finite state space non parametric hidden Markov models and applications. *Statistics and Computing* **26**, 1–2, 61–71. MR3439359

[23] Harchaoui, Z. and Lévy-Leduc, C. (2010). Multiple change-point estimation with a total variation penalty. *Journal of the American Statistical Association* **105**, 492. MR2796565

[24] Hughes, N. P., Tarassenko, L., and Roberts, S. J. (2003). Markov models for automated ECG interval analysis. *Advances in Neural Information Processing Systems 16*.

[25] Johnson, N., Kemp, A., and Kotz, S. (2005). Univariate discrete distributions. *John Wiley & Sons, Inc.*.

[26] Kakade, S. M., Shamir, O., Sridharan, K., and Tewari, A. (2009). Learning exponential families in high-dimensions: Strong convexity and sparsity. *arXiv preprint arXiv:0911.0054*.

[27] Killick, R., Fearnhead, P., and Eckley, I. (2012). Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association* **107**, 500, 1590–1598. MR3036418

[28] Lai, W. R., Johnson, M. D., Kucherlapati, R., and Park, P. J. (2005). Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. *Bioinformatics* **21**, 19, 3763–3770.

[29] Lebarbier, E. (2005). Detecting multiple change-points in the mean of Gaussian process by model selection. *Signal Processing* **85**, 4 (Apr.), 717–736.

[30] Lee, J. D., Sun, Y., and Taylor, J. E. (2013). On model selection consistency of m-estimators with geometrically decomposable penalties. *Advances in Neural Processing Information Systems*.

[31] Maidstone, R., Hocking, T., Rigaill, G., and Fearnhead, P. (2016). On optimal multiple changepoint algorithms for large data. *Statistics and Computing*, 1–15. http://dx.doi.org/10.1007/s11222-016-9636-3. MR3599687

[32] Massart, P. (2007). *Concentration inequalities and model selection*. Springer Verlag. MR2319879

[33] Matteson, D. S. and James, N. A. (2014). A nonparametric approach for multiple change point analysis of multivariate data. *Journal of the American Statistical Association* **109**, 505, 334–345. MR3180567

[34] Muri, F. (1998). Modelling bacterial genomes using hidden Markov mod-

els. *Compstat98. Proceedings in Computational Statistics, Eds R. Payne and P. Green*, 89–100.

[35] Rigaill, G. (2010). Pruned dynamic programming for optimal multiple change-point detection. *Arxiv:1004.0887*. http://arxiv.org/abs/1004.0887.

[36] Rigaill, G., Lebarbier, E., and Robin, S. (2012). Exact posterior distributions and model selection criteria for multiple change-point detection problems. *Statistics and Computing* **22**, 4, 917–929.

[37] Wainwright, M. J. and Jordan, M. I. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning* **1**, 1–2, 1–305.

[38] Yao, Y.-C. (1988). Estimating the number of change-points via Schwarz' criterion. *Statistics & Probability Letters* **6**, 3 (February), 181–189.

[39] Zhang, N. R. and Siegmund, D. O. (2007). A modified Bayes information criterion with applications to the analysis of comparative genomic hybridization data. *Biometrics* **63**, 1, 22–32.