ESAIM: PS 18 (2014) 750–769
DOI: 10.1051/ps/2014005

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

SEGMENTATION OF THE POISSON AND NEGATIVE BINOMIAL RATE MODELS: A PENALIZED ESTIMATOR

Alice Cleynen[1] and Emilie Lebarbier[2]

Abstract. We consider the segmentation problem of Poisson and negative binomial (*i.e.* overdispersed Poisson) rate distributions. In segmentation, an important issue remains the choice of the number of segments. To this end, we propose a penalized log-likelihood estimator where the penalty function is constructed in a non-asymptotic context following the works of L. Birgé and P. Massart. The resulting estimator is proved to satisfy an oracle inequality. The performances of our criterion is assessed using simulated and real datasets in the RNA-seq data analysis context.

Mathematics Subject Classification. 62G05, 62G07, 62P10.

Received February 21, 2013. Revised December 16, 2013.

1. Introduction

We consider a multiple change-point detection setting for count datasets, which can be written as follows: we observe a finite sequence $\{y_t\}_{t\in\{1,\dots,n\}}$ realisation of independent variables Y_t. These variables are supposed to be drawn from a probability distribution $\mathcal{G}$ which depends on a set of parameters. Here two types of parameters are distinguished:

$$Y_t \sim \mathcal{G}(\theta_t, \phi) = s(t),\ 1 \leq t \leq n,$$

where ϕ is a constant parameter while the θs are point-specific. In many contexts, we might want to consider that the θs are piece-wise constant and so subject to an unknown number $K-1$ of abrupt changes (for instance with climatic or financial data). Thus, we want to assume the existence of partition of $\{1,\dots,n\}$ into K segments within which the observations follow the same distribution and between which observations have different distributions, *i.e.* θ is constant within a segment and differ from a segment to another. A motivating example is sequencing data analysis. For instance, the output of RNA-seq experiments is the number of reads (*i.e.* short portions of the genome) which first position maps to each location of a genome of reference. Supposing that we dispose of such a sequence, we expect to observe a stationarity in the amount of reads falling in different areas of the genome: expressed genes, intronic regions, *etc.* We wish to localize those regions that are biologically significant. In our context, we consider for $\mathcal{G}$ the Poisson and negative binomial distributions, adapted to RNA-seq experiment analysis [24].

Keywords and phrases. Distribution estimation, change-point detection, count data (RNA-seq), poisson and negative binomial distributions, model selection.

[1] AgroParisTech UMR518, Paris 5e, France. alice.cleynen@agroparistech.fr

[2] INRA UMR518, Paris 5e, France. emilie.lebarbier@agroparistech.fr

Article published by EDP Sciences

© EDP Sciences, SMAI 2014

https://doi.org/10.1051/ps/2014005 Published online by Cambridge University Press

Change-point detection problems are not new and many methods have been proposed in the literature. For count data-sets, [12] provide a detailed bibliography of methods in the particular case of the segmentation of the DNA sequences that includes Bayesian approaches, scan statistics, likelihood-ratio tests, binary segmentation and numerous other methods such as penalized contrast estimation procedures. In a Bayesian framework, [6] proposes to use an exact "ICL" criterion for the choice of K, while its approximation is computed in the constrained HMM approach of [19]. In this paper, we consider a penalized contrast estimation method which consists first, for every fixed K, in finding the best segmentation in K segments by minimizing the contrast over all the partitions with K segments, and then in selecting a convenient number of segments K by penalizing the contrast. Choosing the number of segments, *i.e.* choosing a "good" penalty, is a crucial issue and not so easy. The most basic examples of penalty are the Akaike Information Criterion (AIC [1]) and the Bayes Information Criterion (BIC [25]) but these criteria are not well adapted in the segmentation context and tend to overestimate the number of change-points (see [10,26] for theoretical explanations). In this particular context, some modified versions of these criteria have been proposed. For instance, [11,26] have proposed modified versions of the BIC criterion (shown to be consistent) in the segmentation of Gaussian processes and DNA sequences respectively. However, these criteria are based on asymptotic considerations. In the last years there has been an extensive literature influenced by [5,8] introducing non-asymptotic model selection procedures, in the sense that the size of the models as well as the size of the list of models are allowed to be large when n is large. 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. This approach has been now considered in various function estimation contexts. In particular, [2] proposed a penalty for estimating the density of independent categorical variables in a least-squares framework, while [7,21], or [4], focused on the estimation of the density of a Poisson process.

When the number of models is large, as in the case of an exhaustive search in segmentation problem, it can be shown that penalties which only depend on the number of parameters of each model, as for the classical criteria, are theoretically (and also practically) not adapted. This was suggested by [10, 18] who show that the penalty term needs to be well defined, and in particular needs to depend on the complexity of the list of models, *i.e.* the number of models having the same dimension. For this reason, following the work of [8] and in particular [14] in the density estimation framework, we consider a penalized log-likelihood procedure to estimate the true distribution s of a Poisson or negative binomial-distributed sequence $\mathbf{y}$. We prove that, up to a $\log n$ factor, the resulting estimator satisfies an oracle inequality. In our framework, instead of considering n independent identically distributed variables drawn from a distribution s (to be estimated) as in the histogram density estimation context [14], we have only one observation with size n, the key point being the piece-wise constant nature of this distribution which allows repetitions on each piece. Still, in this case, s depends on n, (to be more precise the support of the function s depends on n), meaning that when n increases, the support of our function to be estimated also increases.

The paper is organized as follows. The general framework is described in Section 2. More precisely, we present our proposed penalized maximum-likelihood estimator, the form of the penalty and give some non-asymptotic risk bounds for the resulting estimator. The studies of the two considered models (Poisson and negative binomial) are done in parallel along the paper. Some exponential bounds are derived in Section 3. A simulation study is performed to compare our proposed criterion with others and an application to the segmentation of RNA-seq data illustrates the procedure in Section 4. The proof of the main result is given in Section 5 for which the proofs of some intermediate results are given in the Appendix A.

2. Model selection procedure

2.1. Penalized maximum-likelihood estimator

Let us denote by m a partition of $[\![1, n]\!]$, $m = \{[\![1, \tau_1[\![, [\![\tau_1, \tau_2[\![, \ldots, [\![\tau_k, n]\!]\}$ and by $\mathcal{M}_n$ a set of partitions of $[\![1, n]\!]$. In our framework we want to estimate the distribution s defined by $s(t) = \mathcal{G}(\theta_t, \phi)$, $1 \leq t \leq n$, and we

https://doi.org/10.1051/ps/2014005 Published online by Cambridge University Press

consider the two following models:

$$\begin{array}{ll} \mathcal{G}(\theta_t, \phi) = \mathcal{P}(\lambda_t) & (\mathcal{P}) \\ \mathcal{G}(\theta_t, \phi) = \mathcal{NB}(p_t, \phi) & (\mathcal{NB}). \end{array}$$

In the $(\mathcal{NB})$ case, we suppose that the over-dispersion parameter ϕ is known. We define the collection of models:

Definition 2.1. The collection of models associated to partition m is $\mathcal{S}_m$ the set of distribution of sequences of length n such that for each element s_m of $\mathcal{S}_m$, for each segment J of m, and for each t in J, $s_m(t) = \mathcal{G}(\theta_J, \phi)$:

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

We shall denote by $|m|$ the number of segments in partition m, and by $|J|$ the length of segment J.

We consider the log-likelihood contrast $\gamma(u) = \sum_{t=1}^{n} -\log \mathbf{P}_u(Y_t)$, namely respectively for $u(t) = \mathcal{P}(\mu_t)$ and $u(t) = \mathcal{NB}(q_t, \phi)$,

$$\begin{array}{ll} \gamma(u) = \sum_{t=1}^{n} \mu_t - Y_t \log(\mu_t) + \log(Y_t!), & (\mathcal{P}) \\ \gamma(u) = \sum_{t=1}^{n} -\phi \log q_t - Y_t \log(1-q_t) - \log\left(\frac{\Gamma(\phi+Y_t)}{\Gamma(\phi) Y_t!}\right). & (\mathcal{NB}) \end{array}$$

Then the minimal contrast estimator $\hat{s}_m$ of s on the collection $\mathcal{S}_m$ is

$$\hat{s}_m = \arg\min_{u \in \mathcal{S}_m} \gamma(u), \tag{2.1}$$

so that, noting $\bar{Y}_J = \dfrac{\sum_{t \in J} Y_t}{|J|}$, for all $J \in m$ and $t \in J$

$$\hat{s}_m(t) = \mathcal{P}(\bar{Y}_J) \text{ for } (\mathcal{P}) \quad \text{and } \hat{s}_m(t) = \mathcal{NB}\left(\frac{\phi}{\phi + \bar{Y}_J}, \phi\right) \text{ for } (\mathcal{NB}). \tag{2.2}$$

Therefore, for each partition m of $\mathcal{M}_n$ we can obtain the best estimator $\hat{s}_m$ as in equation (2.2), and thus define a collection of estimators $\{(\hat{s}_m)_{m \in \mathcal{M}_n}\}$. Ideally, we would wish to select the estimator $\hat{s}_{m(s)}$ amongst this collection with the minimum given risk. In the log-likelihood framework, it is natural to consider the Kullback−Leibler risk, with $K(s,u) = \mathbf{E}[\gamma(u) - \gamma(s)]$. In the following we note $\mathbf{E}$ and $\mathbf{P}$ the expectation and the probability under the true distribution s respectively (otherwise the underlying distribution is mentioned). In our models, the Kullback−Leibler between distributions s and u can be developed into

$$\begin{array}{ll} K(s,u) = \sum_{t=1}^{n} \left(\mu_t - \lambda_t - \lambda_t \log \dfrac{\mu_t}{\lambda_t}\right), & (\mathcal{P}) \\ K(s,u) = \phi \sum_{t=1}^{n} \log\left(\dfrac{p_t}{q_t}\right) + \dfrac{1-p_t}{p_t} \log\left(\dfrac{1-p_t}{1-q_t}\right). & (\mathcal{NB}) \end{array}$$

Unfortunately, minimizing this risk requires the knowledge of the true distribution s, and is unreachable. We will therefore want to consider the estimator $\hat{s}_{\hat{m}}$ where $\hat{m}$ minimizes $\gamma(\hat{s}_m) + pen(m)$ for a well-chosen function pen (depending on the data). By doing so, we hope to select an estimator $\hat{s}_{\hat{m}}$ whose risk is as close as possible to the risk of $\hat{s}_{m(s)} = \arg\min_{m \in \mathcal{M}_n} \mathbf{E}_s[K(s, \hat{s}_m)]$ in the sense that

$$\mathbf{E}[K(s, \hat{s}_{\hat{m}})] \leq C\, \mathbf{E}\left[K\left(s, \hat{s}_{m(s)}\right)\right],$$

https://doi.org/10.1051/ps/2014005 Published online by Cambridge University Press

where C is a nonnegative constant hopefully close to 1. We therefore introduce the following definition:

Definition 2.2. Let $\mathcal{M}_n$ be a collection of partitions of $[\![1, n]\!]$ constructed on a partition m_f (*i.e.* m_f is a refinement of every m in $\mathcal{M}_n$). Given a nonnegative, increasing in the size of m penalty function pen: $\mathcal{M}_n \to \mathbf{R}_+$, and choosing

$$\hat{m} = \arg \min_{m \in \mathcal{M}_n} \{\gamma(\hat{s}_m) + pen(m)\},$$

we define the penalized maximum-likelihood estimator as $\hat{s}_{\hat{m}}$.

In the following Section we provide a choice of penalty function, and show that the resulting estimator satisfies an oracle inequality.

2.2. Choice of the penalty function

Main result

The following result shows that for an appropriate choice of the penalty function, we have a non-asymptotic risk bound for the penalized maximum-likelihood estimator.

Theorem 2.3. *Let $\mathcal{M}_n$ be a collection of partitions constructed on a partition m_f such that there exist absolute positive constants $\rho_{\min}$, $\rho_{\max}$ and Γ satisfying:*

- $\forall t, \rho_{\min} \leq \theta_t \leq \rho_{\max}$ *and*
- $\forall J \in m_f, |J| \geq \Gamma(\log(n))^2$.

Let $(L_m)_{m \in \mathcal{M}_n}$ be some family of positive weights satisfying

$$\Sigma = \sum_{m \in \mathcal{M}_n} \exp(-L_m |m|) < +\infty. \tag{2.3}$$

Let $\beta > 1/2$ in the Poisson case, $\beta > 1/2\rho_{\min}$ in the negative binomial case. If for every $m \in \mathcal{M}_n$

$$pen(m) \geq \beta |m| \left(1 + 4\sqrt{L_m}\right)^2, \tag{2.4}$$

then

$$\mathbf{E}\left[h^2(s, \hat{s}_{\hat{m}})\right] \leq C_\beta \inf_{m \in \mathcal{M}_n} \{K(s, \bar{s}_m) + pen(m)\} + C(\phi, \Gamma, \rho_{\min}, \rho_{\max}, \beta, \Sigma), \tag{2.5}$$

with $C_\beta = \dfrac{(16\beta)^{1/3}}{(2\beta)^{1/3} - 1}$ *in model* $(\mathcal{P})$ *and* $C_\beta = \dfrac{(2\rho_{\min}\beta)^{1/3}}{(2\rho_{\min}\beta)^{1/3} - 1}$ *in model* $(\mathcal{NB})$.

We note $h^2(s, u)$ the squared Hellinger distance between distribution s and u and $\bar{s}_m$ is the projection of s onto the collection $\mathcal{S}_m$ according to the Kullback–Leibler distance. The proof of this Theorem is given in Section 5.

Denoting $\bar{s}_m = \arg\min_{u \in \mathcal{S}_m} K(s, u)$, we have for $J \in m$ and $t \in J$,

$$\begin{aligned} &\bar{s}_m(t) = \mathcal{P}(\bar{\lambda}_J) \qquad \text{where } \bar{\lambda}_J = \frac{\sum_{t \in J} \lambda_t}{|J|} \qquad (\mathcal{P}) \\ &\bar{s}_m(t) = \mathcal{NB}(p_J, \phi) \quad \text{where } p_J = \frac{|J|}{\sum_{t \in J} 1/p_t}. \qquad (\mathcal{NB}) \end{aligned} \tag{2.6}$$

Remark 2.4. Before proving the theorem, we make a few remarks:

- The risk of the penalized estimator $\hat{s}_{\hat{m}}$ is treated in terms of Hellinger distance instead of the Kullback–Leibler information. This is due to the fact that the Kullback–Leibler is possibly infinite, and so difficult to control. It is possible to obtain a risk bound in term of Kullback–Leibler if we have a uniform control of $||\log(s/\bar{s}_m)||_\infty$ (see [20] for more explanation).
- The constraints on the θ_ts are imposed to derive the exponential bounds which will be necessary in the proof of the theorem. We would like to recall that whith this parametrization large θs increase both the intensity and the variance of the signal in the Poisson case while it is the contrary in the negative binomial case.

https://doi.org/10.1051/ps/2014005 Published online by Cambridge University Press

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

The penalty function depends on the family $\mathcal{M}_n$ through the choice of the weights L_m which satisfy (2.3). We consider for $\mathcal{M}_n$ the set of all possible partitions of $[\![1, n]\!]$ constructed on a partition m_f which satisfies, for all segment J in m_f, $|J| \geq \Gamma(\log n)^2$. Classically (see [9]) the weights are chosen as a function of the dimension of the model s, which is here $|m|$. The number of partitions of $\mathcal{M}_n$ having dimension D being bounded by $\binom{n}{D}$, we have

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

So with the choice $L_D = 1 + \kappa + \log\left(\frac{n}{D}\right)$ with $\kappa > 0$, condition (2.3) is satisfied. Choosing, say $\kappa = 0.1$, the penalty function can be chosen of the form

$$pen(m) = \beta|m| \left(1 + 4\sqrt{1.1 + \log\left(\frac{n}{|m|}\right)}\right)^2, \tag{2.7}$$

where β is a constant to be calibrated.

Integrating this penalty in Theorem 2.3 leads to the following control:

$$\mathbf{E}\left[h^2(s, \hat{s}_{\hat{m}})\right] \leq C_\beta \inf_{m\in\mathcal{M}_n} \left\{ K(s, \bar{s}_m) + \beta|m| \left(1 + 4\sqrt{1.1 + \log\left(\frac{n}{|m|}\right)}\right)^2 \right\} + C(\phi, \Gamma, \rho_{\min}, \rho_{\max}, \beta, \Sigma) \tag{2.8}$$

The following proposition gives a bound on the Kullback–Leibler risk associated to $\hat{s}_m$:

Proposition 2.5. *Let m be a partition of $\mathcal{M}_n$, $\hat{s}_m$ be the minimum contrast estimator and $\bar{s}_m$ be the projection of s given by equations (2.2) and (2.6) respectively. Assume that there exists some positive absolute constants $\rho_{\min}$, $\rho_{\max}$ and Γ such that $\forall t$, $\rho_{\min} \leq \theta_t \leq \rho_{\max}$ and $|J| \geq \Gamma(\log n)^2$. Then $\forall \varepsilon > 0, \forall a > 2$*

$$K(s, \bar{s}_m) - \frac{C_1(\phi, \Gamma, \rho_{\min}, \rho_{\max}, \varepsilon, a)}{n^{a/2-\alpha}} + C_2(\varepsilon)|m| \leq \mathbf{E}[K(s, \hat{s}_m)],$$

where $\alpha < 1$ is a constant that can be expressed according to n, $C_2(\varepsilon) = \frac{1}{2}\frac{1-\varepsilon}{(1+\varepsilon)^2}$ in the Poisson model $(\mathcal{P})$ and $C_2(\varepsilon) = \rho_{\min}^2 \frac{(1-\varepsilon)^2}{(1+\varepsilon)^4}$ in the negative binomial model $(\mathcal{NB})$.

The proof is given in Appendix A.1.

https://doi.org/10.1051/ps/2014005 Published online by Cambridge University Press

Combining roposition 2.5 and equation (2.8), we obtain the following oracle-type inequality:

Corollary 2.6. *Let $\mathcal{M}_n$ be a collection of partitions constructed on a partition m_f such that there exist absolute positive constants $\rho_{\min}$, $\rho_{\max}$ and Γ verifying:*

- $\forall t, \rho_{\min} \leq \theta_t \leq \rho_{\max}$*; and*
- $\forall J \in m_f, |J| \geq \Gamma(\log n)^2$.

There exists some constant C such that

$$\mathbf{E}\left[h^2(s, \hat{s}_{\hat{m}})\right] \leq C \log(n) \inf_{m \in \mathcal{M}_n} \{\mathbf{E}[K(s, \hat{s}_m)]\} + C(\phi, \Gamma, \rho_{\min}, \rho_{\max}, \beta, \Sigma). \tag{2.9}$$

Remark 2.7. The Hellinger risk of our estimator and the Kullback−Leibler distance between the true model and the best projection are both of order of n. On the contrary, the constant $C(\phi, \Gamma, \rho_{\min}, \rho_{\max}, \beta, \Sigma)$ depends on the constraints used in the assumptions of Theorem 2.3 and is in $\mathcal{O}(1)$. Specifically, in the case of the Poisson distribution, denoting W_1 the constant $C(\phi, \Gamma, \rho_{\min}, \rho_{\max}, \beta, \Sigma)$ in (2.5), we have

$$W_1 = 4\Sigma\left[\frac{20\beta - 1 + 7(2\beta)^{2/3}}{(2\beta)^{2/3} - 1}\right] + \frac{2}{n^{a-1}} + C_\beta\sqrt{2\rho_{\max}}\log\left(\frac{\rho_{\max}}{\rho_{\min}}\right)\frac{1}{n^{(a-1)/2}}.$$

The constant $C(\phi, \Gamma, \rho_{\min}, \rho_{\max}, \beta, \Sigma)$ in (2.9) can then be expressed as

$$C(\phi, \Gamma, \rho_{\min}, \rho_{\max}, \beta, \Sigma) = W_1 + A\frac{\rho_{\max}}{\Gamma\rho_{\min}}\frac{\log(n)}{n^{a/2-\alpha}},$$

where A is a positive constant (independent of the constraints). Consequently, $C(\phi, \Gamma, \rho_{\min}, \rho_{\max}, \beta, \Sigma)$ is negligible compared to $\mathbf{E}[K(s, \hat{s}_m)]$ when n is large meaning that our estimator satisfies an oracle inequality up to a $\log n$ factor. It is now well known that this logarithm term is necessary when considering a large collection of models as is the case here (see [10]). This inequality remains valid as soon as $\rho_{\min}$ is larger than $\rho/\log(n)$.

3. Exponential bounds

In order to prove Theorem 2.3, the general procedure in this model selection framework (see for example [9]) is the following: by definitions of $\hat{m}$ and $\hat{s}_m$ (see Def. 2.2 and Eq. (2.1)), we have, $\forall m \in \mathcal{M}_n$

$$\gamma(\hat{s}_{\hat{m}}) + pen(\hat{m}) \leq \gamma(\hat{s}_m) + pen(m) \leq \gamma(\bar{s}_m) + pen(m).$$

Then, with $\bar{\gamma}(u) = \gamma(u) - \mathbf{E}[\gamma(u)]$,

$$K(s, \hat{s}_{\hat{m}}) \leq K(s, \bar{s}_m) + \bar{\gamma}(\bar{s}_m) - \bar{\gamma}(\hat{s}_{\hat{m}}) - pen(\hat{m}) + pen(m).$$

The idea is therefore to control $\bar{\gamma}(\bar{s}_m) - \bar{\gamma}(\hat{s}_{m'})$ uniformly over $m' \in \mathcal{M}_n$. This is more complicated when dealing with different models m and m'. Thus, following the work of [14] (see proof of Thm. 3.2, also recalled in [20]), we propose the following decomposition

$$\bar{\gamma}(\bar{s}_m) - \bar{\gamma}(\hat{s}_{m'}) = (\bar{\gamma}(\bar{s}_{m'}) - \bar{\gamma}(\hat{s}_{m'})) + (\bar{\gamma}(s) - \bar{\gamma}(\bar{s}_{m'})) + (\bar{\gamma}(\bar{s}_m) - \bar{\gamma}(s)), \tag{3.1}$$

and control each term separately. The first term is the most delicate to handle, and requires the introduction and the control of a chi-square statistic. The main difficulty here is the non-bounded characteristic of the objects we are dealing with. Indeed, in the classic density estimation context such as that of [14], the objects are probabilities which are bounded and so facilitate the direct use of concentration inequalities.
In our case, the chi-square statistic we introduce is denoted χ^2_m and defined by

$$\chi^2_m = \chi^2(\bar{s}_m, \hat{s}_m) = \sum_{J \in m} |J| \frac{(\bar{Y}_J - \bar{E}_J)^2}{\bar{E}_J}, \tag{3.2}$$

https://doi.org/10.1051/ps/2014005 Published online by Cambridge University Press

where we recall that $\bar{Y}_J = \frac{\sum_{t\in J} Y_t}{|J|}$ and use the notation $\bar{E}_J = \frac{E_J}{|J|}$ with $E_J = \sum_{t\in J} E_t$. Respectively for $(\mathcal{P})$ and $(\mathcal{NB})$, we have $E_t = \lambda_t$ and $E_t = \phi\frac{1-p_t}{p_t}$. The purpose is thus to control χ^2_m uniformly over $\mathcal{M}_n$. To this effect, we need to obtain an exponential bound of $Y_J = \sum_{t\in J} Y_t$ around its expectation. In Section 3.1, we recall a result of [4] that we use to derive an exponential bound for χ^2_m (Sect. 3.2).

3.1. Control of Y_J

First we recall a large deviation results established by [4] (Lem. 3) that we apply in the Poisson and negative binomial frameworks.

Lemma 3.1. *Let $Y_1, \dots, Y_n$ be n independent centered random variables.*
If $\log(\mathbf{E}[e^{zY_i}]) \le \kappa\frac{z^2\theta_i}{2(1-z\tau)}$ for all $z \in [0, 1/\tau[$, and $1 \le i \le n$, then

$$\mathbf{P}\left[\sum_{i=1}^n Y_i \ge \left(2\kappa x\sum_{i=1}^n \theta_i\right)^{1/2} + \tau x\right] \le e^{-x} \text{ for all } x > 0.$$

If for $1 \le i \le n$ and all $z > 0$ $\log(\mathbf{E}[e^{-zY_i}]) \le \kappa z^2\theta_i/2$, then

$$\mathbf{P}\left[\sum_{i=1}^n Y_i \le -\left(2\kappa x\sum_{i=1}^n \theta_i\right)^{1/2}\right] \le e^{-x} \text{ for all } x > 0.$$

To apply this lemma we therefore need a majoration of $\log \mathbf{E}\left[e^{z(Y_t-E_t)}\right]$ and $\log \mathbf{E}\left[e^{-z(Y_t-E_t)}\right]$ for $z > 0$.

Poisson case.
With $E_t = \lambda_t$, we have:

$$\log \mathbf{E}\left[e^{z(Y_t-\lambda_t)}\right] = -z\lambda_t + \log \mathbf{E}\left[e^{zY_J}\right] = -z\lambda_t + \log e^{(\lambda_t(e^z-1))} = \lambda_t(e^z - z - 1).$$

Using $e^z - z - 1 \le \frac{z^2}{2(1-z)}$ for $0 < z < 1$ and $e^z - z - 1 \le \frac{z^2}{2}$ for $z < 0$, we have

$$\log \mathbf{E}\left[e^{z(Y_t-E_t)}\right] \le E_t\frac{z^2}{2(1-z)} \quad \text{and} \quad \log \mathbf{E}\left[e^{-z(Y_t-E_t)}\right] \le E_t\frac{z^2}{2}.$$

Negative binomial case. In this case $E_t = \phi\frac{1-p_t}{p_t}$ and we have

$$\begin{aligned}\log \mathbf{E}\left(e^{z\left(Y_t-\phi\frac{1-p_t}{p_t}\right)}\right) &= \frac{z^2}{2}\sum_{k\ge 0}\frac{2\kappa_{k+2}}{(k+2)!}z^k \text{ for } z \le -\log(1-p_t)\\ &\le E_t\frac{z^2}{2}\frac{2}{p_t}\sum_{k\ge 0}\left(\frac{z}{p_t}\right)^k\end{aligned}$$

where the κ_k are the cumulants of the negative binomial distribution.

Then

$$\begin{aligned}\log \mathbf{E}\left(e^{z\left(Y_t-\phi\frac{1-p_t}{p_t}\right)}\right) &\le E_t\frac{z^2}{2}\frac{2}{\rho_{\min}}\frac{1}{1-\frac{z}{\rho_{\min}}} \quad \text{for } z \le \rho_{\min}\\ &\le E_t\frac{z^2}{2}\frac{2}{\rho_{\min}} \qquad\qquad \text{for } -1 \le 0 \le z\end{aligned}$$

https://doi.org/10.1051/ps/2014005 Published online by Cambridge University Press

Finally, with $\kappa = 1$ in the Poisson case and $\kappa = 2/\rho_{\min}$ in the negative binomial case, we get

$$P\left[Y_J - E_J \geq \sqrt{2\kappa x E_J} + \kappa x\right] \leq \mathrm{e}^{-x},$$

leading to

$$P\left[Y_J - E_J \geq x\right] \leq \mathrm{e}^{-\frac{x^2}{2\kappa(E_J+x)}} \quad \text{and} \quad P\left[|Y_J - E_J| \geq x\right] \leq 2\mathrm{e}^{-\frac{x^2}{2\kappa(E_J+x)}}. \tag{3.3}$$

3.2. Exponential bound for χ^2_m

We first introduce the following set Ω_m defined by:

$$\Omega_m(\varepsilon) = \bigcap_{J\in m} \left\{\left|\frac{Y_J}{E_J} - 1\right| \leq \varepsilon\right\}, \tag{3.4}$$

for all $\varepsilon \in]0,1[$ and all segmentations m such that each segment J verifies $|J| \geq \Gamma(\log(n))^2$. This set has a large probability since we obtain

$$\begin{aligned}
\mathbf{P}(\Omega_m(\varepsilon)^C) &\leq \sum_{J\in m} \mathbf{P}\left(|Y_J - E_J| > \varepsilon E_J\right) \leq 2\sum_{J\in m} \mathrm{e}^{-\frac{\varepsilon^2 E_J}{2\kappa(1+\varepsilon)}} \\
&\leq \begin{cases} 2\sum_{J\in m} \mathrm{e}^{-|J|\varepsilon'\rho_{\min}} & \leq 2|m|\exp(-\varepsilon'\Gamma\rho_{\min}(\log n)^2) \qquad (\mathcal{P}) \\ 2\sum_{J\in m} \mathrm{e}^{-|J|\varepsilon'\phi(1-\rho_{\max})\rho_{\min}/2} & \leq 2|m|\exp\left(-\varepsilon'\Gamma\phi(1-\rho_{\max})\rho_{\min}/2(\log n)^2\right) \ (\mathcal{NB}) \end{cases}
\end{aligned}$$

by applying equation (3.3) with $x = \varepsilon E_J$ and where $\varepsilon' = \varepsilon^2/(2(1+\varepsilon))$. Thus

$$\mathbf{P}(\Omega_m(\varepsilon)^C) \leq \frac{2}{n^a}, \tag{3.5}$$

with $a > 2$ as soon as $n > \exp\left(\frac{4(1+\varepsilon)}{\varepsilon^2\Gamma\rho_{\min}}\left(1 \vee 1/(\phi(1-\rho_{\max}))\right)\right)$.

The reason for introducing this set is double: in addition to enable the control of χ^2_m given by equation (3.2) on this restricted set, it allows us to link $K(\hat{s}_m, \bar{s}_m)$ to V_m^2 (see (A.2) for the control of the first term in the decomposition) and so to χ^2_m, relation that we use to evaluate the risk of one model (see (A.4)).

Let m_f be a partition of $\mathcal{M}_n$ such that $\forall J \in m_f, |J| \geq \Gamma(\log(n))^2$ and assume that all considered partitions in $\mathcal{M}_n$ are constructed on this grid m_f. The following proposition gives an exponential bound for χ^2_m on the restricted event $\Omega_{m_f}(\epsilon)$.

Proposition 3.2. *Let $Y_1, \ldots, Y_n$ be independent random variables with distribution $\mathcal{G}$ (Poisson or negative binomial distribution). Let m be a partition of $\mathcal{M}_n$ with $|m|$ segments and χ^2_m the statistic given by (3.2). For any positive x, we have*

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

with $C(\rho_{\min}) = 1$ in the Poisson case and $2/\rho_{\min}$ in the negative binomial case.

Proof. As in the density estimation framework, this quantity can be controlled using the Bernstein inequality. In our context, noting $\chi^2_m = \sum_{J\in m} Z_J$ where

$$Z_J = \frac{(Y_J - E_J)^2}{E_J},$$

https://doi.org/10.1051/ps/2014005 Published online by Cambridge University Press

we need

- the calculation (or bounds) of the expectation of χ^2_m:

  Poisson case.

  Y_J is distributed according to a Poisson distribution with parameter λ_J so that

  $$\mathbf{E}\left[\chi^2_m\right] = |m|. \tag{3.6}$$

  Negative binomial case.

  We have

  $$\mathbf{E}\left[\chi^2_m\right] = \sum_{J\in m}\frac{1}{|J|}\frac{\sum_{t\in J} Var(Y_t)}{\phi\frac{1-p_J}{p_J}} = \sum_{J\in m}\frac{1}{|J|}\frac{\sum_{t\in J}\phi\frac{1-p_t}{p_t^2}}{\phi\frac{1-p_J}{p_J}},$$

  and thus

  $$|m| \leq \mathbf{E}\left[\chi^2_m\right] \leq \frac{1}{\rho_{\min}}|m|. \tag{3.7}$$

- an upper bound of $\sum_{J\in m}\mathbf{E}[Z_J^p]$. For every $p \geq 2$ we have,

  $$\begin{aligned}\mathbf{E}\left[Z_J^p \mathbf{1}_{\Omega_{m_f}(\epsilon)}\right] &= \frac{1}{E_J^p}\int_0^{+\infty} 2p\, x^{2p-1} P\left[\{|Y_J - E_J| \geq x\} \cap \Omega_{m_f}(\epsilon)\right] \mathrm{d}x \\ &\leq \frac{1}{E_J^p}\int_0^{\varepsilon E_J} 2p\, x^{2p-1} P\left[|Y_J - E_J| \geq x\right] \mathrm{d}x.\end{aligned}$$

  Using equation (3.3) and since $x \leq \varepsilon E_J$, we obtain the exponential bound $P\left[|Y_J - E_J| \geq x\right] \leq 2\mathrm{e}^{-\frac{x^2}{2\kappa E_J(1+\varepsilon)}}$. Therefore

  $$\begin{aligned}\mathbf{E}\left[Z_J^p \mathbf{1}_{\Omega_{m_f}(\epsilon)}\right] &\leq \frac{1}{E_J^p}\int_0^{\varepsilon E_J} 4p\, x^{2p-1}\mathrm{e}^{-\frac{x^2}{2\kappa E_J(1+\varepsilon)}}\mathrm{d}x \\ &\leq 4p\kappa^p(1+\varepsilon)^p\int_0^{+\infty} u^{2p-1}\mathrm{e}^{-\frac{u^2}{2}}\mathrm{d}u \\ &\leq 4p\kappa^p(1+\varepsilon)^p\int_0^{+\infty}(2t)^{p-1}\,\mathrm{e}^{-t}\mathrm{d}t \\ &\leq 2^{p+1}p\kappa^p(1+\varepsilon)^p\,p!,\end{aligned}$$

  and

  $$\sum_{J\in m}\mathbf{E}\left[Z_J^p \mathbf{1}_{\Omega_{m_f}(\epsilon)}\right] \leq 2^{p+1}p\kappa^p(1+\varepsilon)^p\,p!|m|.$$

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

  $$\sum_{J\in m}\mathbf{E}\left[Z_J^p \mathbf{1}_{\Omega_{m_f}(\epsilon)}\right] \leq \frac{p!}{2}\times\left[2^5\left(\kappa(1+\varepsilon)\right)^2|m|\right]\times\left[4\left(\kappa(1+\varepsilon)\right)\right]^{p-2}.$$

We conclude by taking $v = 2^5\left(\kappa(1+\varepsilon)\right)^2|m|$ and $c = 4\left(\kappa(1+\varepsilon)\right)$ (see Prop. 2.9 of [20] for the definition of the Bernstein's inequality). □

https://doi.org/10.1051/ps/2014005 Published online by Cambridge University Press

TABLE 1. *Properties of segmentation algorithms.* The first column indicates the name of the algorithm and the criterion used for the choice of K. In the second column, NB stands for the negative binomial distribution and P for Poisson. The time of each algorithm is given (column "Complexity") and column "Exact" precises if the exact solution is reached.

Algorithm	Dist	Complexity	Inference	Pen	Exact	Reference
Penalized PDPA	NB	$n \log n$	frequentist	external	exact	[15]
PDPA with BIC	NB	$n \log n$	frequentist	external	exact	[15]
Penalized PDPA	P	$n \log n$	frequentist	external	exact	[15]
PDPA with BIC	P	$n \log n$	frequentist	external	exact	[15]
PELT with BIC	P	n	frequentist	internal	exact	[17]
CART with BIC	P	$n \log n$	frequentist	external	heuristic	[13]
postCP with ICL	NB	n	frequentist	external	exact	[19]
EBS with ICL	NB	n^2	Bayesian	external	exact	[23]

4. SIMULATIONS AND APPLICATION

In the context of RNA-seq experiments, an important question is the (re)-annotation of the genome, that is, the precise localisation of the transcribed regions on the chromosomes. In an ideal situation, when considering the number of reads starting at each position, one would expect to observe a uniform coverage over each gene (proportional to its expression level), separated by regions of null signal (corresponding to non-transcribed regions of the genome). In practice however, those experiments tend to return very noisy signals that are best modelled by the negative binomial distribution.

In this Section, we first study the performance of the proposed penalized criterion by comparing it with others model selection criteria on a resampling dataset (Sect. 4.1). Then we provide an application on real data (Sect. 4.2). Since the penalty depends on the partition only through its size, the segmentation procedure is two-steps: first we estimate, for all number of segments K between 1 and $K_{\max}$, the optimal partition with K segments (*i.e.* construct the collection of estimators $\{\hat{s}_K\}_{1 \leq K \leq K_{\max}}$ where $\hat{s}_K = \arg\min_{\hat{s}_m, m \in \mathcal{M}_K} \{\gamma(\hat{s}_m)\}$). The optimal solution is obtained using a fast segmentation algorithm such as the Pruned Dynamic Programming Algorithm (PDPA, [22]) implemented for the Poisson and negative binomial losses or contrasts in the R package `Segmentor3IsBack` [15]. Then, we choose K using our penalty function which requires the calibration of the constant β that can be tuned according to the data by using the slope heuristic (see [3,10]). Using the negative binomial distribution requires the knowledge of parameter ϕ. We propose to estimate it using a modified version of the Jonhson and Kotz's estimator [16].

4.1. Simulation study

We have assessed the performances of the proposed method (called Penalized PDPA) on a simulation scenario by comparing to five other procedures both its choice in the number of segments and the quality of the obtained segmentation using the Rand-Index $\mathcal{I}$. This index is defined as follows: let C_t be the true index of the segment to which base t belongs and let $\hat{C}_t$ be the corresponding estimated index, then

$$\mathcal{I} = \frac{2 \sum_{t>s} \left[\mathbf{1}_{C_t = C_s} \mathbf{1}_{\hat{C}_t = \hat{C}_s} + \mathbf{1}_{C_t \neq C_s} \mathbf{1}_{\hat{C}_t \neq \hat{C}_s} \right]}{n(n-1)}.$$

The characteristics of the different algorithms are described in Table 1.

The data we considered comes from a resampling procedure using real RNA-seq data. The original data, from an RNA-Seq experiment on the yeast organism by the Sherlock Genomics laboratory at Stanford University, is publicly available on the NCBIs Sequence Read Archive (SRA, `url:http://www.ncbi.nlm.nih.gov/sra`) with the accession number SRA048710. After pooling the original data into four categories according to the biological annotation, we created an artificial gene, inspired from the *Drosophila* inr-a gene, resulting in a

https://doi.org/10.1051/ps/2014005 Published online by Cambridge University Press

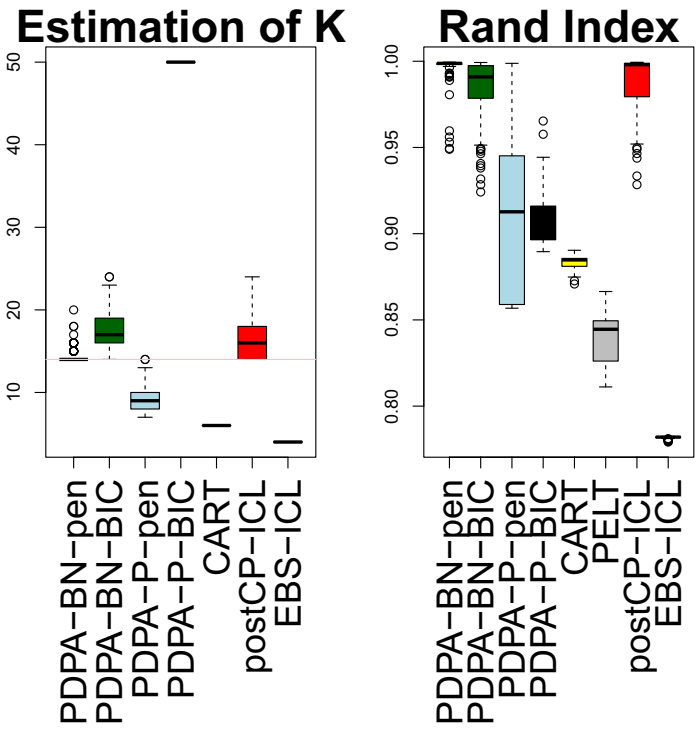


FIGURE 1. *Estimation of K on resampled datasets.* Left: boxplot of the estimation of K on datasets simulated by resampling on artificial gene Inr-a. PELT's estimates average at 427 segments and are not shown. The pink horizontal line indicates the true value of K. Right: boxplot of the Rand-Index values for the proposed estimators.

14-segment signal with unregular intensities mimicking a differentially transcribed gene. 100 datasets are thus created. Because the profile is sampled from a true experiment, its true segmentation is given by the artificial gene boundaries while the distribution of the data is unavailable. The Rand-Index criterion is therefore most appropriate to assess the quality of the proposed segmentation. Results are presented using boxplots in Figure 1. Because PELT's estimate of K averaged around 427 segments, we did not show its corresponding boxplot.

We can see that with the negative binomial distribution, not only do we perfectly recover the true number of segments, but our procedure outperforms all other approaches. Moreover, the impressive results in terms of Rand-Index prove that our choice of number of segments also leads to the almost perfect recovery of the true segmentation. However, the use of the Poisson loss leads to a constant underestimation of the number of segments, which is reflected on the Rand-Index values. This is due to the inappropriate choice of distribution (confirmed by the other algorithms implemented for the Poisson loss which perform worse than the others). It however underlines the need for the development of methods for the negative binomial distribution. Moreover, in terms of computational time, the fast algorithm [15] is in $\mathcal{O}(n \log n)$, allowing its use on long signals (such as a whole-genome analysis), even though it is not as fast as CART or PELT.

4.2. Segmentation of RNA-Seq data

We apply our proposed procedure for segmenting chromosome 1 of the S. Cerevisiae (yeast) using RNA-Seq data from the Sherlock Laboratory at Stanford University [24] and publicly available from the NCBI's Sequence Read Archive (SRA, `url:http://www.ncbi.nlm.nih.gov/sra`, accession number SRA048710). An existing

https://doi.org/10.1051/ps/2014005 Published online by Cambridge University Press

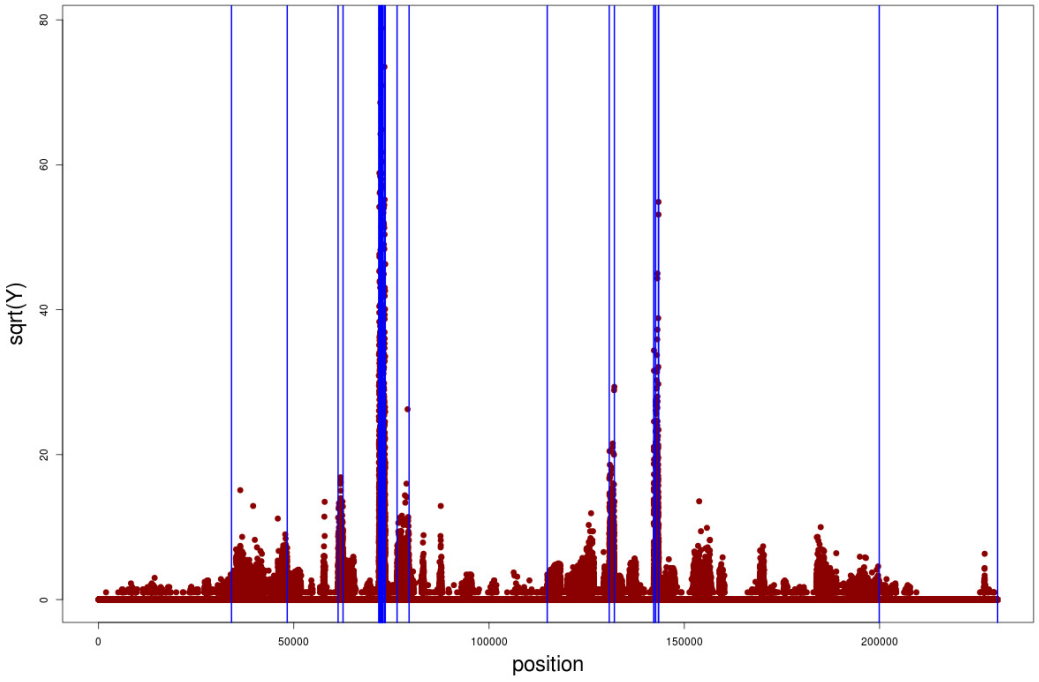


FIGURE 2. *Segmentation of the yeast chromosome 1 using Poisson loss.* Read-count are represented on a root-squared scale. The model selection procedure chooses $K = 106$ segments.

annotation is available on the Saccharomyces Genome Database (SGD) at `url:http://www.yeastgenome.org`, which allows us to validate our results. The two distributions (Poisson and negative binomial) are considered here to show the difference.

In the Poisson distribution case, we select 106 segments of which only 19 are related to the SGD annotation. Indeed, as illustrated by Figure 2, 36 of the segments have a length smaller than 10: the Poisson loss is note adapted to this kind of data with high variability and it tends to select outliers as segment. On the contrary, we select 103 segments in the negative binomial case most of which (all but 3) surround known genes from the SGD. Figure 3 illustrates the result. However, almost none of those change-points correspond exactly to annotated boundaries. Discussion with biologists has increased our belief in the need for genome (re-)annotation using RNA-seq data, and in the validity of our approach.

5. Proof of Theorem 2.3

Recall that we want to control the three terms in the decomposition given by (3.1). All the proofs of the different propositions are given in Section 5.

- The control of the term $\bar{\gamma}(\hat{s}_{m'}) - \bar{\gamma}(\bar{s}_{m'})$ is obtained with the following proposition where the set $\Omega_1(\xi)$ is defined by

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

https://doi.org/10.1051/ps/2014005 Published online by Cambridge University Press

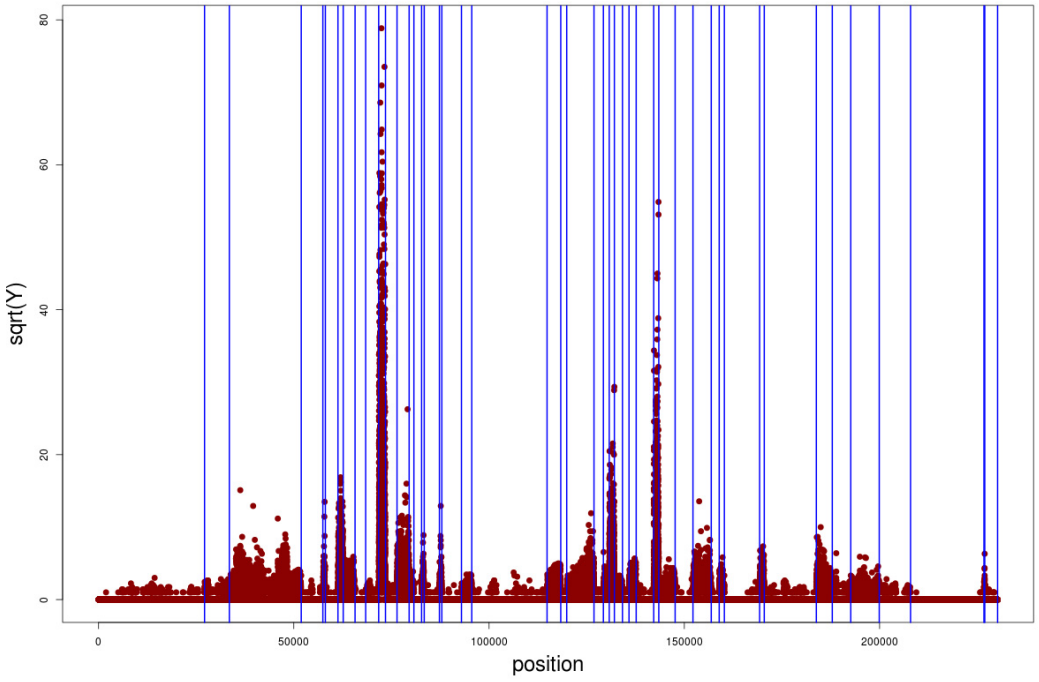


FIGURE 3. *Segmentation of the yeast chromosome 1 using the negative binomial loss.* The model selection procedure chooses $K = 103$ segments, most of which surround genes given by the SGD annotation.

Proposition 5.1. *Let m' be a partition of $\mathcal{M}_n$. Then*

$$(\bar{\gamma}(\hat{s}_{m'}) - \bar{\gamma}(\bar{s}_{m'}))\,\mathbf{1}_{\Omega_{m_f}(\epsilon)\cap\Omega_1(\xi)} \leq C(\varepsilon)\Big[|m'| + 8(1+\varepsilon)\sqrt{(L_{m'}|m'|+\xi)|m'|} + 4(1+\varepsilon)(L_{m'}|m'|+\xi)\Big] + \frac{1}{1+\varepsilon}K(\bar{s}_{m'}, \hat{s}_{m'}),$$

with $C(\varepsilon) = \frac{1}{2}\left(\frac{1+\varepsilon}{1-\varepsilon}\right)$ in the Poisson case and $C(\varepsilon) = \frac{1+\varepsilon}{2\rho_{\min}}$ in the negative binomial case.

- The control of the term $\bar{\gamma}(\bar{s}_m) - \bar{\gamma}(s)$, or more precisely its expectation, is given by the following proposition:

Proposition 5.2.

$$|\mathbf{E}[(\bar{\gamma}(\bar{s}_m) - \bar{\gamma}(s))\mathbf{1}_{\Omega_{m_f}(\epsilon)}]| \leq \frac{C(\phi, \Gamma, \rho_{\min}, \rho_{\max}, \varepsilon, a)}{n^{(a-1)/2}}. \tag{5.1}$$

- To control $\bar{\gamma}(s) - \bar{\gamma}(\bar{s}_{m'})$, we use the following proposition which gives an exponential bound for $\bar{\gamma}(s) - \bar{\gamma}(u)$.

Proposition 5.3. *Let s and u be two distributions of a sequence Y. Let γ be the log-likelihood contrast, $\bar{\gamma}(u) = \gamma(u) - \mathbf{E}[\gamma(u)]$, and $K(s,u)$ and $h^2(s,u)$ be respectively the Kullback–Leibler and the squared Hellinger distances between distributions s and u. Then $\forall x > 0$,*

$$\mathbf{P}\left[\bar{\gamma}(s) - \bar{\gamma}(u) \geq K(s,u) - 2h^2(s,u) + 2x\right] \leq \mathrm{e}^{-x}.$$

https://doi.org/10.1051/ps/2014005 Published online by Cambridge University Press

Applying it to $u = \bar{s}_{m'}$ yields:

$$\mathbf{P}\left[\bar{\gamma}(s) - \bar{\gamma}(\bar{s}_{m'}) \geq K(s, \bar{s}_{m'}) - 2h^2(s, \bar{s}_{m'}) + 2x\right] \leq \mathrm{e}^{-x}. \tag{5.2}$$

We then define

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

Let $\Omega(\varepsilon, \xi) = \Omega_{m_f}(\epsilon) \cap \Omega_1(\xi) \cap \Omega_2(\xi)$. Then, combining equation (5.2) and Proposition 5.1, we get for $m' = \hat{m}$,

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

with $R = \bar{\gamma}(\bar{s}_m) - \bar{\gamma}(s)$. So that

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

And since

- $K(s, \hat{s}_{\hat{m}}) = K(s, \bar{s}_{\hat{m}}) + K(\bar{s}_{\hat{m}}, \hat{s}_{\hat{m}})$ (see Eq. (A.1));
- $K(s, u) \geq 2h^2(s, u)$ (see Lem. 7.23 in [20]);
- $h^2(s, \hat{s}_{\hat{m}}) \leq 2\left(h^2(s, \bar{s}_{\hat{m}}) + h^2(\bar{s}_{\hat{m}}, \hat{s}_{\hat{m}})\right)$ (using inequality $2ab \leq \kappa a^2 + \kappa^{-1}b^2$ with $\kappa = 1$).

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

But

$$\begin{aligned}
C(\varepsilon)\left[1 + (1+\varepsilon)\left(8\sqrt{L_{\hat{m}}} + \varepsilon + 4L_{\hat{m}}\right)\right] + 2L_{\hat{m}} &\leq C(\varepsilon)\left[1 + (1+\varepsilon)\left(\varepsilon + 8\sqrt{L_{\hat{m}}} + 8L_{\hat{m}}\right)\right] \\
&\leq C_2(\varepsilon)\left[1 + 8\sqrt{L_{\hat{m}}} + 8L_{\hat{m}}\right].
\end{aligned}$$

with $C_2(\varepsilon) = \frac{1}{2}\left(\frac{1+\varepsilon}{1-\varepsilon}\right)^3$ for $(\mathcal{P})$ and $C_2(\varepsilon) = \frac{1}{2\rho_{\min}}(1+\varepsilon)^3$ for $(\mathcal{NB})$. So we have

$$\begin{aligned}
\frac{\varepsilon}{1+\varepsilon}h^2(s, \hat{s}_{\hat{m}})\mathbf{1}_{\Omega(\epsilon,\xi)} &\leq K(s, \bar{s}_m)\mathbf{1}_{\Omega(\epsilon,\xi)} + R\mathbf{1}_{\Omega(\epsilon,\xi)} - pen(\hat{m}) + pen(m) \\
&\quad + |\hat{m}|C_2(\varepsilon)\left(1 + 4\sqrt{L_{\hat{m}}}\right)^2 + 2\xi\left[1 + (1+\varepsilon)C(\varepsilon)\left(\frac{8}{\varepsilon} + 2\right)\right].
\end{aligned}$$

https://doi.org/10.1051/ps/2014005 Published online by Cambridge University Press

By assumption, $pen(\hat{m}) \geq \beta|\hat{m}|\left(1+4\sqrt{L_{\hat{m}}}\right)^2$. Choosing $\beta = C_2(\varepsilon)$ yields

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

Then, using Propositions 5.2 and 5.1, we have $\mathbf{P}\left(\Omega_1(\xi)^C\right) \leq \sum_{m'\in\mathcal{M}_n} \mathrm{e}^{-L_{m'}|m'|+\xi}$ and $\mathbf{P}\left(\Omega_2(\xi)^C\right) \leq \sum_{m'\in\mathcal{M}_n} \mathrm{e}^{-L_{m'}|m'|+\xi}$. So that using hypothesis (2.3),

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

and thus $\mathbf{P}\left(\Omega_1(\xi)\cap\Omega_2(\xi)\right) \geq 1-2\Sigma\mathrm{e}^{-\xi}$. We now integrate over ξ, and using equation (5.1), we obtain:

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

And since $\mathbf{E}\left[h^2(s,\hat{s}_{\hat{m}})\mathbf{1}_{\Omega_{m_f}(\epsilon)^C}\right] \leq \dfrac{C(\phi,\Gamma,\rho_{\min},\rho_{\max},\beta,a)}{n^{a-1}}$, we have

$$\mathbf{E}\left[h^2(s,\hat{s}_{\hat{m}})\right] \leq C_\beta\left[K(s,\bar{s}_m) + pen(m)\right] + C'\left(\phi,\Gamma,\rho_{\min},\rho_{\max},\beta,\Sigma\right).$$

Finally, by minimizing over $m\in\mathcal{M}_n$, we get

$$\mathbf{E}\left[h^2(s,\hat{s}_{\hat{m}})\right] \leq C_\beta \inf_{m\in\mathcal{M}_n}\left\{K(s,\bar{s}_m) + pen(m)\right\} + C'\left(\phi,\Gamma,\rho_{\min},\rho_{\max},\beta,\Sigma\right).$$

Appendices A.

A.1. Proof of Proposition 2.5

Using Pythagore-type identity, we obtain the following decomposition (see for example [14]):

$$K(s,\hat{s}_m) = K(s,\bar{s}_m) + K(\bar{s}_m,\hat{s}_m). \tag{A.1}$$

The objective is then to obtain a lower bound of $\mathbf{E}[K(\bar{s}_m,\hat{s}_m)]$ in the two considered distribution cases.

Poisson case.

We have

$$K(\bar{s}_m,\hat{s}_m) = \sum_{J\in m}|J|\left(\bar{Y}_J - \bar{\lambda}_J - \bar{\lambda}_J\log\frac{\bar{Y}_J}{\bar{\lambda}_J}\right) = \sum_{J\in m}|J|\bar{\lambda}_J\Phi\left(\log\frac{\bar{Y}_J}{\bar{\lambda}_J}\right).$$

where $\Phi(x) = \mathrm{e}^x - 1 - x$. Since $\dfrac{1}{2}x^2(1\wedge\mathrm{e}^x) \leq \Phi(x) \leq \dfrac{1}{2}x^2(1\vee\mathrm{e}^x)$, then on $\Omega_{m_f}(\epsilon)$, we have

$$\frac{1}{2}\log^2\frac{\bar{Y}_J}{\bar{\lambda}_J}\left(1\wedge\frac{\bar{Y}_J}{\bar{\lambda}_J}\right) \leq \Phi\left(\log\frac{\bar{Y}_J}{\bar{\lambda}_J}\right) \leq \frac{1}{2}\log^2\frac{\bar{Y}_J}{\bar{\lambda}_J}\left(1\vee\frac{\bar{Y}_J}{\bar{\lambda}_J}\right),$$
$$\frac{1-\varepsilon}{2}\log^2\frac{\bar{Y}_J}{\bar{\lambda}_J} \leq \Phi\left(\log\frac{\bar{Y}_J}{\bar{\lambda}_J}\right) \leq \frac{1+\varepsilon}{2}\log^2\frac{\bar{Y}_J}{\bar{\lambda}_J}.$$

So

$$\frac{1-\varepsilon}{2}V_m^2 \leq K(\bar{s}_m,\hat{s}_m) \leq \frac{1+\varepsilon}{2}V_m^2, \tag{A.2}$$

https://doi.org/10.1051/ps/2014005 Published online by Cambridge University Press

where

$$V_m^2 = V^2(\bar{s}_m, \hat{s}_m) = \sum_{J\in m} |J|\bar{\lambda}_J \log^2 \frac{\bar{Y}_J}{\bar{\lambda}_J} = \sum_{J\in m} |J| \frac{(\bar{Y}_J - \bar{\lambda}_J)^2}{\bar{\lambda}_J} \left(\frac{\log \frac{\bar{Y}_J}{\bar{\lambda}_J}}{\frac{\bar{Y}_J}{\bar{\lambda}_J} - 1} \right)^2. \tag{A.3}$$

And using, for $x > 0$, $\frac{1}{1\vee x} \le \frac{\log x}{x-1} \le \frac{1}{1\wedge x}$, we get, on $\Omega_{m_f}(\epsilon)$

$$\frac{1}{(1+\varepsilon)^2}\, \chi_m^2 \le V_m^2 \le \frac{1}{(1-\varepsilon)^2}\, \chi_m^2. \tag{A.4}$$

So

$$\frac{1-\varepsilon}{2(1+\varepsilon)^2}\, \chi_m^2 \mathbf{1}_{\Omega_{m_f}(\epsilon)} \le K(\bar{s}_m, \hat{s}_m)\mathbf{1}_{\Omega_{m_f}(\epsilon)} \le \frac{1+\varepsilon}{2(1-\varepsilon)^2}\, \chi_m^2 \mathbf{1}_{\Omega_{m_f}(\epsilon)}.$$

On one hand, $\mathbf{E}\left[\chi_m^2\right] = |m|$, and

$$\frac{1-\varepsilon}{2(1+\varepsilon)^2}|m| - \mathbf{E}\left[\chi_m^2 \mathbf{1}_{\Omega_{m_f}(\epsilon)^C}\right] \le \mathbf{E}\left[K(\bar{s}_m, \hat{s}_m)\mathbf{1}_{\Omega_{m_f}(\epsilon)}\right] \le \frac{1+\varepsilon}{2(1-\varepsilon)^2}|m|.$$

Since $\chi_m^2 \le \frac{1}{\Gamma(\log(n))^2\rho_{\min}} \sum_{J\in m}(Y_J - \lambda_J)^2 \le \frac{1}{\Gamma(\log(n))^2\rho_{\min}} \left(\sum_t Y_t - \sum_t \lambda_t\right)^2$, using Cauchy–Schwarz Inequality, we get

$$\begin{aligned}
\mathbf{E}\left[\chi_m^2 \mathbf{1}_{\Omega_{m_f}(\epsilon)^C}\right] &\le \frac{1}{\Gamma(\log(n))^2\rho_{\min}} \left[3\left(\sum_t \lambda_t\right)^2 + \sum_t \lambda_t\right] P(\Omega_{m_f}(\epsilon)^C)^{1/2} \\
&\le C\left(\Gamma, \rho_{\min}, \rho_{\max}\right) \frac{n}{(\log(n))^2} P(\Omega_{m_f}(\epsilon)^C)^{1/2} \\
&\le C\left(\Gamma, \rho_{\min}, \rho_{\max}\right) n^{\alpha} P(\Omega_{m_f}(\epsilon)^C)^{1/2} \\
&\le \frac{C\left(\phi, \Gamma, \rho_{\min}, \rho_{\max}, \varepsilon, a\right)}{n^{a/2-\alpha}},
\end{aligned}$$

where $\alpha = 1 - 2\frac{\log(\log(n))}{\log(n)}$, $n \ge 2$. For example, $\alpha = 0.62$ for $n = 10^6$.

On the other hand, using $\log 1/x \ge 1 - x$ for all $x > 0$, $\mathbf{E}\left[K(\bar{s}_m, \hat{s}_m)\mathbf{1}_{\Omega_{m_f}(\epsilon)^C}\right] \ge 0$. Finally, we have

$$K(s, \bar{s}_m) + \frac{1-\varepsilon}{2(1+\varepsilon)^2}|m| - \frac{C_1(\Gamma, \rho_{\min}, \rho_{\max}, \varepsilon, a)}{n^{a/2-\alpha}} \le \mathbf{E}[K(s, \hat{s}_m)],$$

Negative binomial case.

We have $K(\bar{s}_m, \hat{s}_m) = \phi \sum_{J\in m} \frac{|J|}{p_J} h_{\frac{\phi}{\phi+\bar{Y}_J}}(p_J)$, and $\forall 0 < a < 1$, $\quad h_a(x) \ge \frac{1-x}{1-a}\log^2\left(\frac{1-x}{1-a}\right)$.

Then on $\Omega_{m_f}(\epsilon)$

$$K(\bar{s}_m, \hat{s}_m) \ge \phi \sum_{J\in m} \frac{|J|}{p_J} \frac{1-p_J}{\frac{\bar{Y}_J}{\phi+\bar{Y}_J}} \log^2\left(\frac{\frac{\bar{Y}_J}{\phi+\bar{Y}_J}}{1-p_J}\right).$$

https://doi.org/10.1051/ps/2014005 Published online by Cambridge University Press

Introducing

$$V_m^2 = \sum_{J\in m} \phi|J| \frac{1-p_J}{p_J} \log^2\left(\frac{\frac{\bar{Y}_J}{\phi+\bar{Y}_J}}{1-p_J}\right), \tag{A.5}$$

we get

$$K(\bar{s}_m, \hat{s}_m) \geq V_m^2, \tag{A.6}$$

and since $\bar{Y}_J - \phi\frac{1-p_J}{p_J} = \frac{\phi+\bar{Y}_J}{p_J}\left(\frac{\bar{Y}_J}{\phi+\bar{Y}_J} - (1-p_J)\right)$, we have

$$V_m^2 = \sum_{J\in m} |J| \left(\frac{\phi}{\phi+\bar{Y}_J}\right)^2 \frac{\left(\bar{Y}_J - \phi\frac{1-p_J}{p_J}\right)^2}{\phi\frac{1-p_J}{p_J}} \left[\frac{\log\left(\frac{\frac{\bar{Y}_J}{\phi+\bar{Y}_J}}{1-p_J}\right)}{\frac{\frac{\bar{Y}_J}{\phi+\bar{Y}_J}}{1-p_J} - 1}\right]^2.$$

And finally,

$$K(\bar{s}_m, \hat{s}_m)\mathbf{1}_{\Omega_{m_f}(\epsilon)} \geq \rho_{\min}^2 \frac{(1-\varepsilon)^2}{(1+\varepsilon)^4}\ \chi_m^2 \mathbf{1}_{\Omega_{m_f}(\epsilon)}.$$

Moreover, on one hand we have $|m| \leq \mathbf{E}\left[\chi_m^2\right] \leq \frac{1}{\rho_{\min}}|m|$. On the other hand, since $\chi_m^2 \leq \frac{1}{\Gamma(\log(n))^2\phi(1-\rho_{\max})}\left(\sum_t Y_t - \sum_t E_t\right)^2$, using Cauchy–Schwarz Inequality, we get

$$\begin{aligned}
\mathbf{E}\left[\chi_m^2 \mathbf{1}_{\Omega_{m_f}(\epsilon)^C}\right] &\leq \frac{\left[\sum_t \mathbf{E}\left(Y_t - E_t\right)^4 + 6\phi^2 \sum_{(t,l),l\neq t} \frac{1-p_t}{p_t^2}\frac{1-p_l}{p_l^2}\right]^{1/2}}{\Gamma(\log(n))^2\phi(1-\rho_{\max})} P(\Omega_{m_f}(\epsilon)^C)^{1/2}, \\
&\leq C(\phi, \Gamma, \rho_{\min}, \rho_{\max}) n^{\alpha} P(\Omega_{m_f}(\epsilon)^C)^{1/2}, \\
&\leq \frac{C(\phi, \Gamma, \rho_{\min}, \rho_{\max}, \varepsilon, a)}{n^{a/2-\alpha}},
\end{aligned}$$

where $\alpha = 1 - 2\frac{\log(\log(n))}{\log(n)}$, $n \geq 2$. Finally, we have

$$K(s, \bar{s}_m) + \rho_{\min}^2 \frac{(1-\varepsilon)^2}{(1+\varepsilon)^4}|m| - \frac{C(\phi, \Gamma, \rho_{\min}, \rho_{\max}, \varepsilon, a)}{n^{a/2-\alpha}} \leq \mathbf{E}[K(s, \hat{s}_m)].$$

A.2. Proof of Proposition 5.1

Poisson case.

The term to be controlled is $\bar{\gamma}(\bar{s}_{m'}) - \bar{\gamma}(\hat{s}_{m'}) = \sum_{J\in m'} |J|\left(\bar{Y}_J - \bar{\lambda}_J\right)\log\frac{\bar{Y}_J}{\bar{\lambda}_J}$. Using Cauchy–Schwarz inequality, we have

$$\bar{\gamma}(\bar{s}_{m'}) - \bar{\gamma}(\hat{s}_{m'}) \leq \sqrt{\chi_{m'}^2}\ \sqrt{V_{m'}^2},$$

with $\chi_{m'}^2$ and $V_{m'}^2$ defined as in equations (3.2) and (A.3). Then, using equation (A.2)

$$\left(\bar{\gamma}(\bar{s}_{m'}) - \bar{\gamma}(\hat{s}_{m'})\right)\mathbf{1}_{\Omega_{m_f}(\epsilon)} \leq \sqrt{\chi_{m'}^2}\ \sqrt{\frac{2}{1-\varepsilon}K(\bar{s}_{m'}, \hat{s}_{m'})},$$

https://doi.org/10.1051/ps/2014005 Published online by Cambridge University Press

and using $2ab \leq \kappa a^2 + \kappa^{-1} b^2$ for all $\kappa > 0$, we get

$$\left(\bar{\gamma}(\bar{s}_{m'}) - \bar{\gamma}(\hat{s}_{m'})\right) \mathbf{1}_{\Omega_{m_f}(\epsilon)} \leq \frac{\kappa}{2}\chi^2_{m'} + \frac{\kappa^{-1}}{1-\varepsilon} K(\bar{s}_{m'}, \hat{s}_{m'}). \tag{A.7}$$

And with Proposition 3.2, we get, for $\kappa = \dfrac{1+\varepsilon}{1-\varepsilon} = 2C(\varepsilon)$,

$$\begin{aligned}
&\left(\bar{\gamma}(\bar{s}_{m'}) - \bar{\gamma}(\hat{s}_{m'})\right) \mathbf{1}_{\Omega_{m_f}(\epsilon)\cap\Omega_1(\xi)} \\
&\leq \frac{1+\varepsilon}{2(1-\varepsilon)} \left[|m'| + 8(1+\varepsilon)\sqrt{(L_{m'}|m'| + \xi)|m'|} + 4(1+\varepsilon)(L_{m'}|m'| + \xi)\right] + \frac{1}{1+\varepsilon} K(\bar{s}_{m'}, \hat{s}_{m'}).
\end{aligned}$$

Negative binomial case.

In this case we can write $\bar{\gamma}(\bar{s}_{m'}) - \bar{\gamma}(\hat{s}_{m'}) = \sum_{J\in m'} |J| \left(\bar{Y}_J - \bar{E}_J\right) \log \frac{\frac{\bar{Y}_J}{\phi+\bar{Y}_J}}{1-p_J}$. Again, using Cauchy–Schwarz inequality, and with χ^2_m and V^2_m defined by equations (3.2) and (A.5), we get

$$\bar{\gamma}(\bar{s}_{m'}) - \bar{\gamma}(\hat{s}_{m'}) \leq \sqrt{\chi^2_{m'}} \quad \sqrt{V^2_{m'}},$$

so that with equation (A.6) and $2ab \leq \kappa a^2 + \kappa^{-1} b^2$ for all $\kappa > 0$

$$\left(\bar{\gamma}(\bar{s}_{m'}) - \bar{\gamma}(\hat{s}_{m'})\right) \mathbf{1}_{\Omega_{m_f}(\epsilon)} \leq \frac{\kappa}{2}\chi^2_{m'} + \frac{\kappa^{-1}}{2} K\left(\bar{s}_{m'}, \hat{s}_{m'}\right). \tag{A.8}$$

Finally, with Proposition 3.2 and $\kappa = \dfrac{1+\varepsilon}{2} = 2C(\varepsilon)$,

$$\begin{aligned}
&\left(\bar{\gamma}(\bar{s}_{m'}) - \bar{\gamma}(\hat{s}_{m'})\right) \mathbf{1}_{\Omega_{m_f}(\epsilon)\cap\Omega_1(\xi)} \\
&\leq C\left(\rho_{\min}\right) \frac{1+\varepsilon}{4} \left[|m'| + 8(1+\varepsilon)\sqrt{(L_{m'}|m'| + \xi)|m'|} + 4(1+\varepsilon)(L_{m'}|m'| + \xi)\right] + \frac{1}{1+\varepsilon} K\left(\bar{s}_{m'}, \hat{s}_{m'}\right).
\end{aligned}$$

A.3. Proof of Proposition 5.2

Poisson case.

Noting that $\mathbf{E}[(\bar{\gamma}(\bar{s}_m) - \bar{\gamma}(s))\mathbf{1}_{\Omega_{m_f}(\epsilon)}] = -\mathbf{E}[(\bar{\gamma}(\bar{s}_m) - \bar{\gamma}(s))\mathbf{1}_{\Omega_{m_f}(\epsilon)^C}]$, we have

$$\begin{aligned}
\left|\mathbf{E}\left[(\bar{\gamma}(\bar{s}_m) - \bar{\gamma}(s))\mathbf{1}_{\Omega_{m_f}(\epsilon)}\right]\right| &\leq \left|\mathbf{E}\left[(\bar{\gamma}(\bar{s}_m) - \bar{\gamma}(s))\mathbf{1}_{\Omega_{m_f}(\epsilon)^C}\right]\right| \leq \mathbf{E}\left[\left|(\bar{\gamma}(\bar{s}_m) - \bar{\gamma}(s))\right| \mathbf{1}_{\Omega_{m_f}(\epsilon)^C}\right] \\
&\leq \mathbf{E}\left[\left|\left(\sum_J \sum_t (Y_t - E_t) \log\left(\rho_{\max}/\rho_{\min}\right)\right)\right| \mathbf{1}_{\Omega_{m_f}(\epsilon)^C}\right] \\
&\leq \log\left(\rho_{\max}/\rho_{\min}\right) \times \mathbf{E}\left[\left|\sum_t (Y_t - E_t)\right| \mathbf{1}_{\Omega_{m_f}(\epsilon)^C}\right] \\
&\leq \log\left(\rho_{\max}/\rho_{\min}\right) \times \left(\left[\mathbf{E}\left(\sum_t (Y_t - E_t)\right)^2\right]^{1/2} \times \left(P(\Omega_{m_f}(\epsilon)^C\right)^{1/2}\right) \\
&\leq \left(n\rho_{\max}\right)^{1/2} \times \log\left(\rho_{\max}/\rho_{\min}\right) \times \left(P\left(\Omega_{m_f}(\epsilon)^C\right)^{1/2},
\end{aligned}$$

which concludes the proof.

https://doi.org/10.1051/ps/2014005 Published online by Cambridge University Press

Negative binomial case.

Once again, $\mathbf{E}[(\bar{\gamma}(\bar{s}_m) - \bar{\gamma}(s))\mathbf{1}_{\Omega_{m_f}(\epsilon)}] = -\mathbf{E}[(\bar{\gamma}(\bar{s}_m) - \bar{\gamma}(s))\mathbf{1}_{\Omega_{m_f}(\epsilon)^C}]$, and

$$\begin{aligned}
|\mathbf{E}[(\bar{\gamma}(\bar{s}_m) - \bar{\gamma}(s))\mathbf{1}_{\Omega_{m_f}(\epsilon)}]| &\leq |\mathbf{E}[(\bar{\gamma}(\bar{s}_m) - \bar{\gamma}(s))\mathbf{1}_{\Omega_{m_f}(\epsilon)^C}]| \leq \mathbf{E}[|(\bar{\gamma}(\bar{s}_m) - \bar{\gamma}(s))|\mathbf{1}_{\Omega_{m_f}(\epsilon)^C}] \\
&\leq \mathbf{E}\left[\left|\left(\sum_J \sum_t \left(Y_t - \phi\frac{1-p_t}{p_t}\right)\log\left(1/(1-\rho_{\min})\right)\right)\right| \mathbf{1}_{\Omega_{m_f}(\epsilon)^C}\right] \\
&\leq \log\left(1/(1-\rho_{\min})\right) \times \mathbf{E}\left[\left|\sum_t (Y_t - E_t)\right| \mathbf{1}_{\Omega_{m_f}(\epsilon)^C}\right] \\
&\leq \left(n\phi\frac{1}{\rho_{\min}^2}\right)^{1/2} \times \log\frac{1}{1-\rho_{\min}} \times \left(P\left(\Omega_{m_f}(\epsilon)^C\right)\right)^{1/2}
\end{aligned}$$

which concludes the proof.

A.4. Proof of Proposition 5.3

Using the Markov inequality $\mathbf{P}\left[\bar{\gamma}(s) - \bar{\gamma}(u) \geq b\right] \leq \inf_a \left[\mathrm{e}^{-ab}\mathbf{E}\left(\mathrm{e}^{a(\bar{\gamma}(s)-\bar{\gamma}(u))}\right)\right]$ with $a = \frac{1}{2}$, we get

$$\begin{aligned}
\mathbf{P}\left[\bar{\gamma}(s) - \bar{\gamma}(u) \geq b\right] &\leq \exp\left[-\frac{b}{2} + \log\mathbf{E}\left[\exp\left(\frac{1}{2}\left(\gamma(s) - \gamma(u)\right) + \frac{1}{2}\mathbf{E}\left[\gamma(u) - \gamma(s)\right]\right)\right]\right] \\
&\leq \exp\left[-\frac{b}{2} + \frac{1}{2}K(s,u) + \log\mathbf{E}\left[\exp\left(-\frac{1}{2}\sum_t \log\mathbf{P}_s(X_t = Y_t) + \log\mathbf{P}_u(X_t = Y_t)\right)\right]\right] \\
&\leq \exp\left[-\frac{b}{2} + \frac{1}{2}K(s,u) + \sum_t \log\mathbf{E}\sqrt{\frac{\mathbf{P}_u(X_t = Y_t)}{\mathbf{P}_s(X_t = Y_t)}}\right] \\
&\leq \exp\left[-\frac{b}{2} + \frac{1}{2}K(s,u) + \sum_t \mathbf{E}\sqrt{\frac{\mathbf{P}_u(X_t = Y_t)}{\mathbf{P}_s(X_t = Y_t)}} - n\right] \\
&\leq \exp\left[-\frac{b}{2} + \frac{1}{2}K(s,u) - h^2(s,u)\right]
\end{aligned}$$

where $\mathbf{P}_s = \mathbf{P}$ denote the probability under the distribution s. Thus

$$\mathbf{P}\left[\bar{\gamma}(s) - \bar{\gamma}(u) \geq K(s,u) - 2h^2(s,u) + 2x\right] \leq \mathrm{e}^{-x}.$$

Acknowledgements. The authors wish to thank Stéphane Robin for more than helpful discussions on the statistical aspect and Gavin Sherlock for his insight on the biological applications. The authors also wish to thank the anonymous referee for his helpful suggestions.

References

[1] H. Akaike, Information Theory and Extension of the Maximum Likelihood Principle. *Second int. Symp. Inf. Theory* (1973) 267–281.

[2] N. Akakpo, Estimating a discrete distribution via histogram selection. *ESAIM: PS* **15** (2011) 1–29.

[3] S. Arlot and P. Massart, Data-driven calibration of penalties for least-squares regression. *J. Mach. Learn. Res.* **10** (2009) 245–279.

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

https://doi.org/10.1051/ps/2014005 Published online by Cambridge University Press

[5] A. Barron, L. Birgé and P. Massart, Risk bounds for model selection via penalization. *Probab. Theory Relat. Fields* **113** (1999) 301–413.
[6] C. Biernacki, G. Celeux, G. Govaert, Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **22** (2000) 719–725.
[7] L. Birgé, Model selection for Poisson processes. In Asymptotics: particles, processes and inverse problems, Vol. 55 of *IMS Lect. Notes Monogr. Ser.. Beachwood, OH: Inst. Math. Statist.* (2007) 32–64.
[8] L. Birgé and P. Massart, From model selection to adaptive estimation, in *Festschrift for Lucien Le Cam*. New York, Springer (1997) 55–87.
[9] L. Birgé and P. Massart, Gaussian model selection. *J. Eur. Math. Soc.* **3** (2001) 203–268.
[10] L. Birgé and P. Massart, Minimal penalties for Gaussian model selection. *Probab. Theory Relat. Fields* (2007) **138** 33–73.
[11] J.V. Braun, R. Braun and H.G. Müller, Multiple changepoint fitting via quasilikelihood, with application to DNA sequence segmentation. *Biometrika* **87** (2000) 301–314.
[12] J.V. Braun, H.G. Muller, Statistical methods for DNA sequence segmentation. *Stat. Sci.* (1998) 142–162.
[13] Breiman, Friedman, Olshen, Stone: Classification and Regression Trees. Wadsworth and Brooks (1984).
[14] G. Castellan, Modified Akaikes criterion for histogram density estimation. Technical Report#9961 (1999).
[15] A. Cleynen, M. Koskas, E. Lebarbier, G. Rigaill and S. Robin, Segmentor3IsBack, an R package for the fast and exact segmentation of Seq-data. Algorithms for Molecular Biology (2014)
[16] N. Johnson, A. Kemp and S. Kotz, Univariate Discrete Distributions. John Wiley & Sons, Inc. (2005).
[17] R. Killick and I.A. Eckley, Changepoint: an R package for changepoint analysis. Lancaster University (2011).
[18] E. Lebarbier, Detecting multiple change-points in the mean of Gaussian process by model selection. *Signal Process.* **85** (2005) 717–736.
[19] T.M. Luong, Y. Rozenholc and G. Nuel, Fast estimation of posterior probabilities in change-point analysis through a constrained hidden Markov model. *Comput. Stat. Data Anal.* (2013).
[20] P. Massart, Concentration inequalities and model selection. In *Lect. Notes Math.* Springer Berlin/Heidelberg (2007).
[21] P. Reynaud-Bouret, Adaptive estimation of the intensity of inhomogeneous Poisson processes via concentration inequalities. *Probab. Theory Relat. Fields* **126** (2003) 103–153.
[22] G. Rigaill, Pruned dynamic programming for optimal multiple change-point detection. `ArXiv:1004.0887` 2010, [http://arxiv.org/abs/1004.0887].
[23] G. Rigaill, E. Lebarbier and S. Robin, Exact posterior distributions and model selection criteria for multiple change-point detection problems. *Stat. Comput.* **22** (2012) 917–929.
[24] D. Risso, K. Schwartz, G. Sherlock and S. Dudoit, GC-Content Normalization for RNA-Seq Data. *BMC Bioinform.* **12** (2011) 480.
[25] Y.C. Yao, Estimating the number of change-points via Schwarz' criterion. *Stat. Probab. Lett.* **6** (1988) 181–189.
[26] N.R. Zhang and D.O. Siegmund, A modified Bayes information criterion with applications to the analysis of comparative genomic hybridization data. *Biometrics* **63** (2007) 22–32.

https://doi.org/10.1051/ps/2014005 Published online by Cambridge University Press