Stat Comput
DOI 10.1007/s11222-014-9492-y

Comparing change-point location in independent series

A. Cleynen · S. Robin

Received: 25 March 2014 / Accepted: 30 June 2014

Abstract We are interested in the comparison of the positions of the change-points in the segmentation of independent series. We consider a Bayesian framework with conjugate priors to perform exact inference on the change-point model. This work is motivated by the comparison of transcript boundaries in yeast grown under different conditions. When comparing two series, we derive the posterior credibility interval of the shift between the locations. When comparing more than two series, we compute the posterior probability for a given change-point to have the same location in all series. All calculations are made in an exact manner in a quadratic time. The performances of those approaches are assessed via a simulation study. When applied to yeast genes, this approach reveals different behavior between internal and external exon boundaries.

Keywords Segmentation · Change-point comparison · Credibility intervals · Bayesian inference · Negative binomial

1 Introduction

Segmentation problems arise in a large range of domains such as economy, biology or meteorology, to name a few. Many methods have been developed and proposed in the literature in the last decades to detect change-points in the distribution of the signal along one single series. Yet, more and more applications require the analysis of several series at a time to better understand a complex underlying phenomenon. Such situations refer for example to the analysis of meteorological series observed in different locations (Ehsanzadeh et al. 2011), of the genomic profiles of a cohort of patients (Picard et al. 2011), or of sets of astronomical series of photons abundance (Dobigeon et al. 2007).

When dealing with multiple series, two approaches can be typically considered. The first consists in the *simultaneous* segmentation of all series, looking for changes that are common to all of them. This approach amounts to the segmentation of one single multivariate series but might permit the detection of change-points in series with too low a signal to allow their detection independently. The second approach consists in the *joint* segmentation of all the series, each having its specific number and location of changes. This allows to account for dependence between the series without imposing that the changes occur simultaneously.

We are interested here in a third kind of statistical problem, which is the comparison of the locations of the change-points between several series that have been observed independently. To our knowledge, this problem has not yet been fully addressed. In this framework, we consider series that have been observed along the same period of time, but not necessarily simultaneously, so that they can be assumed to be independent. This assumption typically makes sense when 'time' actually refers to some non temporal one-dimensional organization of the data, such as the genome location in our motivating example, or when interested in the location of copy number variation boundaries. It is also valid for neuroscience experiments where patients are independently submitted to a sequence of stimuli with same time intervals, like in the application presented in Wager et al. (2009). In this case, the question is to know if all patients react to a given stimulus with the same delay. Similar situations may for instance happen when studying abrupt changes in the

A. Cleynen · S. Robin (✉)
AgroParisTech, MIA 518, Paris, France
e-mail: Stephane.Robin@agroparistech.fr

A. Cleynen · S. Robin
INRA, MIA 518, Paris, France

Published online: 24 September 2014

stream flow of a river over a year and wondering if these changes occur at the same time every year, etc. In such contexts, change-point comparison makes sense, although the series at hand are independent.

The comparison of change-points is connected to the evaluation of the uncertainty of their positions. An important point is that the standard likelihood-based inference is very intricate, since the required regularity conditions for the change-point parameters are not satisfied (Feder 1975). Most methods to obtain confidence intervals on the locations of the change-points are based on their limit distribution estimators (Feder 1975; Bai and Perron 2003) or on the asymptotic use of a likelihood-ratio statistic (Muggeo 2003). Bootstrap techniques have also been proposed (see Hukov and Kirch (2008) and references therein). Comparison studies of some of these methods can be found in Reeves et al. (2007) for climate applications or in Toms and Lesperance (2003) for ecology. Recently, Rigaill et al. (2012) proposed a Bayesian framework to derive the exact posterior distributions of various quantities of interest—including change-point locations—in the context of exponential family distributions with conjugate priors.

Our motivation comes from the comparison of the transcription profiles of yeast genes, when grown under different conditions. Indeed, in complex organisms such as human, the transcribed regions are known to vary according to external conditions (Tian et al. 2005). Based on transcriptional data observed along the genome, we apply our methodology to detect genes for which such boundary-shift occurs.

Contribution In this paper we develop a Bayesian approach to compare the positions of the change-points in the segmentation of independent series. In Sect. 2, we recall the Bayesian segmentation model introduced in Rigaill et al. (2012) and its adaptation to our framework. In Sect. 3 we derive the posterior distribution of the shift between the positions of the change-points in two independent series, while in Sect. 4 we introduce the calculation of the posterior probability for change-points to share the same location in different series. The performances are assessed in Sect. 5 via a simulation study on count data. We finally apply the proposed methodology to study the existence of differential splicing in yeast in Sect. 6. Our approach is implemented in an R package `EBS` which is available on the CRAN repository.

All the results we provide are given conditional on the number of segments in each profile. Indeed comparing the location of, say, the second change-point in each series implicitly refers to a total number of change-points in each of them. Yet, we provide in the package two criteria for choosing this number (namely the Bayesian Information Criterion, BIC, and the Integrated Completed Likelihood, ICL) and most of the results we provide can be marginalized over the number of segments. This aspect is not discussed in this paper as the interpretation of the comparison of change-points when integrating over possible number of segments is delicate.

2 Model for one series

We now introduce the general Bayesian framework for the segmentation of one series and recall preceding results on the posterior distribution of change-points.

2.1 Bayesian framework for one series

The general segmentation problem consists in partitioning a signal of n data-points $\{y_t\}_{t\in[\![1,n]\!]}$ into K adjacent segments. The model is defined as follows: the observed data $\{y_t\}_{t\in[\![1,n]\!]}$ are supposed to be a realization of an independent random process $Y = \{Y_t\}_{t=1,\dots,n}$. This process is drawn from a probability distribution $\mathcal{G}$ which depends on a set of parameters among which one parameter θ is assumed to be affected by $K-1$ abrupt changes, called change-points and denoted τ_k $(1 \le k \le K-1)$. A partition m is defined as a set of change-points: $m = (\tau_0, \tau_1, \dots, \tau_K)$ with conventions $\tau_0 = 1$ and $\tau_K = n+1$ and a segment J is said to belong to m if $J = [\![\tau_{k-1}; \tau_k[\![$ for some k.

The Bayesian model is fully specified with the following distributions:

- The prior distribution of the number of segments $P(K)$;
- The conditional distribution of partition m given K: $P(m|K)$;
- The parameters θ_J for each segment J from m are supposed to be independent with same distribution $P(\theta_J)$;
- The observed data $Y = (Y_t)$ data are independent conditional on m and (θ_J) with distribution depending on the segment:

$$(Y_t|m, J \in m, \theta_J, t \in J) \sim \mathcal{G}(\theta_J).$$

We will further denote $Y_J = (Y_{\tau_{k-1}}, \dots, Y_{\tau_k-1})$ the vector of random variables Y belonging to segment J defined by $[\![\tau_{k-1}; \tau_k[\![$. As a consequence, we denote $P(Y_J|\theta_J) = \prod_{t\in J} g(Y_t; \theta_J)$ where $g(\cdot; \theta_J)$ stands for the pdf of the distribution $\mathcal{G}$.

2.2 Exact calculation of posterior distributions

Rigaill et al. (2012) show that if distribution $\mathcal{G}$ possesses conjugate priors for θ_J, and if the model satisfies the factorability assumption, that is, if

$$P(Y, m) = C \prod_{J\in m} a_J P(Y_J|J),$$

$$\text{where} \qquad P(Y_J|J) = \int P(Y_J|\theta_J) P(\theta_J) d\theta_J, \tag{1}$$

C is a normalization constant and a_J are segment-specific normalization constants imposed by $P(m|K)$, quantities such as $P(Y, K)$, the posterior distributions of the location of the change-points or the posterior entropy can be computed exactly and in quadratic time. Examples of satisfying distributions are

- The Gaussian heteroscedastic:

$$\mathcal{G}(\theta_J) = \mathcal{N}(\mu_J, \sigma_J^2) \text{ with } \theta_J = (\mu_J, \sigma_J^2)$$

- The Gaussian homoscedastic with known variance σ^2:

$$\mathcal{G}(\theta_J) = \mathcal{N}(\mu_J, \sigma^2) \text{ with } \theta_J = \mu_J,$$

- The Poisson:

$$\mathcal{G}(\theta_J) = \mathcal{P}(\lambda_J) \text{ with } \theta_J = \lambda_J$$

- Or the negative binomial homoscedastic with known dispersion ϕ:

$$\mathcal{G}(\theta_J) = \mathcal{NB}(p_J, \phi) \text{ with } \theta_J = p_J$$

Note that the Gaussian homoscedastic does not satisfy the factorability assumption if σ is unknown, and that the negative binomial heteroscedastic does not belong to the exponential family when ϕ is unknown. In both cases, all computations such as that of Eq. 1 or the approaches proposed in the next sections are done conditional on the value of the known parameter σ or ϕ. Note that, $\mathcal{G}(\theta_J)$ should be denoted as $\mathcal{G}(\theta_J, \phi)$, and idem for $g(\cdot; \theta_J, \phi)$ and $P(Y_J|\theta_J, \phi)$. The parameter ϕ will be dropped in the sequel for the sake of readability.

The factorability assumption (1) also induces some constraints on the distribution of the segmentation $P(m|K)$. In this paper, we will limit ourselves to the uniform prior:

$$P(m|K) = \mathcal{U}\left(\mathcal{M}_K^{1,n+1}\right)$$

where $\mathcal{M}_K^{1,n+1}$ stands for the set of all possible partitions of $[\![1, n+1[\![$ into K non-empty segments. Note that the marginal distribution of m (after integration over K) would not be uniform in general.

3 Posterior distribution of the shift

The framework described above allows to compute a set of quantities of interest in an exact manner. In this paper, we are mostly interested in the location of change-points. We first remind how posterior distributions can be computed and then propose a first exact comparison strategy.

3.1 Posterior distribution of the change-points

The key ingredient for most of the calculations is the $(n+1) \times (n+1)$ upper triangular matrix A that contains the probabilities of all segments. Specifically, denoting $[A]_{i,j}$ the generic elements of matrix A, we have $[A]_{i,j} = 0$ if $i \geq j$ and

$$\forall 1 \leq i < j \leq n+1, \qquad [A]_{i,j} = P(Y_{[\![i,j[\![}|[\![i, j[\![) \tag{2}$$

where $P(Y_J|J)$ is given in (1).

The posterior distribution of change-points can be deduced from this matrix in a quadratic time with the following proposition:

Proposition 1 *Denoting $p_k(t; Y; K) = P(\tau_k = t|Y, K)$ the posterior distribution of the kth change-point, we have*

$$p_k(t; Y; K) = \frac{\left[(A)^k\right]_{1,t}\left[(A)^{K-k}\right]_{t,n+1}}{\left[(A)^K\right]_{1,n+1}}.$$

where $(A)^k$ denotes the kth power of matrix A and $\left[(A)^k\right]_{i,j}$ its generic term.

This result and the proof we give here are a reformulation of a result which can be found in Rigaill et al. (2012).

Proof We have

$$p_k(t; Y; K) = \frac{\sum_{m \in \mathcal{B}_{K,k}(t)} p(Y|m)p(m|K)}{P(Y|K)}$$

where $\mathcal{B}_{K,k}(t)$ is the set of partitions of $\{1, \ldots, n\}$ in K segments with kth change-point at location t. Note that $\mathcal{B}_{K,k}(t) = \mathcal{M}_k^{1,t} \otimes \mathcal{M}_{K-k}^{t,n+1}$ (i.e. all $m \in \mathcal{B}_{K,k}(t)$ can be decomposed uniquely as $m = m_1 \cup m_2$ with $m_1 \in \mathcal{M}_k^{1,t}$ and $m_2 \in \mathcal{M}_{K-k}^{t,n+1}$ and reciprocally). Then using the factorability assumption, we can write

$$p_k(t; Y; K) = \frac{\sum_{m_1 \in \mathcal{M}_k^{1,t}} p(Y|m_1) \sum_{m_2 \in \mathcal{M}_{K-k}^{t,n+1}} p(Y|m_2)\ p(m|K)}{\sum_{m \in \mathcal{M}_K^{1,n+1}} p(Y|m)\ p(m|K)}$$

where the term $p(m|K)$ vanishes thanks to the uniform prior and each sum can be computed using the results of Rigaill et al. (2012), which amounts at computing the k-th power of matrix A. □

3.2 Comparison of two series

We now propose a first procedure to compare the location of two change-points in two independent series. Consider two independent series Y^1 and Y^2 with same length n and respective number of segments K^1 and K^2. The aim is to compare the locations of the k_1th change-point from series Y^1 (denoted $\tau_{k_1}^1$) with the k_2th change-point from series Y^2 (denoted $\tau_{k_2}^2$). The posterior distribution of the difference between the locations of the two change-points, which we will call the shift, can be derived with the following Proposition.

Proposition 2 *Denoting* Δ *the shift* $\tau^1_{k_1} - \tau^2_{k_2}$ *and* $\delta_{k_1,k_2}(d; K^1, K^2) = P(\Delta = d|Y^1, Y^2, K^1, K^2)$ *its posterior distribution, we have*

$$\delta_{k_1,k_2}(d; K^1, K^2) = \sum_t p_{k_1}(t; Y^1; K^1) p_{k_2}(t-d; Y^2; K^2).$$

Proof This simply results from the convolution between the two posterior distributions p_{k_1} and p_{k_2}. □

The posterior distribution of the shift can therefore be computed exactly and in a quadratic time. The non-difference between the positions of the two change-points $\tau^1_{k_1}$ and $\tau^2_{k_2}$ can then be assessed, looking at the position of 0 with respect to the posterior distribution δ. Examples of such credibility intervals are given in Fig. 5.

4 Comparison of the locations of change-points

We now consider the comparison of the locations of change-points between more than 2 series. In this case, the convolution method described above does not apply anymore so we propose a comparison based on the exact computation of the posterior probability for the change-points under study to have the same location.

4.1 Model for I series

We now consider I independent series Y^ℓ (with $1 \le \ell \le I$) with same length n. We denote m^ℓ, their respective partitions and K^ℓ their respective number of segments. We further denote τ^ℓ_k the kth change-point in Y^ℓ so that $m^\ell = (\tau^\ell_0, \tau^\ell_1, \ldots, \tau^\ell_{K^\ell})$. Similarly, θ^ℓ_J denotes the parameter for the series ℓ within segment J provided that J belongs to m^ℓ, and ϕ^ℓ denotes the constant parameter of series ℓ. In the following, the set of series will be referred to as $\mathbf{Y}$ and respectively for the vector of segment numbers ($\mathbf{K}$), the set of all partitions ($\mathbf{m}$) and the set of all parameters ($\boldsymbol{\theta}$).

In the perspective of change-point comparison, for a given set of change-point indexes, we introduce the following event:

$$E_0 = E_0(k_1, \ldots, k_I) := \{\tau^1_{k_1} = \cdots = \tau^I_{k_I}\}.$$

We further denote E_1 its complementary and define the binary random variable

$$E = \mathbb{I}\{E_1\} = 1 - \mathbb{I}\{E_0\}.$$

The complete hierarchical model is displayed in Fig. 1 and is defined as follows:

- The random variable E is drawn conditionally on $\mathbf{K}$ as a Bernoulli $\mathcal{B}(1 - p_0(\mathbf{K}))$ where $p_0(\mathbf{K}) = P(E_0|\mathbf{K})$;

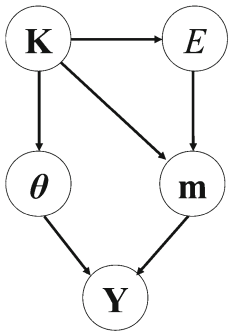

Fig. 1 Graphical model. Hierarchical model for the comparison of I series

- The parameters $\boldsymbol{\theta}$ are drawn independently according to $P(\boldsymbol{\theta}|\mathbf{K})$;
- The partitions are drawn conditionally on E according to $P(\mathbf{m}|\mathbf{K}, E)$;
- The observations are generated according to the conditional distribution $P(\mathbf{Y}|\mathbf{m}, \boldsymbol{\theta})$.

More specifically, denoting $\mathcal{M}^{1,n+1}_{\mathbf{K}} = \bigotimes_\ell \mathcal{M}^{1,n+1}_{K^\ell}$, the partitions are assumed to be uniformly distributed, conditional on E, that is

$$P(\mathbf{m}|\mathbf{K}, E_0) = \mathcal{U}(\mathcal{M}^{1,n+1}_{\mathbf{K}} \cap E_0), \quad P(\mathbf{m}|\mathbf{K}, E_1) = \mathcal{U}(\mathcal{M}^{1,n+1}_{\mathbf{K}} \cap E_1).$$

4.2 Posterior probability for the existence of a common change-point

We propose to assess the existence of a common change-point location between the I series based on the posterior probability of this event, namely $P(E_0|\mathbf{Y}, \mathbf{K})$.

Proposition 3 *The posterior probability of* E_0 *can be computed in* $O(Kn^2)$ *as*

$$P(E_0|\mathbf{Y}, \mathbf{K}) = \frac{p_0(\mathbf{K})}{q_0(\mathbf{K})} Q(\mathbf{Y}, E_0|\mathbf{K}) \Big/ \left[\frac{p_0(\mathbf{K})}{q_0(\mathbf{K})} Q(\mathbf{Y}, E_0|\mathbf{K}) + \frac{1 - p_0(\mathbf{K})}{1 - q_0(\mathbf{K})} Q(\mathbf{Y}, E_1|\mathbf{K})\right]$$

where

$$Q(\mathbf{Y}, E_0|\mathbf{K}) = \sum_t \prod_\ell \left[(A_\ell)^{k_\ell}\right]_{1,t} \left[(A_\ell)^{K_\ell - k_\ell}\right]_{t+1,n+1},$$

$$Q(\mathbf{Y}, E_1|\mathbf{K}) = \prod_\ell \left[(A_\ell)^{K_\ell}\right]_{1,n+1} - Q(\mathbf{Y}, E_0|\mathbf{K}),$$

and

$$q_0(\mathbf{K}) = Q(E_0|\mathbf{K}) = \sum_t \prod_\ell \binom{t-2}{k_\ell - 1} \binom{n-t}{K_\ell - k_\ell - 1} \Big/ \binom{n-1}{K_\ell - 1}.$$

and A_ℓ *stands for the matrix* A *as defined in (2), corresponding to series* ℓ.

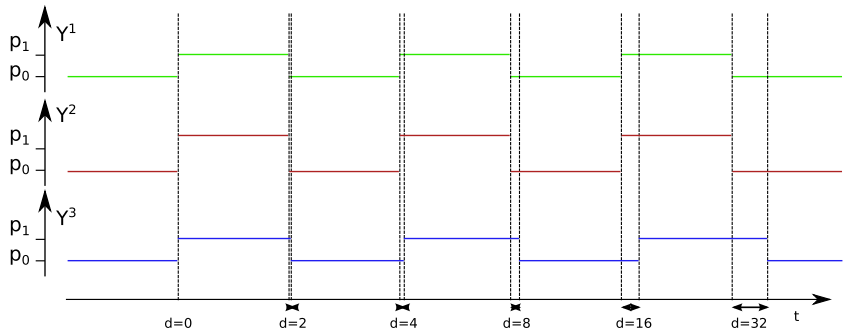

Fig. 2 Simulation design. Representation of the parameter values along time: each profile is simulated with 7 segments using same parameter value, but the location of the change-points is increasingly shifted in series Y^3 by value d

Proof We consider the surrogate model where the partition $\mathbf{m}$ is drawn uniformly and independently from E, namely $Q(\mathbf{m}|\mathbf{K}) = \mathcal{U}(\mathcal{M}_{\mathbf{K}}^{1,n+1})$ (note that this corresponds to choosing $p_0(\mathbf{K}) = q_0(\mathbf{K})$). All probability distributions under this model are denoted by Q along the proof and their formulas derive from Rigaill et al. (2012). For instance, we obtain

$$
\begin{aligned}
Q(\mathbf{Y}|\mathbf{K}) &= \prod_{\ell} \left[(A_\ell)^{K_\ell} \right]_{1,n+1} \text{ and} \\
Q(\mathbf{Y}, E_0|\mathbf{K}) &= \sum_{t} Q(\mathbf{Y}, \tau_{k_1}^1 = \cdots = \tau_{k_I}^I = t|\mathbf{K}) \\
&= \sum_{t} \prod_{\ell} Q\left(Y^\ell, \tau_{k_\ell}^\ell = t|K^\ell\right) \\
&= \sum_{t} \prod_{\ell} \left[(A_\ell)^{k_\ell} \right]_{1,t} \left[(A_\ell)^{K_\ell - k_\ell} \right]_{t+1,n+1},
\end{aligned}
$$

It then suffices to apply the probability change as

$$
\begin{aligned}
P(\mathbf{Y}, E_0|\mathbf{K}) &= \frac{p_0(\mathbf{K})}{q_0(\mathbf{K})} Q(\mathbf{Y}, E_0|\mathbf{K}), \\
P(\mathbf{Y}, E_1|\mathbf{K}) &= \frac{1 - p_0(\mathbf{K})}{1 - q_0(\mathbf{K})} Q(\mathbf{Y}, E_1|\mathbf{K}),
\end{aligned}
$$

where $Q(Y, E_1|K) = Q(Y|K) - Q(Y, E_0|K)$. The result then follows from the decomposition of $P(\mathbf{Y}|\mathbf{K})$ as $P(\mathbf{Y}, E_0|\mathbf{K}) + P(\mathbf{Y}, E_1|\mathbf{K})$. □

The Bayes factor is sometimes preferred for model comparison; it can be computed exactly in a similar way:

Corollary 4 *The Bayes factor can be computed in* $O(Kn^2)$ *as*

$$
\frac{P(\mathbf{Y}|E_0, \mathbf{K})}{P(\mathbf{Y}|E_1, \mathbf{K})} = \frac{1 - q_0(\mathbf{K})}{q_0(\mathbf{K})} \frac{Q(\mathbf{Y}, E_0|\mathbf{K})}{Q(\mathbf{Y}, E_1|\mathbf{K})}
$$

using the same notations as in Proposition 3.

Proof The proof follows this of Proposition 3. □

All probabilities are evaluated conditionally on $\mathbf{K}$. We remind that all the equivalent marginal probabilities can be evaluated with the same complexity. However, dealing with such marginal probabilities would raise interpretation issues.

5 Simulation study

5.1 Simulation design

We designed a simulation study to identify the influence of various parameters on the performances of our approaches. The design is illustrated in Fig. 2: we compared 3 independent series with 7 segments, with all odd (respectively even) segments sharing the same distribution. The first two series share the same segmentation given by $m = (1, 101, 201, 301, 401, 501, 601, 701)$ and the change-point locations of the third one are progressively shifted apart as $\tau_k^3 = \tau_k^1 + 2^{k-1}$, for each $1 \leq k \leq 6$. We shall denote $d_k = \tau_k^3 - \tau_k^1$ and drop the index k when there is no ambiguity on it.

In view of the application presented in the next section, we focused on the negative binomial distribution. Our purpose is to mimic data obtained by sequence-based transcriptomic experiments, so that the parameters for the negative binomial distribution were chosen to fit typical real-data. Considering the model where odd segments are sampled with distribution $\mathcal{NB}(p_0, \phi)$, and even with $\mathcal{NB}(p_1, \phi)$, we chose two different values of p_0, 0.8 and 0.5, and for each of them, we made p_1 vary so that the odds ratio $s := p_1/(1-p_1)/[p_0/(1-p_0)]$ is 4, 8 and 16. Finally, we used different values of ϕ as detailed in Table 1 in order to explore a wide range of possible dispersions while keeping a signal/noise ratio not too high. Note that the higher ϕ, the less overdispersed the signal.

In practice there is little chance that the overdispersion is known. We propose to estimate this parameter from the data

Table 1 Values of parameters used in the simulation study

$p_0 = 0.8$		$p_0 = 0.5$	
p_1	ϕ	p_1	ϕ
0.5	5	0.2	$0.08^{1/8}$
0.33	$\sqrt{5}$	0.1	$0.08^{1/4}$
0.2	0.8	0.05	$0.08^{1/2}$
	0.64		0.08

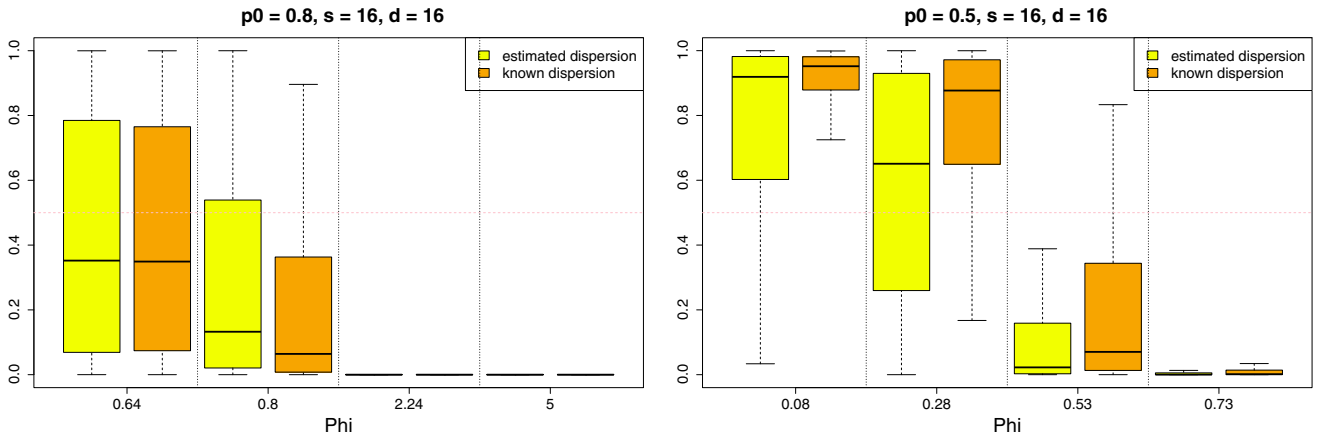

Fig. 3 Impact of estimating the dispersion parameter. *Boxplot* of the posterior probability of E_0 for $s = 16$ and $d = 16$ when estimating the value of ϕ (*left boxplot* of each subdivision) or when using the known value (*right boxplot* of each subdivision)

and use the obtained value in the analysis. For the simulation study, we used the estimator inspired from Johnson et al. (2005): starting from sliding window of size 15, we compute the method of moments estimator of ϕ, using the formula $\phi = E^2(X)/(V(X) - E(X))$, and retain the median over all windows. When this median is negative (which is likely to happen in datasets with many zeros), we double the size of the window. In practice however, results are very similar when using maximum likelihood or quasi-maximum likelihood estimators on sliding windows. An empirical Bayes estimator can also be considered, but it requires several evaluation of the likelihood function and can therefore not be used in an extensive simulation study. This estimator will be presented in the next section in the context of the yeast application.

5.2 Results

We compute the posterior probability $P(E_0|\mathbf{Y}, \mathbf{K})$ for each simulation and each value of d. Figures 7 and 8 in Appendix 1 represent the boxplots of this probability across all simulations for each configuration. Hence, the variability depicted by each boxplot comes from the sampling of different $\mathbf{Y}$. Note that in each figure, the first boxplot corresponds to $d = 0$ and thus to model E_0, while $d \neq 0$ for other boxplots so that the true model is E_1. These plots can be understood as abacus for the detection power of the proposed approach. For example, the ideal situation corresponds to $s = 16$ in Fig. 9 obtained in a similar simulation study performed with the Poisson distribution and presented in "Appendix" 2, where $p(E_0|\mathbf{Y}, \mathbf{K})$ is always one for $d = 0$ and zero otherwise.

As expected, these results show that the lower the value of ϕ, the more difficult the decision becomes. The trend is identical for decreasing values of the odd-ratio s and decreasing values of d. In the most difficult scenario of very high dispersion compared to signal value, the method fails to provide satisfying decisions whatever the level of odds ratio or distance between change-points. However, in most configurations, the method is adequate as soon as $d \geq 16$.

An important question is the impact of the estimation of the dispersion parameter. Interestingly, in the simulation study with $p_0 = 0.8$, our estimator tended to underestimate ϕ (and thus over-estimate the dispersion) while it was the contrary in the simulation study with $p_0 = 0.5$. This affects the performance of the decision rule, which behaves better when ϕ is higher. For instance, Fig. 3 shows, for $s = 16$ and $d = 16$, that knowing the true value of ϕ improves the results when $p_0 = 0.8$ but worsens them when $p_0 = 0.5$.

6 Comparison of transcribed regions in yeast

Differential splicing in yeast Differential splicing is one of the mechanism that living cells use to modify the transcription of their genome in response to some change in their environment, such as a stress. More precisely, differential splicing refers to the ability for the cell to choose between versions (called isoforms) of a given gene by changing the boundaries of the regions to be transcribed.

New sequencing technologies give access to a measure of the transcription at the nucleotide resolution. When applied to the set of transcripts present in a cell, the signal provided by these technologies consists in a count associated to each nucleotide along the genome, which is proportional to the transcription level of the nucleotide. More precisely, Y_t corresponds to the number of aligned reads which first nucleotide aligns to position t of the genome. This technology therefore allows to locate precisely the boundaries of the transcribed regions, to possibly revise the known annotation of the genomes and to study the variation of these boundaries across conditions.

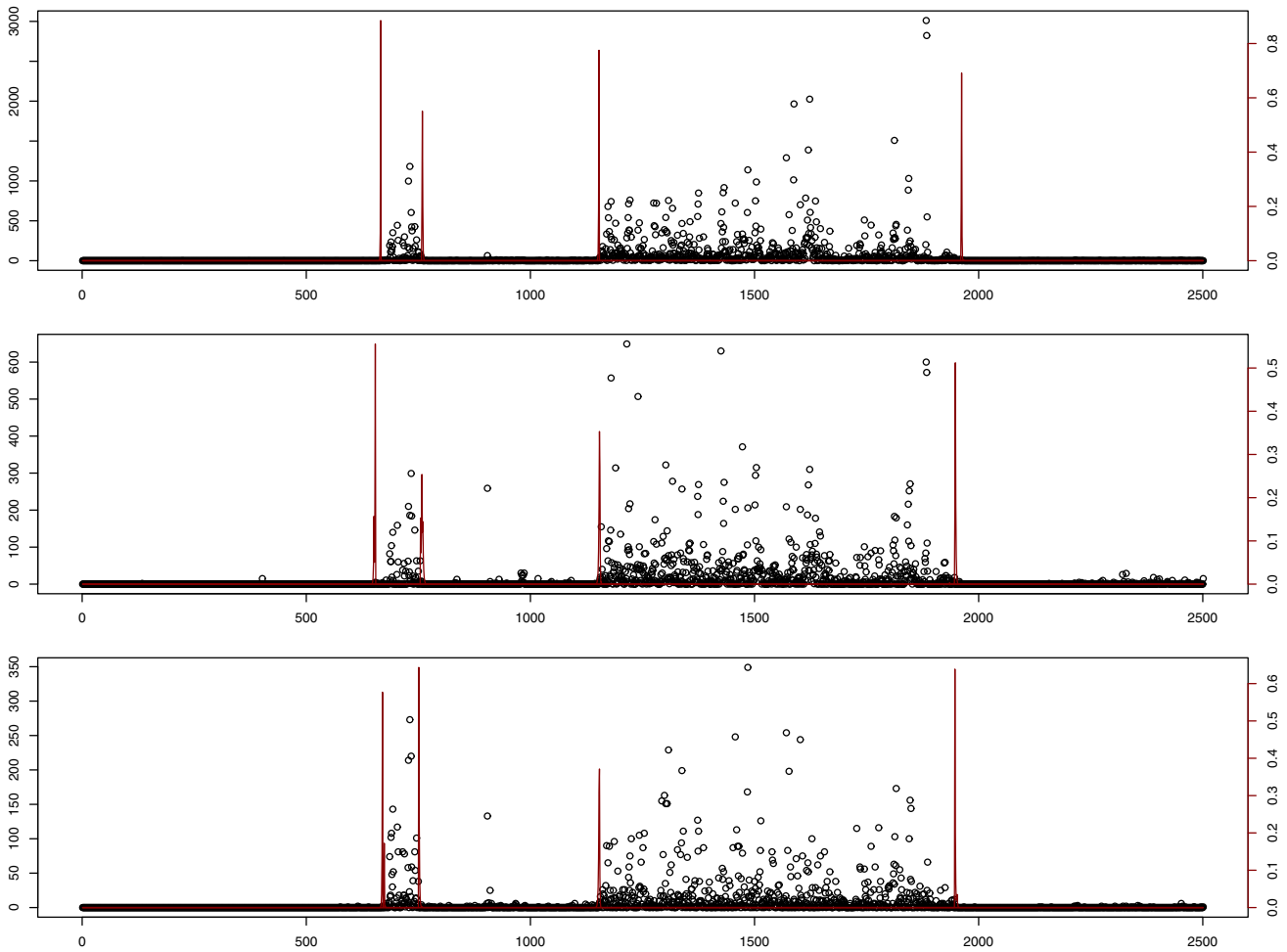

Fig. 4 Posterior distribution of change-point location. Segmentation in 5 segments of gene YAL013W in three different media: ypd (*top*), glycerol (*middle*) and delft (*bottom*). *Black dots* represent the number of reads starting at each position of the genome (*left scale*) while *red curves* are the posterior distribution of the change-point location (*right scale*). (Color figure online)

Experimental design. In this example we consider a study from the Sherlock lab in Stanford (Risso et al. 2011) on a yeast strain, *Saccharomyce Cerevisiae*, grown in three different environments: ypd, which is the traditional (rich) media for yeast, delft, a similar but poorer media, and glycerol. In the last decade many studies [see for instance Proudfoot et al. (2002); Tian et al. (2005)] have showed that a large proportion of genes can express multiple transcripts with different right-end position. Similarly, the 5' capping process is dependent on environment conditions (Mandal et al. 2004), and the left-end position of the transcript may vary according to stress factors. We may therefore expect that the yeast cells grown in different conditions (they ferment in the first two media, while they respire in glycerol) will produce transcripts with different boundaries.

In this motivating study, we are interested in the differential splicing mechanism in yeast, and whether it differs in extremal and internal boundaries. To this effect, we focused on genes for which the region to be transcribed is not contiguous on the genome, but shattered into so-called exons (the regions separating the exons being called introns).

Change-point locations We applied our procedure using the negative binomial distribution to gene YAL003W which has two exons. The corresponding series were segmented into 5 segments to allow one segment per transcribed region separated by segments of non-coding regions. Figure 4 illustrates the posterior distribution of each change-point in each profile.

Estimation of the overdispersion parameter Our methodology requires a prior estimate of the parameter ϕ. For the present study on yeast, we used an empirical Bayes approach. We considered the likelihood function

$$P(\mathbf{Y}|\mathbf{K};\phi) = \sum_{\mathbf{m}\in\otimes_\ell \mathcal{M}_{K_\ell}} P(\mathbf{Y}|\mathbf{m};\phi)P(\mathbf{m}|\mathbf{K}).$$

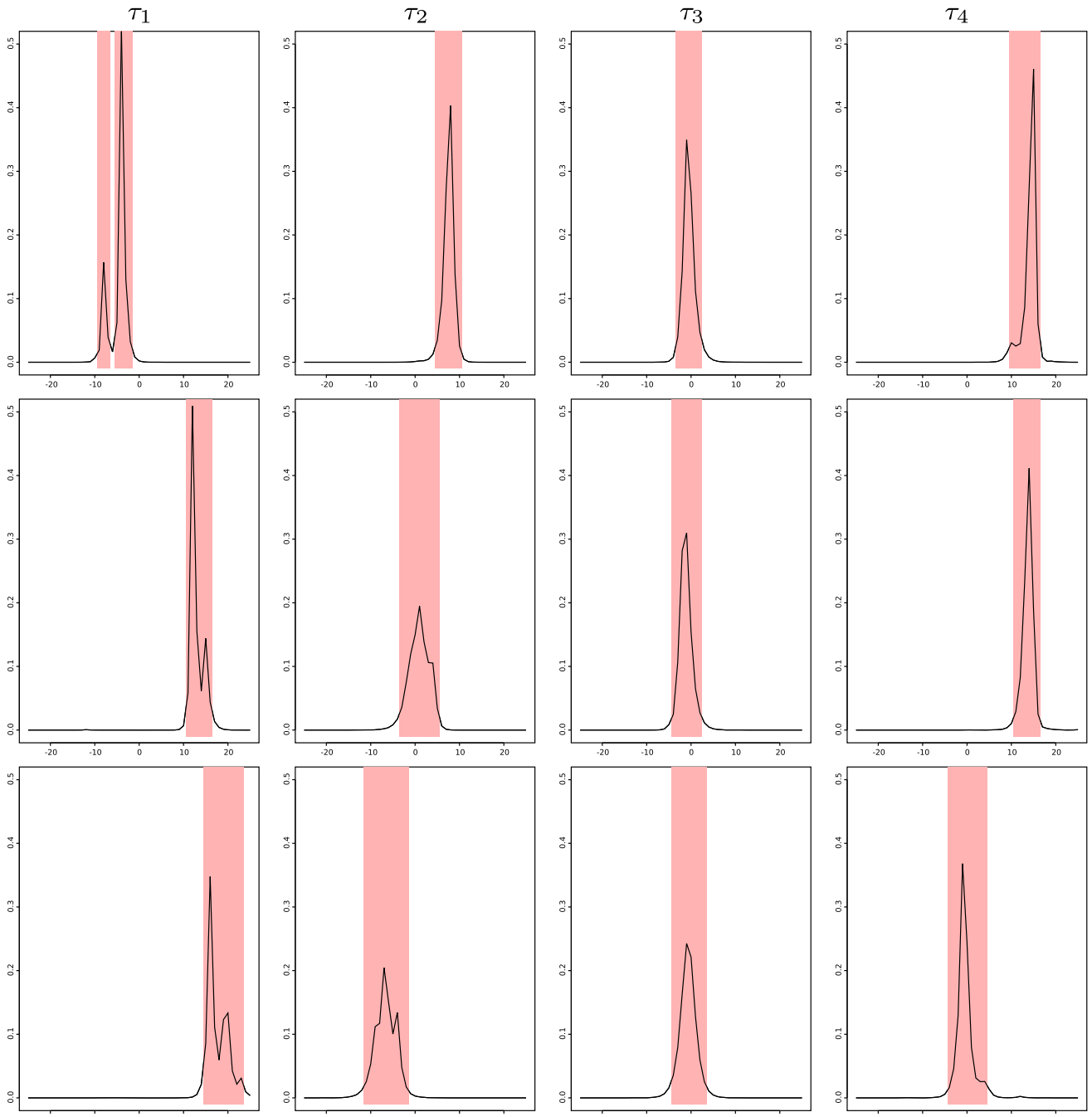

Fig. 5 Distribution of change-point location and 95 % credibility intervals. For each of the two-by-two comparisons (*top*: ypd-delft; *middle*: ypd-glycerol; *bottom* delft-glycerol), posterior distribution of the difference of location between each of the first to the fourth change-points

We chose to estimate ϕ apart from any prior knowledge on E_0, namely keeping its prior probability to $p_0(\mathbf{K}) = q_0(\mathbf{K})$ as defined in Proposition 3. This amounts to assume that segmentations are drawn independently (conditional on $\mathbf{K}$), so we get $P(\mathbf{Y}|\mathbf{K};\phi) = Q(\mathbf{Y}|\mathbf{K};\phi)$ (as defined in the proof of Proposition 3), which can be computed with quadratic complexity for any value of ϕ. An estimate of ϕ can then be chosen as $\widehat{\phi} = \arg\max_\phi P(\mathbf{Y}|\mathbf{K};\phi)$, which consists in a one-dimensional optimization problem that can efficiently be solved using iterative algorithms such as Armijo's (Armijo 1966) which we propose to use in this illustration. Note that such approaches require several evaluations of the likelihood function $\log P(\mathbf{Y}|\mathbf{K};\phi)$ and may raise computational issues for very long series.

Table 2 Posterior probability of a common change point across conditions for gene YAL013W

Comparison	Change-point			
	τ_1	τ_2	τ_3	τ_4
All media	10^{-2}	1	1	$1.9\ 10^{-2}$
Ypd-delft	0.92	0.89	0.99	10^{-2}
Ypd-glycerol	10^{-3}	0.99	0.99	0.62
Delft-glycerol	$2\ 10^{-2}$	0.94	0.99	0.99

Credibility intervals on the shift For each of the first to the fourth change-point, and for each pair of conditions, we computed the posterior distribution of the difference between the locations of the change-points. For the biological reasons stated above, we expect to observe more differences for the first and last change-points than for the other two, which can be used as a verification of the decision rule.

Figure 5 provides the posterior distribution of these differences, as well as the 95 % credibility intervals. It can for instance be noted that even though the locations seemed very similar on Fig. 4, only in a few cases does 0 fall inside the 95 % credibility intervals of the posterior distribution of the shift.

Posterior probability of common change-point. We then computed the probability that the change-point is the same across several series, taking $p_0 = 1/2$. Table 2 provides, for the simultaneous comparison of the three conditions and for each pair of conditions, the value of the posterior probability of E_0 at each change-point (τ_1^ℓ is associated with the left-end of the gene, τ_2^ℓ to the left boundary of intron, τ_3^ℓ to the right boundary of the intron and τ_4^ℓ to the right-end of the gene). Reassuringly, our method confirms that the change-point location is identical when corresponding to intron boundaries. On the contrary, extremal boundaries seem more subject to variations in locations from one condition to another.

Differential splicing in yeast. We finally applied our comparison procedure to a set of 50 genes from the yeast genome which all possess two exons and which were expressed in all three conditions at the time of the experiment. The left figure of Fig. 6 shows the distribution of the posterior probability of E_0 for the simultaneous comparison of the three conditions when $p_0(\mathbf{K}) = 1/2$. Once again the results strengthen the expectation that intron boundaries should not vary between conditions while more difference is observed for the external boundaries of the genes. A closer look at the genes for which we have evidence of either the second or third change-point difference reveals that in almost all of them (5 out of 6) one of the two exons was not expressed in the Glycerol medium.

A discussion with Dr Sherlock suggests that about 10 % of the genes should be liable to differential splicing affecting almost specifically the extremal boundaries. The Bayesian framework allows to account for this prior knowledge through the prior probability p_0. We therefore performed the analysis over again setting $p_0 = 0.9$ for τ_1 and τ_4 and $p_0 = 0.99$ for the other two, and removing the 5 genes with one non-expressed exon. Results are illustrated in the right panel of Fig. 6. For these new prior values, we observe that 5 genes have a left external boundary which varies, and 12 for the right external boundary.

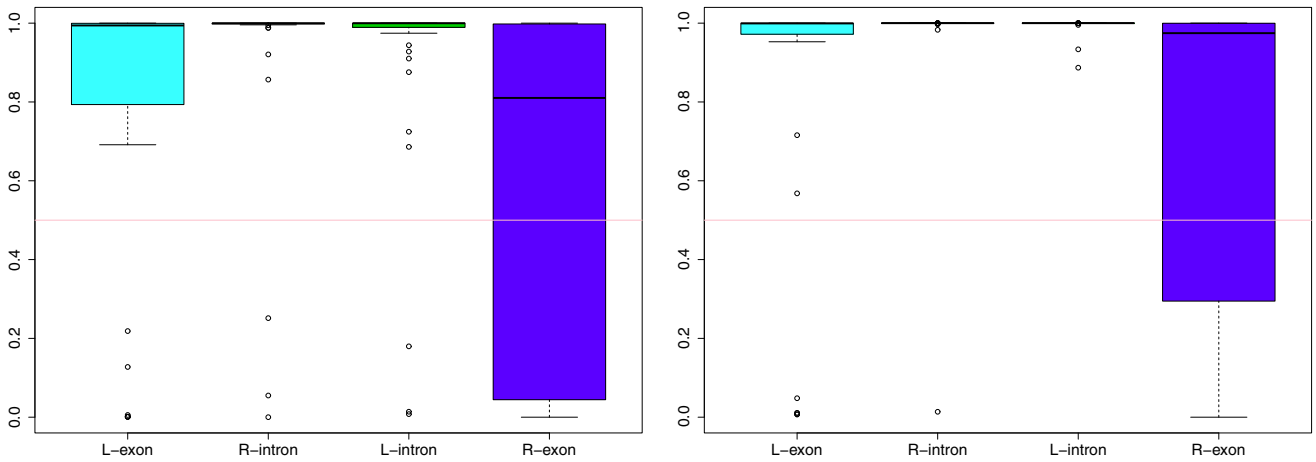

Fig. 6 Distribution of $P(E_0|\mathbf{Y}, \mathbf{K})$ for a set of 50 genes with two values of p_0. We set $p_0 = 1/2$ for all change-points in the *left figure*, and $p_0 = 0.9$ for τ_1 (=L-exon) and τ_4 (=R-exon), and $p_0 = 0.999$ for the intron [Left (L-) and Right (R-)] boundaries in the *right figure*

7 Conclusion

We have proposed two exact approaches for the comparison of the locations of change-points. The first is based on the posterior distribution of the shift between two series, while the second is adapted to the comparison of multiple series and based on the posterior probability of having a common change-point. When applied to transcription profiles of yeast, this methodology confirmed the expectation that transcription starting and ending sites may vary between growth conditions while the localization of introns remains the same.

While we have illustrated these procedures with count datasets, they can be adapted to all distributions from the exponential family verifying the factorability assumption as described in Sect. 2.2. They are in fact implemented in an R package `EBS` for the negative binomial, Poisson, Gaussian heteroscedastic and Gaussian homoscedastic with known variance parameter. This package is available on the CRAN repository at http://cran.r-project.org/web/packages/EBS/index.html.

Acknowledgments The authors deeply thank Sandrine Dudoit, Marie-Pierre Etienne, Emilie Lebarbier, Eric Parent and Gavin Sherlock for helpful conversations and comments on this work. Part of this work was supported by the ABS4NGS ANR project (ANR-11-BINF-0001-06).

Appendix 1

Abacus of posterior probabilities

See Figs. 7 and 8.

Appendix 2

Poisson simulation study

Provided that the ratio $\lambda = \phi(1-p)/p$ remains constant, the negative binomial distribution with dispersion parameter ϕ going to infinity converges to the Poisson distribution $\mathcal{P}(\lambda)$. We have performed a simulation study identical to that presented in Sect. 5 using the Poisson distribution for the comparison with non-dispersed datasets. Specifically, we used for λ_0 the values 1.25 and 0.73 so that the odds ratios $s = 4; 8; 16$ corresponded to the respective values $\lambda_1 = 5; 10; 20$ and $2.92; 5.83; 11.7$ (Fig. 9).

This simulation study confirms that for undispersed data the decision is easier, and that the influence of odds-ratio s and shift d is identical.

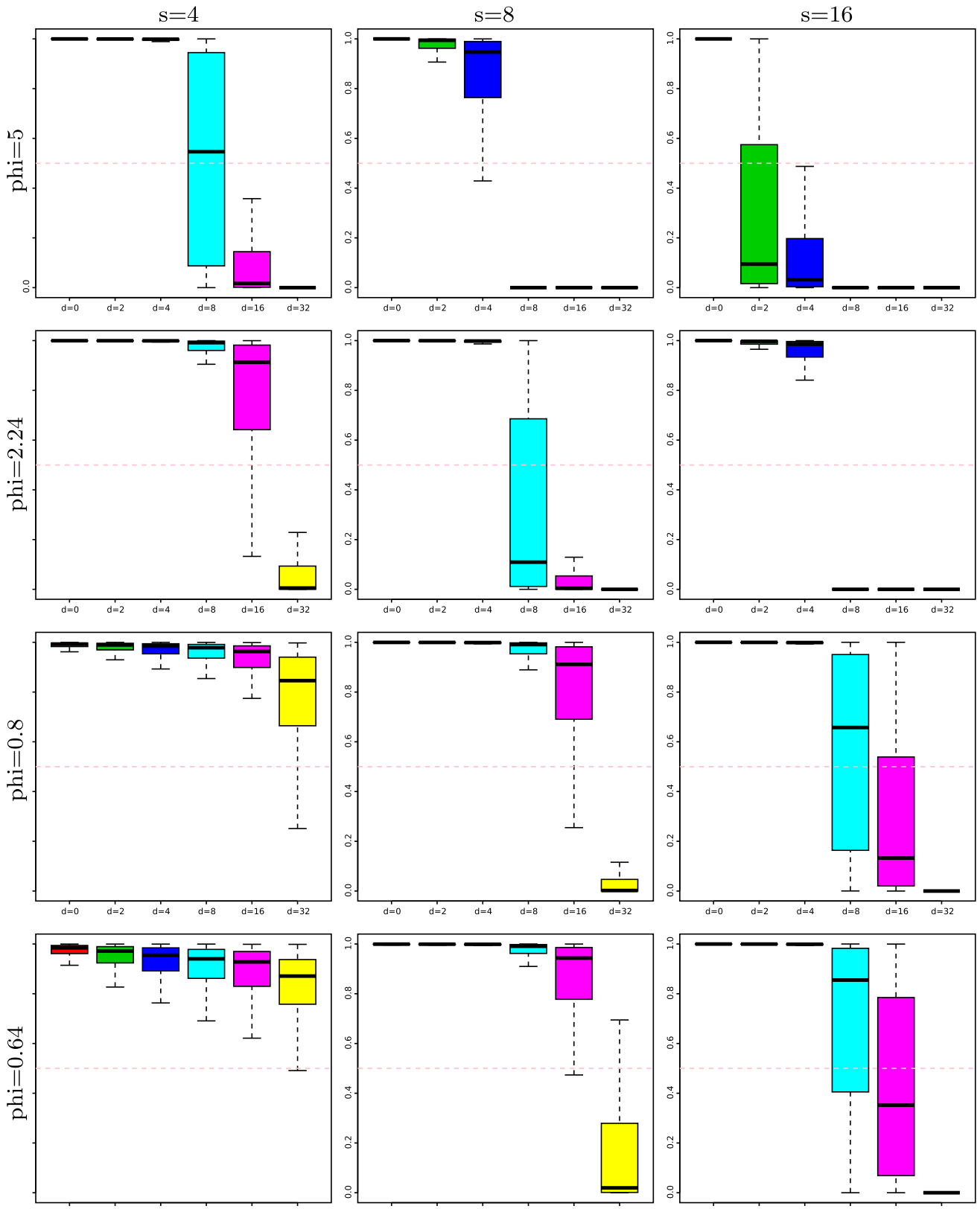

Fig. 7 *Boxplot* of posterior probabilities of E_0 for negative Binomial, with $p_0 = 0.8$. Plotted as d increases in simulation studies for the negative binomial distribution with $p_0 = 0.8$ and for each value of s (*in columns*) and each value of ϕ (*in rows*) as detailed in the left side of Table 1. The overdispersion is estimated as detailed in Sect. 5.1

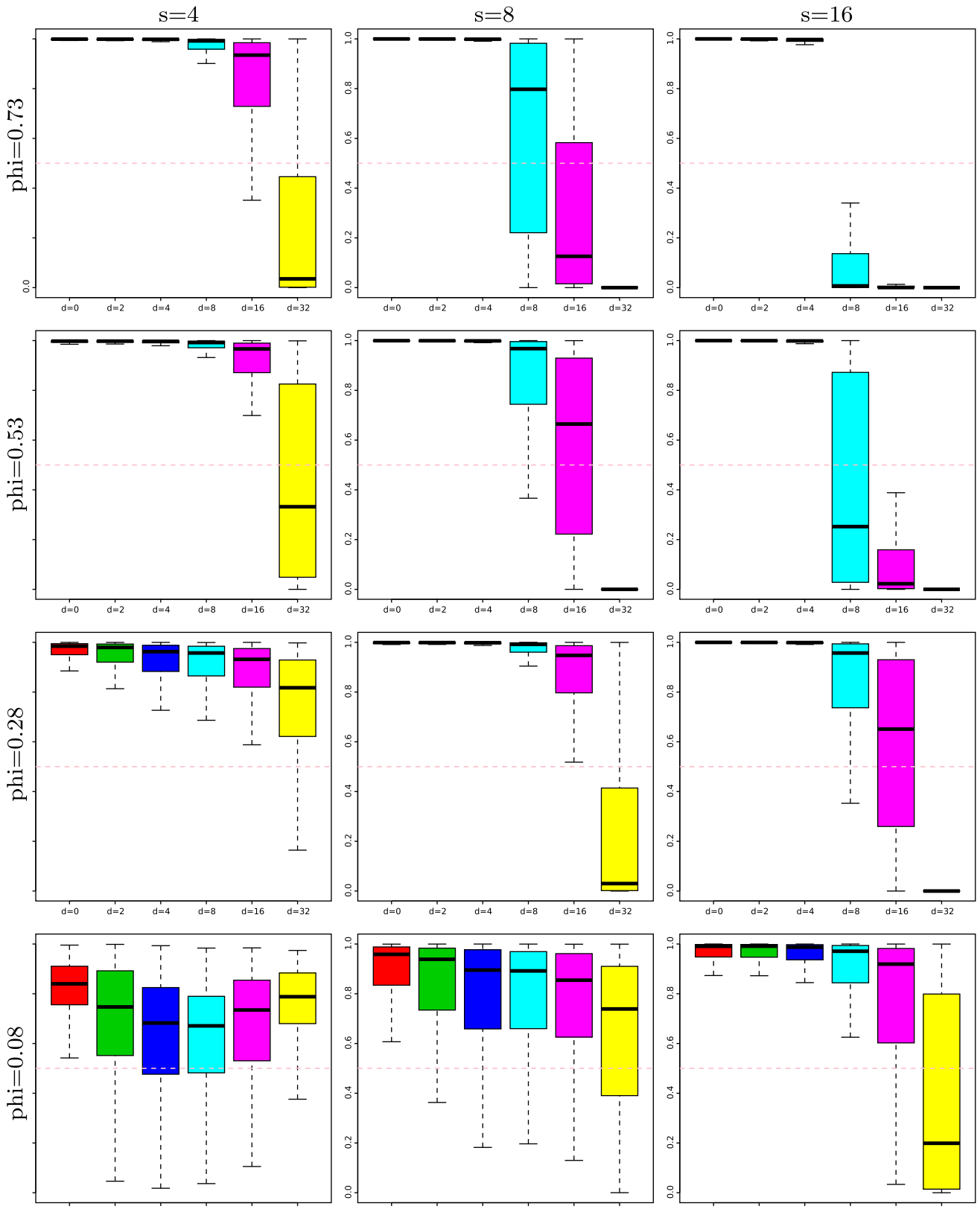

Fig. 8 *Boxplot* of posterior probabilities of E_0 for negative Binomial, with $p_0 = 0.5$. Plotted as d increases in simulation studies for the negative binomial distribution with $p_0 = 0.5$ and for each value of s (*in columns*) and each value of ϕ (*in rows*) as detailed in the right side of Table 1. The overdispersion is estimated as detailed in Sect. 5.1

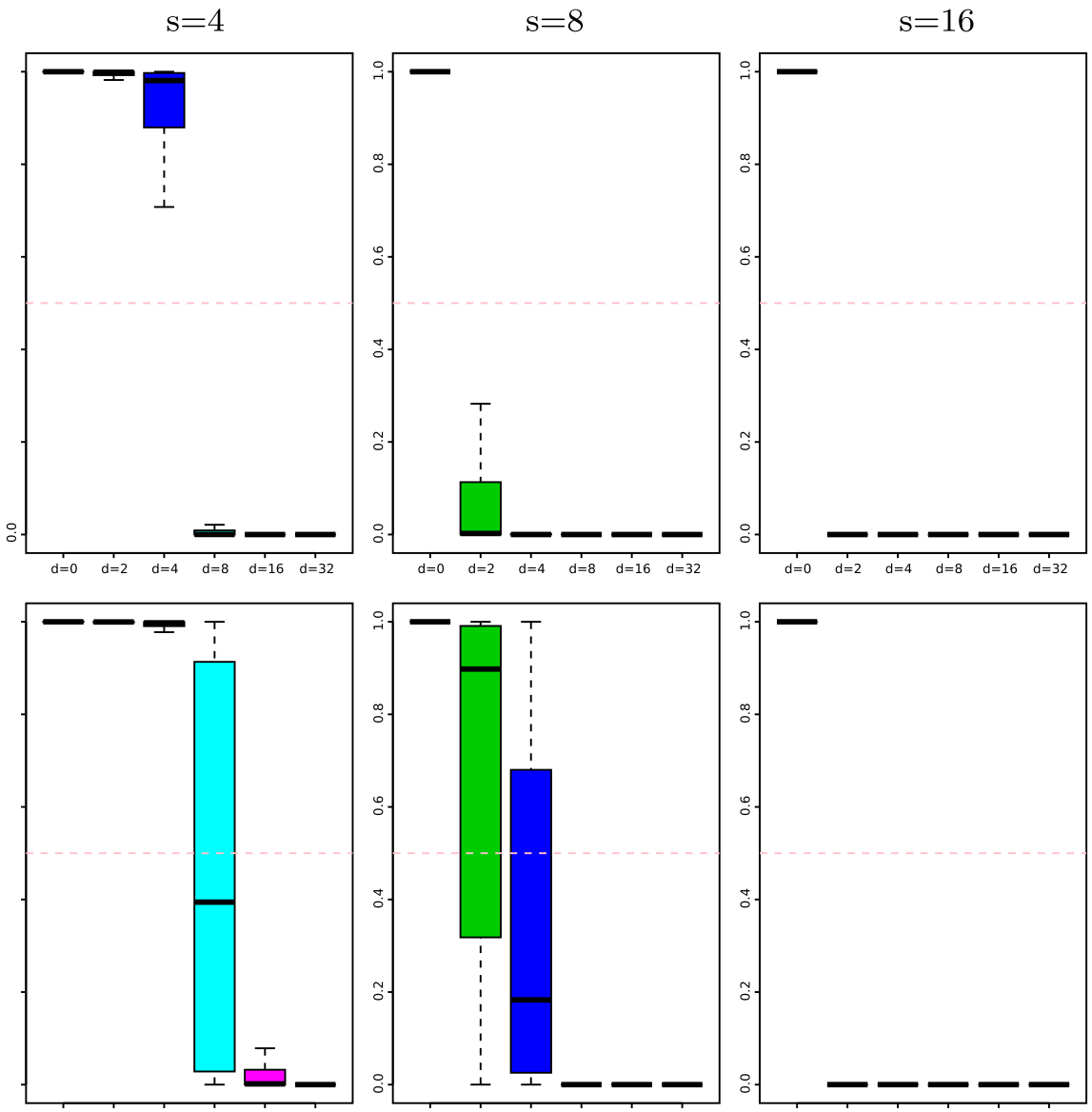

Fig. 9 *Boxplot* of posterior probabilities of E_0 for Poisson. Plotted as d increases in simulation studies for the Poisson distribution with $\lambda_0 = 0.73$ (*Top*) and $\lambda_0 = 2.92$ (*Bottom*) and for each value of s (*in columns*)

References

Armijo, L.: Minimization of functions having Lipschitz continuous first partial derivatives. Pac. J. Math. **16**(1), 1–3 (1966)

Bai, J., Perron, P.: Computation and analysis of multiple structural change models. J. Appl. Econ. **18**, 1–22 (2003)

Dobigeon, N., Tourneret, J.-Y., Scargle, J.D.: Joint segmentation of multivariate astronomical time series: Bayesian sampling with a hierarchical model. IEEE Trans. Signal Process. **55**(2), 414–423 (2007)

Ehsanzadeh, E., Ouarda, T.B., Saley, H.M.: A simultaneous analysis of gradual and abrupt changes in Canadian low streamflows. Hydrol. Process. **25**(5), 727–739 (2011)

Feder, P.I.: The log likelihood ratio in segmented regression. Ann. Stat. **3**, 84–97 (1975)

Hukov, M., Kirch, C.: Bootstrapping confidence intervals for the change-point of time series. J. Time Ser. Anal. **29**(6), 947–972 (2008)

Johnson, N., Kemp, A., Kotz, S.: Univariate Discrete Distributions. Wiley, New York (2005)

Mandal, S.S., Chu, C., Wada, T., Handa, H., Shatkin, A.J., Reinberg, D.: Functional interactions of RNA-capping enzyme with factors that positively and negatively regulate promoter escape by RNA polymerase II. Proc. Natl. Acad. Sci. USA **101**(20), 7572–7577 (2004)

Muggeo, V.M.: Estimating regression models with unknown break-points. Stat. Med. **22**, 3055–3071 (2003)

Picard, F., Lebarbier, E., Hoebeke, M., Rigaill, G., Thiam, B., Robin, S.: Joint segmentation, calling, and normalization of multiple CGH profiles. Biostatistics **12**(3), 413–428 (2011)

Proudfoot, N., Furger, A., Dye, M.: Integrating mRNA processing with transcription. Cell **108**, 501–512 (2002)

Reeves, J., Chen, J., Wang, X.L., Lund, R., QiQi, L.: A review and comparison of change-point detection techniques for climate data. J. Appl. Meteorol. Climatol. **46**(6), 900–915 (2007)

Rigaill, G., Lebarbier, E., Robin, S.: Exact posterior distributions and model selection criteria for multiple change-point detection problems. Stat. Comput. **22**, 917–929 (2012)

Risso, D., Schwartz, K., Sherlock, G., Dudoit, S.: GC-content normalization for RNA-Seq data. BMC Bioinform. **12**(1), 480 (2011)

Tian, B., Hu, J., Zhang, H., Lutz, C.: A large-scale analysis of mRNA polyadenylation of human and mouse genes. Nucleic Acids Res. **33**, 201–212 (2005)

Toms, J., Lesperance, M.: Piecewise regression: a tool for identifying ecological thresholds. Ecology **84**(8), 2034–2041 (2003)

Wager, T.D., Waugh, C.E., Lindquist, M., Noll, D.C., Fredrickson, B.L., Taylor, S.F.: Brain mediators of cardiovascular responses to social threat: part i: reciprocal dorsal and ventral sub-regions of the medial prefrontal cortex and heart-rate reactivity. NeuroImage **47**(3), 821–835 (2009)