Statistica Sinica **22** (2012), 1507-1538
doi:http://dx.doi.org/10.5705/ss.2010.257

MODEL SELECTION FOR HIGH-DIMENSIONAL, MULTI-SEQUENCE CHANGE-POINT PROBLEMS

Nancy R. Zhang and David O. Siegmund

Stanford University

Abstract: Change-point models have been widely applied for segmentation of spatial or time-series data. Some recent applications in genomics motivate multi-sequence change-point models for shared changes across multiple aligned sequences. These applications frequently involve data where the number of change-points can be large. In a previous paper we derived a Bayes Information Criterion (BIC) for determining the number of changes in the mean of a sequence of independent normal observations when the number of change-points m is assumed to remain bounded as the number of observations increases. Here we extend that result to the case where m can increase with the sample size and to simultaneous change-points in multiple sequences. Stochastic terms that enter into the new criteria involve integrals and maxima of two-sided random walks with negative drift. The new criteria are applied to the analysis of DNA copy number data.

Key words and phrases: Change-point detection, DNA copy number, segmentation, model selection.

1. Introduction

With the ubiquity of high-throughput data collection schemes in various scientific disciplines, one frequently sees data of the following structure: For each of $j = 1, \ldots, N$ "subjects," a linearly ordered sequence of noisy observations $\mathbf{y}_j = \{y_{tj} : t = 1, \ldots, T\}$ is collected. Signals in each sequence appear as intervals where the observations exhibit a change in distribution, and these changed segments may be shared by a subset of the subjects. In this paper our analysis is motivated by experiments that profile DNA copy number, so each sequence contains measurements of the quantity of DNA in a biological sample, mapping to ordered locations along a reference genome. Figure 1 shows a single sample of ($\log_2$) DNA copy number assayed from a breast tumor, while Figure 4 shows, in the form of a heatmap, data containing $N = 62$ samples assayed at $T = 2{,}000$ locations. Both T, the length of each sequence, and N, the number of samples, may be very large. Signals in Figure 4 are visible as streaks of gray that are shared across subjects. More background on DNA copy number is given in Zhang et al. (2010) and in Section 7.

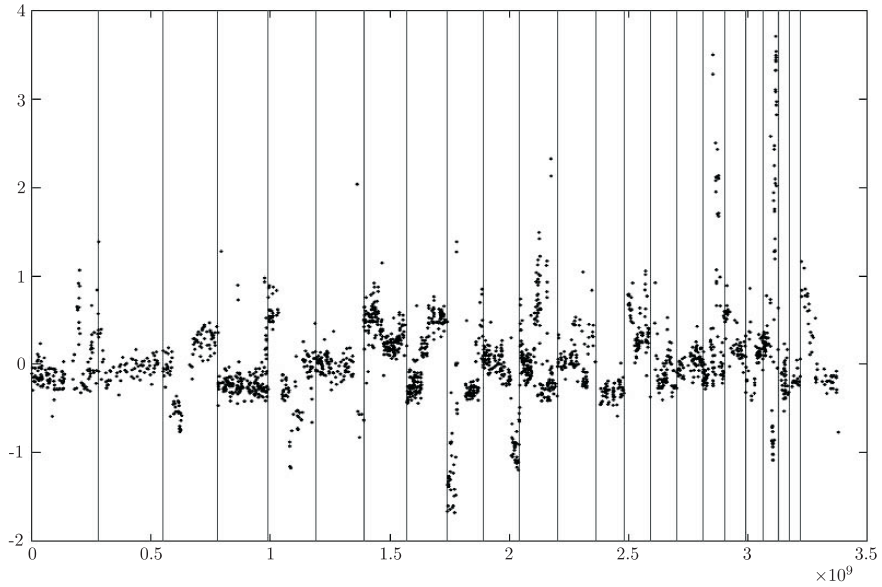

Figure 1. DNA copy number measured by array-based comparative genomic hybridization for tumor sample BT474. Vertical lines denote chromosome boundaries.

We assume that, within each sequence, the data can be modeled as independent observations from a distribution that occasionally experiences abrupt changes. Change-point models have been applied to the segmentation of various types of genome-wide profiles. Earlier models focused on the segmentation problem for one sequence. If a cross-sequence summary is desired, the results can be combined post segmentation. Several recent approaches to DNA copy number analysis advocate pooling data across sequences during the segmentation step (Lipson et al. (2006), Zhang et al. (2010) , Shah et al. (2007)). In this paper we propose model selection methods for single- and multi-sequence change-point problems. Although our methods can be adapted to more complex scenarios, we limit our analysis to the simple model where, within each sequence, we observe independent normally distributed observations with piece-wise constant mean and overall constant variance, and where observations in different sequences are independent. See Zhang et al. (2010) for a description of data normalization designed to insure the approximate validity of these model assumptions.

To establish basic notation and review relevant literature, we start by revisiting the problem of estimating the number of changepoints in the mean of a single sequence of independent observations with normal homoscedastic errors (e.g., Yao (1988), Zhang and Siegmund (2007)). Let

$$0 = \tau_0 < \tau_1 < \cdots < \tau_m < \tau_{m+1} = T$$

be the change-points of the process. Then, suppressing the subject index j,

$$y_t \sim N(\mu_i, \sigma^2) \quad \text{for } \tau_i < t \le \tau_{i+1},\ i = 0, \ldots, m. \tag{1.1}$$

The segment means, the change-point locations, as well as the number of change-points m are all unknown. To simplify our theoretical studies we assume that σ^2 is known. In the very large data sets often encountered in CNV analysis, this parameter can be very accurately estimated, but we also give a slightly more complex result for cases where variability in the estimated variance may play an important role. Given m, the change-point locations and the segment means can be estimated either by maximum likelihood or by Bayesian approaches. The dimension of the model is controlled by m, and thus the determination of m is a model selection problem.

The Bayes information criterion (BIC), proposed by Schwarz (1978), has proved a useful off-the-shelf method for estimating the dimension of parametric models. The original BIC criterion was derived in a Bayesian framework by approximating the log of the posterior probability of the model and neglecting terms that are of constant order when the total number of observations T is assumed to increase indefinitely. The prior distributions play no role in the final approximation if certain assumptions are satisfied. In particular (i) the dimension of each model, as well as the number of possible models must be bounded and (ii) the log likelihood function must be twice differentiable in the parameters of the model. The latter assumption fails to hold for change-point models, while both theoretical considerations and experience suggest that if the number of observations becomes very large, then the number of change-points might also be large.

To deal with (ii), Zhang and Siegmund (2007) assumed that there exists a vector $\boldsymbol{\rho} = (\rho_0, \ldots, \rho_{m+1})$, with

$$0 = \rho_0 < \rho_1 < \cdots < \rho_{m+1} = 1$$

such that $\boldsymbol{\tau}/T \to \boldsymbol{\rho}$, and showed that up to terms of constant order

$$\ell(\hat{\boldsymbol{\tau}}, \hat{\boldsymbol{\mu}}) - 2^{-1} \sum_{i=1}^{m+1} \log(\hat{\tau}_i - \hat{\tau}_{i-1}) - \left(m - \frac{1}{2}\right) \log T \tag{1.2}$$

is an approximation to the log of the posterior model probability and thus is analogous to the original Schwarz criterion. When applied to sequences that contain a large number of change-points, one finds empirically that (1.2) can be quite conservative compared to other methods. This motivates the developments in the first part of this paper, where we consider an approximation to the log of the posterior model probability when the number of change-points m is assumed to

increase with the number of observations T, although $m \ll T$. We refer to these situations as high-dimensional change-point models, and in our approximation to the posterior model probability we keep terms that grow at least as fast as m.

An example motivating our reformulation is provided by the BT474 data studied in Zhang and Siegmund (2007), where 39 change-points were detected, while some methods detected more. With the criterion suggested below we now detect 77. The truth for this example is unknown, so we have relied on related simulations, reported below, to try to evaluate different methods.

When approximating the Bayes factor for high-dimensional models, the prior probabilities of some models are necessarily very small and hence are no longer negligible. In a high-dimensional, but statistically regular, context, Berger, Ghosh, and Mukhopadhyay (2003) proposed a mixture of Gaussian priors. We adapt their prior distributions to our change-point setting to treat the large number of segment mean parameters. This leads to an extra term in the BIC approximation that involves the prior distribution, but which can be estimated by empirical Bayes methods. Additional new terms that involve functionals of random walks also arise in the approximation, due to variability in the estimates $\hat{\boldsymbol{\tau}}$ of the change-points.

We also consider extensions of the single sequence change-point model to estimate shared change-points in multiple aligned sequences. Such extensions have been studied from different points of view in Shah et al. (2007), Lipson et al. (2006), Wang et al. (2008), Zhang et al. (2010) and Siegmund, Yakir, and Zhang (2011). At a change-point, a subset of the sequences, the *carriers*, experience a shift in mean. The magnitude of the shift is allowed to differ across samples for the same change-point, and across change-points for the same sample. The variances of the observations are sequence specific, but constant within sequences. The locations of the change-points, the sets of carriers, and the mean shift for each carrier at each change-point are all unknown. Extending the circular binary segmentation approach (CBS) (Vostrikova (1981), Olshen et al. (2004)) for single sequence segmentation, Zhang et al. (2010) suggested a recursive algorithm for partitioning the multi-sequence data into homogeneous regions. The algorithm contains an ad hoc thresholding step for identifying the carriers of each change-point, and a p-value based criterion for stopping the recursion. The BIC model selection procedures derived in this paper can be used with a similar algorithm that eliminates the need for several user defined tuning parameters.

We adopt a simple Bernoulli prior for the carrier sets, and estimate its hyper-parameters empirically. Such an empirical Bayes approach to the BIC has been used earlier for regular models by George and Foster (2000) and in a regression setting by Chen and Chen (2008).

The new BIC criteria proposed in this paper have a convenient modular property, in the sense that they are comprised of interpretable terms, each of which

is attributable to a separate part of the model. This facilitates understanding of the procedure and generalization to more complex problems, as we illustrate below.

The paper is organized as follows. After introducing our basic notation in Section 2, Section 3 focuses on the analysis of a single sequence for high-dimensional change-point models. After showing that two different priors for the mean value process lead to slightly different BIC approximations, in Section 4 we show via simulations that the two new versions behave about the same and outperform the old one. The problem of multiple aligned sequences is treated in Section 5. After giving in Section 7.1 a more detailed introduction to DNA copy number data, Sections 7.2 and 7.3 contain applications to the segmentation of *inherited* DNA copy number variants and of aberrations in tumor data, respectively. Section 8 contains a discussion, and the Appendix provides some of the technical details in the derivations of the results in Sections 3 and 5.

Remarks. (i) In the case of multiple sequences, a model with simultaneous changepoints in different sequences is scientifically plausible for inherited CNV, many of which arise from mutational events in the history of a population and indicate distant relatedness of individuals exhibiting the same CNV. The model is not obviously appropriate for cancer data. There do, however, appear to be genomic locations more prone to breakage; in cancer data the most interesting changes for their possible role in the pathology of a cancer are those that are shared by different individuals. Although our multi-sequence model is based on the assumption of change-points that are aligned across samples, the iterative nature of our algorithm achieves a reasonable level of robustness against breakdown of this assumption. See the analysis of a complicated CNV region with several varying change-points in Section 7.2, and of a tumor sample in Section 7.3.

(ii) A model selection procedure often considered alongside BIC is AIC, for which a change-point version has been suggested by Ninomiya (2004). Since this version of AIC (like its analogue in statistically regular problems) penalizes by counting parameters, but does not take account of the sample size, it can lead to arbitrarily many false positive errors in very long sequences populated by relatively few change-points.

2. Notation

In this section we introduce notation that will be used throughout the paper. Let

$$\boldsymbol{\tau}_T = (\tau_{T,0}, \ldots, \tau_{T,m+1})$$

denote the true values of the change-points, where $\tau_{T,0} = 0$ and $\tau_{T,m+1} = T$, so $m = m_T$ is the true number of change-points. The number of change-points, as well as their values, are allowed to change with T, but we usually suppress the dependence on T in the notation. We always use i to index the change-points.

We use t to index the observations within a sequence. The sequences all have length T, so $t \in \{1, \ldots, T\}$. We use vectors $\mathbf{t} = (t_0, \ldots, t_{m+1})$ to denote possible values for $\boldsymbol{\tau}$.

When there are $N > 1$ sequences, we use j to index sequences. For example, $y_{\tau_i,j}$ is the observation at the i-th change-point in the j-th sequence.

Let $S = \{S_0, S_1, \ldots, S_T\}$ denote the cumulative sums process, $S_t = S_{t-1} + y_t$, with $S_0 = 0$. When there are multiple sequences, $\{S_{0j}, S_{1j}, \ldots, S_{Tj}\}$ denotes the process of cumulative sums in the j-th sequence.

We denote vectors in bold face, such as $\boldsymbol{\tau}$, $\boldsymbol{\mu}$, and $\mathbf{t}$.

3. Analysis of a Single Sequence

Consider a single sequence of observations where $y_1, \ldots, y_T$ follow model (1.1). The vector of change-points, $\boldsymbol{\tau} = \boldsymbol{\tau}_T$, takes value in the set

$$\mathcal{D} = D(m, T) = \{(t_0, \ldots, t_{m+1}) : \ 0 = t_0 < t_1 < \cdots < t_m < t_{m+1} = T\}.$$

Note that the cardinality of $\mathcal{D}$ is $|D(m,T)| = \binom{T}{m} \approx T^m/m!$ when $m \ll T$. We denote by M_m the model with m change-points. For any $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_m) \in \Re^m$ and $\boldsymbol{\tau} \in \mathcal{D}$, let $\mathrm{P}_{\boldsymbol{\tau},\boldsymbol{\delta}}$ be the measure where the means change by magnitude δ_i at $\tau_i + 1$, $\delta_i = \mu_i - \mu_{i-1}$. Let P_0 denote the measure under the null hypothesis $H_0 : y_t \sim N(\mu_0, \sigma^2), \quad t = 1, \ldots, T$, for some unknown $\mu_0 \in \Re$, and let $\mathrm{E}_{\boldsymbol{\tau},\boldsymbol{\delta}}$ and E_0 denote expectations under $\mathrm{P}_{\boldsymbol{\tau},\boldsymbol{\delta}}$ and P_0, respectively.

Remark. We have parameterized our model by $\boldsymbol{\delta}$, which specifies changes in the mean value at the change-points. While this initially introduces some complications in computing the likelihood function compared to using the means μ_i, for change-point problems the parameterization in terms of $\boldsymbol{\delta}$ is more appropriate, since the sizes of the changes, not the means themselves, are critical to the success of any segmentation method.

For a candidate change-point vector $\mathbf{t}$, it will be convenient to let $X(\mathbf{t})$ denote the $m \times 1$ vector with elements

$$X_i(\mathbf{t}) = t_i \frac{S_T}{T} - S_{t_i}, \quad i = 1, \ldots, m,$$

and let $\Sigma(\mathbf{t})$ be the $m \times m$ covariance matrix of $X(\mathbf{t})$, so

$$\Sigma_{i,j}(\mathbf{t}) = t_i(1 - \frac{t_j}{T}), \quad \text{for } 1 \le i < j \le m. \tag{3.1}$$

A tedious but straightforward calculation shows that

$$\mathrm{E}_{\mathbf{t},\boldsymbol{\delta}}[X(\mathbf{t})] = \Sigma(\mathbf{t})\boldsymbol{\delta}.$$

Let $V(\mathbf{t}) = \Sigma^{-1}(\mathbf{t})X(\mathbf{t})$, so $\mathrm{E}_{\tau,\boldsymbol{\delta}}[V(\boldsymbol{\tau})] = \boldsymbol{\delta}$ and $\mathrm{E}_0[V(\mathbf{t})V'(\mathbf{t})] = \Sigma^{-1}(\mathbf{t})$. The log likelihood ratio of $\mathrm{P}_{\mathbf{t},\boldsymbol{\delta}}$ versus P_0 can be written

$$\ell(\mathbf{t},\boldsymbol{\delta}) = \log\frac{d\mathrm{P}_{\mathbf{t},\boldsymbol{\delta}}}{d\mathrm{P}_0} = \boldsymbol{\delta}'\Sigma(\mathbf{t})V(\mathbf{t}) - \boldsymbol{\delta}'\Sigma(\mathbf{t})\boldsymbol{\delta}/2 = \boldsymbol{\delta}'X(\mathbf{t}) - \boldsymbol{\delta}'\Sigma(\mathbf{t})\boldsymbol{\delta}/2. \tag{3.2}$$

Maximizing (3.2) with respect to $\boldsymbol{\delta}$ gives the likelihood ratio process as a function of $\mathbf{t}$,

$$\ell(\mathbf{t}) = \max_{\boldsymbol{\delta}} \ell(\mathbf{t},\boldsymbol{\delta}) = \frac{1}{2}X(\mathbf{t})'\Sigma^{-1}(\mathbf{t})X(\mathbf{t}). \tag{3.3}$$

The maximum likelihood estimates for the change-points are thus

$$\hat{\boldsymbol{\tau}} = \hat{\boldsymbol{\tau}}_T = \operatorname*{argmax}_{\mathbf{t}\in\mathcal{D}} \ell(\mathbf{t}). \tag{3.4}$$

To proceed, we need prior distributions for the parameters $\boldsymbol{\tau}$, $\boldsymbol{\delta}$, and m. We assume that the τ_i are the discrete analogue of uniform order statistics on $[0,T]$ although order statistics from any prior of the form $T^{-1}p(t/T)$, where p is a continuous density on $[0,1]$ bounded away from 0 and ∞, would give the same results. We consider two priors for $\boldsymbol{\delta}$. For the first, conditional on $\boldsymbol{\tau} = \mathbf{t}$, $\boldsymbol{\delta} \sim N(0, w^{-1}\Sigma^{-1}(\mathbf{t}))$. One can obtain exact formulas for the Bayes factor in Gaussian regression models, where this prior is frequently used and is called the g-prior by Zellner (1986). For the second prior we assume that $\boldsymbol{\delta} \sim N(0, w^{-1}I)$; call this the independence prior for $\boldsymbol{\delta}$.

A complete Bayesian analysis would assign w a prior distribution, say with a positive continuous density $g(w)$ on $[0,\infty)$.

Mixtures of g-priors have been used for high-dimensional Gaussian models in Berger, Ghosh, and Mukhopadhyay (2003) and were shown to have desirable properties in Liang et al. (2008). In the computation of the Bayes Factor when m is large we make a Laplace approximation, so the exact form of $g(\cdot)$ is unimportant. In effect, w is replaced by a maximizing value in the approximation that can be interpreted as an empirical Bayes estimator.

We assume the prior distribution of m decays algebraically, say $\pi(m) \propto m^{-\alpha}$ for some $\alpha > 0$, so $\log\pi(m) = o(m)$ and hence can be neglected.

To state the main results of this section we need to make some assumptions about the configurations of the parameters $\boldsymbol{\tau}, \boldsymbol{\delta}$, which we describe here informally.

I The δ_i are bounded away from 0 and the smallest separation between the change-points must increase with the sequence length, e.g.,

$$\liminf_{T\to\infty} \min_{1\le i\le m_T} \frac{(\tau_{T,i}-\tau_{T,i-1})}{[\log(T)]^2} = \infty.$$

II The values of $\boldsymbol{\delta}$ and $\boldsymbol{\tau}$ must ensure the consistency of the maximum likelihood estimator of $\boldsymbol{\tau}$, e.g., that

$$\lim_{T\to\infty} [\log(T)]^{-2} \max_{1\le i\le m_T} |\hat{\tau}_{T,i}-\tau_{T,i}| \to 0 \text{ in probability.}$$

III The value of the likelihood ratio process $\ell(\boldsymbol{\tau})$ is stochastically of order T.

Taken together, these assumptions ensure that the change-points are estimable, and that, in the limit, the true model can be identified. See the unpublished Stanford University Ph. D. thesis of E. S. Venkatraman, who proved consistency of a number of different segmentation algorithms under slightly weaker conditions.

In view of the length of this paper and the technical calculations involved, we have not derived mathematically rigorous results, but have provided heuristic derivations that illustrate the origins of the various terms in the approximations. An important application is to copy number data, where the conditions given above are admittedly not satisfied. In particular, intervals between change-points can be short. In these problems consistent estimation does not appear to be possible, and justification for use of such a procedure must be based on empirical evidence (i) with simulations, where one knows the answer, and (ii) with scientific data, where one can compare different methods empirically for their consistency with each other, for their capacity to deal with complex data, and for their performance on the occasional set of data where laboratory methods provide some check on the methods' accuracy.

Strictly speaking, the following calculations presuppose that S_t is Brownian motion observed continuously on $[0,T]$, which we are treating as a convenient approximation for our discrete time problem. Otherwise the numerical constants κ_1, κ_2 defined below would involve functionals of a discrete time random walk and would depend on $\boldsymbol{\delta}$. We return to this point in the Appendix.

Let $\{W_t : \ -\infty < t < \infty\}$ be Brownian motion with $W_0 = 0$, and define constants

$$\kappa_1 = \mathrm{E}\left[\max_{-\infty<t<\infty}\{W_t - \frac{|t|}{2}\}\right] = \frac{3}{2}, \tag{3.5}$$

$$\kappa_2 = \mathrm{E}\left[\log \int_{-\infty}^{\infty} \exp\{W_t - \frac{|t|}{2}\}dt\right] \tag{3.6}$$

$$= \mathrm{E}[\log(Y_1^{-1} + Y_2^{-1})] \approx 2.27, \tag{3.7}$$

where Y_1, Y_2 are independent and exponentially distributed with mean value $1/2$ (cf. Pollack and Siegmund (1985)). Define

$$\hat{\delta}_1 = \frac{\hat{\tau}_1 S_{\hat{\tau}_2}/\hat{\tau}_2 - S_{\hat{\tau}_1}}{\hat{\tau}_1(1-\hat{\tau}_1/\hat{\tau}_2)}, \tag{3.8}$$

$$\hat{\delta}_i = \frac{\hat{\tau}_i S_{\hat{\tau}_{i+1}}/\hat{\tau}_{i+1} - S_{\hat{\tau}_i}}{\hat{\tau}_i(1-\hat{\tau}_i/\hat{\tau}_{i+1})} - \left(\frac{\hat{\tau}_{i-1}}{\hat{\tau}_i}\right)\frac{\hat{\tau}_{i-1} S_{\hat{\tau}_i}/\hat{\tau}_i - S_{\hat{\tau}_{i-1}}}{\hat{\tau}_{i-1}(1-\hat{\tau}_{i-1}/\hat{\tau}_i)}, \tag{3.9}$$

for $i = 2, \ldots, m$. Let $\hat{\boldsymbol{\delta}} = (\hat{\delta}_1, \ldots, \hat{\delta}_m)$. By the Law of Large Numbers, it is easy to verify that assuming I and II, as $T \to \infty$,

$$\max_i |\hat{\delta}_i - \delta_i| \to 0 \tag{3.10}$$

in probability, so the $\hat{\delta}_i$ are consistent estimators of the jump parameters δ_i.

BIC-like model selection criteria for the single sequence model under the large m scenario are given in the following proposition, where we evaluate the Bayes factor for the mth model up to terms that are small compared to m.

Proposition 1. Under assumptions (I–III), the distributional and prior assumptions stated above for $\boldsymbol{\tau}$ and y, and the g-prior for $\boldsymbol{\delta}$, as $T \to \infty$,

$$\log \frac{\mathrm{P}(M_m|y)}{\mathrm{P}(M_0|y)} = \ell(\hat{\boldsymbol{\tau}}) - \frac{m}{2}\left\{\log\left[\frac{2\ell(\hat{\boldsymbol{\tau}})}{m}\right]\right\}^+ - \frac{m}{2} - m(\kappa_1 - \kappa_2) - \sum_{i=1}^{m} \log \hat{\delta}_i^2 - \log|D(m,T)| + o_p(m), \tag{3.11}$$

where $\hat{\boldsymbol{\tau}}$ is the maximum likelihood change-point estimate (3.4). For the independence prior for $\boldsymbol{\delta}$, as $T \to \infty$,

$$\log \frac{\mathrm{P}(M_m|y)}{\mathrm{P}(M_0|y)} = \ell(\hat{\boldsymbol{\tau}}) - 2^{-1}\sum_{i=1}^{m+1} \log(\hat{\tau}_i - \hat{\tau}_{i-1}) - \frac{m}{2}\log\left[\sum_{i=1}^{m} \frac{\hat{\delta}_i^2}{m}\right] - \frac{m}{2} - m(\kappa_1 - \kappa_2) - \sum_{i=1}^{m} \log \hat{\delta}_i^2 - \log|D(m,T)| + o_p(m). \tag{3.12}$$

Note that in (3.11) and (3.12), the terms m, $\hat{\boldsymbol{\delta}}$, and $\hat{\boldsymbol{\tau}}$ all depend on T, but we have suppressed the dependence in our notation. We define the right hand sides of (3.11) and (3.12), without the $o_p(m)$ remainder terms, to be the modified Bayes information criteria for the one sequence change-point model. For future reference we call the criteria BIC_1 and BIC_2, respectively.

Remark. The use of (3.11) or (3.12) requires some discretion. There is an implicit assumption that the term involving the maximized log likelihood dominates

the other terms to the extent that the entire expression is positive. Any particular model for which this fails to be true should not be considered. Presumably it is unnecessary to add this condition, although one should note that penalty terms involving the logarithm of the $\hat{\delta}_i$ provides an incentive to look at models with many very small values of the $\hat{\delta}_i$. In practice the overall expressions increase with m to a maximum value and then begin to decrease. This local maximum should be the value of $\hat{m}$, even though in principle the expressions might begin to increase again for still larger values of m.

We sketch the derivation of the first approximation in Proposition 1 here to allow better understanding of the origin of the approximation and an interpretation of the terms therein. Some of the technical details are given in the Appendix. To compute the Bayes Factor on the left hand side of (3.11), we integrate the likelihood ratio $d\mathrm{P}_{\mathbf{t},\boldsymbol{\delta}}/dP_0$ over the prior for $\boldsymbol{\delta}$ and then sum over all possible values of $\mathbf{t}$ to get

$$\begin{aligned}\frac{\mathrm{P}(M_m|y)}{\mathrm{P}(M_0|y)} &= |\mathcal{D}|^{-1}\sum_{\mathbf{t}\in\mathcal{D}}\int_0^\infty g(w)\Big(\frac{w}{2\pi}\Big)^{m/2}|\Sigma(\mathbf{t})|^{-1/2}\\ &\quad\times\int_{\Re^m} e^{-(1/2)\boldsymbol{\delta}'\Sigma(\mathbf{t})\boldsymbol{\delta}(1+w)+\boldsymbol{\delta}'X(\mathbf{t})}d\boldsymbol{\delta}dw\\ &= |\mathcal{D}|^{-1}\sum_{\mathbf{t}\in\mathcal{D}}\int_0^\infty g(w)\Big[\frac{w}{1+w}\Big]^{m/2}e^{(1/2)X(\mathbf{t})'\Sigma(\mathbf{t})^{-1}X(\mathbf{t})(1+w)^{-1}}dw\\ &= |\mathcal{D}|^{-1}\sum_{\mathbf{t}\in\mathcal{D}}e^{(1/2)X(\mathbf{t})'\Sigma(\mathbf{t})^{-1}X(\mathbf{t})}\\ &\quad\times\int_0^\infty g(w)\Big[\frac{w}{1+w}\Big]^{m/2}e^{-\frac{1}{2}X(\mathbf{t})'\Sigma(\mathbf{t})^{-1}X(\mathbf{t})[w/(1+w)]}dw \qquad (3.13)\end{aligned}$$

Putting $\eta = w/(1+w)$, and letting

$$f(\eta) = \eta^{m/2}\exp\Big(-\frac{\eta X(\mathbf{t})'\Sigma(\mathbf{t})^{-1}X(\mathbf{t})}{2}\Big), \qquad (3.14)$$

we see that the integral with respect to w in (3.13) becomes

$$\int_0^1 g[\eta(1-\eta)^{-1}]f(\eta)(1-\eta)^2 d\eta. \qquad (3.15)$$

Let

$$\eta^* = \operatorname*{argmax}_{0\le\eta\le1} f(\eta) = \min\{m[X(\mathbf{t})'\Sigma(\mathbf{t})^{-1}X(\mathbf{t})]^{-1}, 1\}.$$

As $m\to\infty$, $T\to\infty$, the integrand in (3.15) is dominated by $\eta^{m/2}$ for $\eta\to0$ and $\exp(-\eta X(\mathbf{t})'\Sigma(\mathbf{t})^{-1}X(\mathbf{t})/2) = O_p[\exp(-\eta T)]$ for $\eta\to1$. Hence (3.15) is

asymptotically equivalent to $f(\eta^*)\exp(E_1)$, where E_1 is an $O_p(1)$ random factor that can be ignored. Since

$$f(\eta^*) = \exp\Big[-\frac{m}{2} - (\frac{m}{2})\log\Big(\frac{X(\mathbf{t})'\Sigma(\mathbf{t})^{-1}X(\mathbf{t})}{m}\Big)\Big], \tag{3.16}$$

substituting in (3.13) and taking logs leads to

$$\begin{aligned}\log\frac{\mathrm{P}(M_m|y)}{\mathrm{P}(M_0|y)} = {}& \frac{1}{2}X(\hat{\boldsymbol{\tau}})'\Sigma(\hat{\boldsymbol{\tau}})^{-1}X(\hat{\boldsymbol{\tau}}) - \frac{m}{2} - \frac{m}{2}\log\Big[\frac{X(\hat{\boldsymbol{\tau}})'\Sigma(\hat{\boldsymbol{\tau}})^{-1}X(\hat{\boldsymbol{\tau}})}{m}\Big] \\ & - \log|D| + A(\boldsymbol{\tau}) + B(\boldsymbol{\tau}) + E_2,\end{aligned} \tag{3.17}$$

where

$$A(\boldsymbol{\tau}) = \frac{1}{2}[X(\boldsymbol{\tau})'\Sigma(\boldsymbol{\tau})^{-1}X(\boldsymbol{\tau}) - X(\hat{\boldsymbol{\tau}})'\Sigma(\hat{\tau})^{-1}X(\hat{\boldsymbol{\tau}})], \tag{3.18}$$

$$B(\boldsymbol{\tau}) = \log\left\{\sum_{\mathbf{t}\in D}\exp\left[\frac{1}{2}(X(\mathbf{t})'\Sigma(\mathbf{t})^{-1}X(\mathbf{t}) - X(\boldsymbol{\tau})'\Sigma(\boldsymbol{\tau})^{-1}X(\boldsymbol{\tau}))\right]\right\}, \tag{3.19}$$

and E_2 is an $O_p(1)$ error term. The proposition then follows from the approximations

$$A(\boldsymbol{\tau}) = -m\kappa_1 + o_p(m), \tag{3.20}$$

$$B(\boldsymbol{\tau}) = m\kappa_2 - \sum_{i=1}^{m}\log(\hat{\delta}_i^2) + o_p(m). \tag{3.21}$$

As noted above, in (3.20) and (3.21) we have replaced functionals of discrete time random walks by the corresponding Brownian functionals. The derivations of these results are given in the Appendix.

Compared to the classic BIC and the modified BIC in equation (1.2) for low-dimensional change-point models, the leading terms of BIC_1 and BIC_2 are still the maximized log-likelihoood. The differences between these criteria are in the penalty terms for $\boldsymbol{\delta}$ and $\boldsymbol{\tau}$. Comparing BIC_2 given in (3.12) to (1.2), we see that the penalty for each δ_i, $\log(\hat{\tau}_i - \hat{\tau}_{i-1})$, can be interpreted as the effective sample size for estimating δ_i. The additional term $-m\log(\sum_i \hat{\delta}_i^2/2) - m/2$ in (3.12) comes from the mixture prior for $\boldsymbol{\delta}$, and effectively arises from plugging in an empirical estimate for w, the variance of $\boldsymbol{\delta}$. This term is negligible when m is assumed to remain bounded. A similar analysis applies in the case of BIC_1 and the g-prior.

The term $\log|D(m,T)|$, which comes directly from the prior distribution of $\boldsymbol{\tau}$, penalizes for the uncertainty about the m-dimensional change-point parameter over the range $\{1,\ldots,T\}$. In (1.2), this term was replaced by $m\log T$, its

asymptotic equivalent if m were of constant order. An additional penalty of

$$-\kappa_1 + \kappa_2 - \log \hat{\delta}_i^2 \tag{3.22}$$

for each change-point parameter τ_i comes from a more careful analysis of the terms $A(\boldsymbol{\tau})$ and $B(\boldsymbol{\tau})$ that were ignored in Zhang and Siegmund (2007) because they do not grow with T. Since these terms are of order m, which is now assumed to be large, they are included here.

Note that BIC_1 and BIC_2 differ only in the penalty terms attributable to $\boldsymbol{\delta}$ (i.e., the terms in the first lines of (3.11) and (3.12)). This modularity makes it easy to generalize these approximations to more complex models, as we will see in Section 5.

Proposition 1 can be extended to handle the case where the noise variance σ^2 is unknown. To describe the result extended to this case, we assume that σ has an improper uniform prior. This assumption is not necessary for the result, as any smooth prior with support on $[0, \infty)$ would give the same asymptotic approximation. Given σ, we still assume that δ is a mixture of Gaussian distributions, but with its variance scaled by σ^2.

Regarding the sample partitioned by $\mathbf{t}$ as a one-way analysis of variance and writing the total sum of squares as the sum of within-group and between-group sums of squares, we have, in obvious notation,

$$SS_{\mathrm{all}} = \sum_{i=1}^{T} (y_i - \bar{y})^2, \tag{3.23}$$

$$SS_{\mathrm{bg}}(\mathbf{t}) = X(\mathbf{t})' \Sigma^{-1}(\mathbf{t}) X(\mathbf{t}) = \sum_{i=0}^{m} n_i(\mathbf{t}) [\bar{y}_i(\mathbf{t}) - \bar{y}]^2, \tag{3.24}$$

where

$$n_i(\mathbf{t}) = t_{i+1} - t_i, \quad \bar{y}_i(\mathbf{t}) = n_i(\mathbf{t})^{-1} (S_{t_{i+1}} - S_{t_i}), \quad i = 0, \ldots, m,$$

$$SS_{\mathrm{wg}}(\mathbf{t}) = SS_{\mathrm{all}} - SS_{\mathrm{bg}}(\mathbf{t}) = \sum_{i=0}^{m} \sum_{j=t_i+1}^{t_{i+1}} [y_j - \bar{y}_i(\mathbf{t})]^2. \tag{3.25}$$

Given a change-point estimate $\mathbf{t}$, the maximum likelihood estimate of σ^2 is

$$\hat{\sigma}^2 = (T-1)^{-1} SS_{\mathrm{wg}}(\mathbf{t}).$$

The generalized log likelihood ratio of the measure $P_{\mathbf{t},\boldsymbol{\delta},\sigma}$ for the model with change-points at $\mathbf{t}$ versus the measure P_σ of the null model is

$$\ell(\mathbf{t}) = \log \frac{\max_{\boldsymbol{\delta},\sigma} dP_{\mathbf{t},\boldsymbol{\delta},\sigma}}{\max_\sigma dP_\sigma} = \frac{T-1}{2} \log \left(1 + \frac{SS_{\mathrm{bg}}}{SS_{\mathrm{wg}}} \right). \tag{3.26}$$

As before, take $\hat{\boldsymbol{\tau}} = \operatorname{argmax}_{\mathbf{t}} \ell(\mathbf{t})$. Then the appropriately modified BIC is given by (3.11), with the form (3.26) for $\ell(\mathbf{t})$, and with the term involving $\sum_i \hat{\delta}_i^2$ replaced by $\sum_i \hat{\delta}_i^2/\hat{\sigma}^2$. A sketch of the derivation of this result is given in the online Supplementary Material.

4. Simulations

We have conducted simulations to study (i) the effect of the choice of prior distribution for $\boldsymbol{\delta}$ and (ii) the effect of the number of change-points.

In the first simulation we compared BIC_1 and BIC_2 from Proposition 1 under a g-prior for $\boldsymbol{\delta}$ and under the independent prior. In both cases both procedures gave very similar results, which suggests that the nominal dependence on the prior has little impact on the conclusion. Details are given in the online Supplementary Material.

In the second simulation we compared BIC_1 with the version of BIC at (1.2) when m is small and when m is large. In both cases the approximation (3.11) achieved more sensitivity when the signal was weak and higher specificity when the signal was strong. We also compared the case of unknown variance with the case where the variance is assumed known. The results are virtually the same. Details are again available online.

5. Analysis of Multiple Sequences

We now consider models for change-points that are shared across multiple sequences. First, we consider an extension of (1.1) to the multi-sequence case. As an additional example of the BIC method, we consider a model in which the sequences need to return to a "baseline state" after every excursion into a changed state.

Since the two approximations derived from Proposition 3.1 are similar, both in form and in performance, we show extensions of only the first case (g-prior).

5.1. Multiple sequence mixture model

As before let M_m be the model where there are m change-points. For each $i = 1, \ldots, m$, let J_i be the carriers of the ith change-point, defined as the subset of sequences that have a change in mean at $\tau_i + 1$. For $j \in J_i$, let δ_{ij} be the change in mean of sample j. The segment means, change-points, and carrier sets $J = (J_1, \ldots, J_m)$ are all unknown and must be estimated from the data. The dimension of $\boldsymbol{\delta}$ in the model is equal to

$$M = \sum_{i=1}^{m} |J_i|. \tag{5.1}$$

The set of models under consideration is now

$$\mathcal{M} = \{(m, J_1, \ldots, J_m) : \ m = 0, 1, 2, \ldots; \quad J_i \in \{0,1\}^N, i = 1, \ldots, m\}$$

and, since its cardinality can be very large, the choice of prior probabilities is potentially important. For a simple prior on J, we assume that within each interval, each sample becomes a carrier independently of the other samples with probability π. Let $P(J) = P_\pi(J) = \pi^{|J|}(1-\pi)^{(N-|J|)}$. The contribution to the log likelihood of the m carriers $J_1, \ldots, J_m$ is

$$\begin{aligned}\sum_{i=1}^{m} \log P(J_i) &= \sum_{i=1}^{m} [|J_i| \log \pi + (N - |J_i|) \log(1-\pi)] \\ &= M \log \pi + (Nm - M) \log(1-\pi).\end{aligned}$$

The analysis can be easily adapted to the case where each interval is assigned a different carrier probability π_i. An alternative that seems to have some advantages when we are confident that the proportion of carriers of each individual change is not too large is the Poisson prior $P_\pi(J) = \pi^{|J|} \exp(-N\pi)$.

For each $i = 1, \ldots, m$, and $j \in J_i$, define

$$s_{ij} = \min\{k > i : j \in J_k\},$$

that is, $t_{s_{ij}}$ is the last time after the ith change-point before the j-th sample changes again. For each $i = 1, \ldots, m$, $j \in J_i$, and $\mathbf{t} \in \mathcal{D}$, define

$$X_{i,j}(t) = \frac{t_i S_{t_{s_{ij}},j}}{t_{s_{ij}}} - S_{t_i,j}$$

and

$$\delta_{i,j} = \mu_{i,j} - \mu_{i-1,j}.$$

We introduce the mapping

$$\zeta(i,j) = \sum_{l=1}^{i-1} |J_l| + \sum_{l=1}^{j} I\{l \in J_i\}, \quad \text{for } j \in J_i; \ i = 1, \ldots, m.$$

Let X and $\boldsymbol{\delta}$ be vectors of length M, with

$$X_k = X_{\zeta^{-1}(k)}, \quad \delta_k = \delta_{\zeta^{-1}(k)}, \quad k = 1, \ldots, M. \tag{5.2}$$

The covariance matrix of X, $\Sigma(\mathbf{t})$, now has values

$$\Sigma_{\zeta(i,j),\zeta(i',j)}(\mathbf{t}) = \begin{cases} t_i(1 - t_i/t_{i'}), & 1 \le i < i' \le m; \ j \in J_i \cap J_{i'}, \\ 0, & \text{otherwise.} \end{cases} \tag{5.3}$$

The log-likelihood ratios $\ell(\boldsymbol{\delta}, \mathbf{t})$ and $\ell(\mathbf{t})$ for this new model have the same form as (3.2) and (3.3), with the new definitions for $X(\mathbf{t})$ and $\Sigma(\mathbf{t})$. As before, we make the uniform prior assumption for $\boldsymbol{\tau}$, and a hierarchical Gaussian prior assumption for $\boldsymbol{\delta}$. Let $M_{m,J}$ denote the model with m change-points and carrier sets J, and M_0 denote the model with no change-points. It is not difficult to generalize (3.11) to yield the approximation (for the g-prior for $\boldsymbol{\delta}$)

$$\log \frac{P(M_{m,J}|y)}{P(M_0|y)} = \mathrm{BIC}_1^{\pi}(m, J) + o_p(m), \tag{5.4}$$

where

$$\begin{aligned}\mathrm{BIC}_1^{\pi}(m, J) = \ell(\hat{\boldsymbol{\tau}}) - \frac{1}{2} M\{\log[2\ell(\hat{\boldsymbol{\tau}})/M]\}^+ - \frac{M}{2} - \log|D(m, T)| \\ -m(\kappa_1 - \kappa_2) - \sum_{i=1}^{m} \log\left[\sum_{j \in J_i} \hat{\delta}_{i,j}^2\right] + \sum_{i=1}^{m} \log P_{\pi}(J_i).\end{aligned} \tag{5.5}$$

An outline of the derivation of (5.4)-(5.5) is given in the Appendix.

Since the carrier probability π is not known a priori, we estimate it empirically by

$$\hat{\pi} = \hat{\pi}(m, J) = \operatorname*{argmax}_{\pi} \mathrm{BIC}_1^{\pi}(m, J) = \frac{M}{Nm}.$$

If, instead of a global value, we assigned to each change-point τ_i its own carrier probability π_i, the last term of (5.5) would be replaced by $\sum_{i=1}^{m} \log P_{\pi_i}(J_i)$, where the empirical estimate of π_i would be $\hat{\pi}_i = |J_i|/N$. For the Poisson prior, the maximized term would be $|J| \log(M/Nm) - |J|$. Substituting the empirical estimate $\hat{\pi}$ for π in (5.5), we estimate J, m by

$$(\hat{m}, \hat{J}) = \operatorname*{argmax}_{m,J} \mathrm{BIC}_1^{\hat{\pi}(m,J)}(m, J).$$

Under this mixture model, the BIC criterion (5.5) compares models by weighing the total gain in log-likelihood against penalties incurred by the new $\boldsymbol{\delta}$ and $\boldsymbol{\tau}$ parameters, as well as by the uncertainty regarding which new $\boldsymbol{\delta}$ parameters to introduce.

Given m, finding the optimal $\hat{J}$ requires that we search 2^{mN} models. Since this is not feasible in practice if either m or N is moderately large, we designed a greedy approach that seems to work well; it is described in Supplementary Methods.

5.2. Baseline model

The baseline model is a slight variation that can be useful when the intervals of changed mean values are relatively few and short within a largely stable set of

sequences, and between intervals of change the mean level reverts to a "baseline." With a slight change in notation, we take the intervals of changed mean values to be $\{(\tau_{i,1}, \tau_{i,2}] : \ i = 1, \ldots, m\}$, where for $t = \tau_{i,1} + 1, \ldots, \tau_{i,2}$,

$$y_{t,j} \sim N(\mu_{0,j} + \delta_{i,j}, \sigma_j^2) \quad \text{if } j \in J_i \subseteq \{1, \ldots, N\}.$$

Otherwise, $y_{t,j} \sim N(\mu_{0,j}, \sigma_j^2)$. In this model the process is assumed to return to the baseline after visiting a changed state. For simplicity, we assume that the baseline mean values $\{\mu_{0,j} : \ j = 1, \ldots, N\}$ are all 0. Within every changed interval, we allow each carrier to have its own mean value.

The log-likelihood ratio function $\ell(\mathbf{t}, \boldsymbol{\delta})$ for this model is the same as (3.2), but with $X(\mathbf{t})$ changed to

$$X_{i,j}(\mathbf{t}) = S_{t_{i,2},j} - S_{t_{i,1},j}, \ i = 1, \ldots, m; \ j \in J_i, \tag{5.6}$$

and $\Sigma(\mathbf{t}) = \mathrm{Cov}[X(\mathbf{t})]$ adjusted appropriately. Then X and $\boldsymbol{\delta}$ can be vectorized as in (5.2) with the same definition of the function ζ. The estimates for $\boldsymbol{\delta}$ are

$$\hat{\delta}_{i,j} = X_{i,j}(\mathbf{t})/(t_{i,2} - t_{i,1}), \quad i = 1, \ldots, m; \ j \in J_i. \tag{5.7}$$

As before, let $\boldsymbol{\tau}$ follow a uniform prior over D_m and let $\boldsymbol{\delta}$ follow the mixture of Gaussians prior. Then, the approximation for the log of the Bayes factor replaces (5.5) by

$$\begin{aligned} \mathrm{BIC}_1^{\pi}(m, J) = \ell(\hat{\boldsymbol{\tau}}) - \frac{1}{2} M \{\log [2\ell(\hat{\boldsymbol{\tau}})/M]\}^+ - \frac{M}{2} - \log |D(2m, T)| \\ -2m(\kappa_1 - \kappa_2) - 2 \sum_{i=1}^{m} \log [\sum_{j \in J_i} \hat{\delta}_{i,j}^2] + \sum_{i=1}^{m} \log P_{\pi}(J_i). \end{aligned} \tag{5.8}$$

A brief outline of the derivation is given in the Appendix. Comparing (5.8) to (5.5), the only difference is that m is replaced by $2m$ in the penalty for $\boldsymbol{\tau}$,

$$-\log |D(2m, T)| - 2m(\kappa_1 - \kappa_2) - 2 \sum_{i=1}^{m} \log [\sum_{j \in J_i} \hat{\delta}_{i,j}^2].$$

The reason is that in the baseline model m is the number of changed intervals, and there are $2m$ change-points. Note that there is still only one jump parameter per carrier per interval, and thus the penalty for $\boldsymbol{\delta}$ remains unchanged.

6. Multi-sequence Simulation Study

We call our method MSCBS-MBIC for **M**ulti-**s**ample **CBS** with **M**odified **BIC** model selection. The supplementary materials contains more details on the

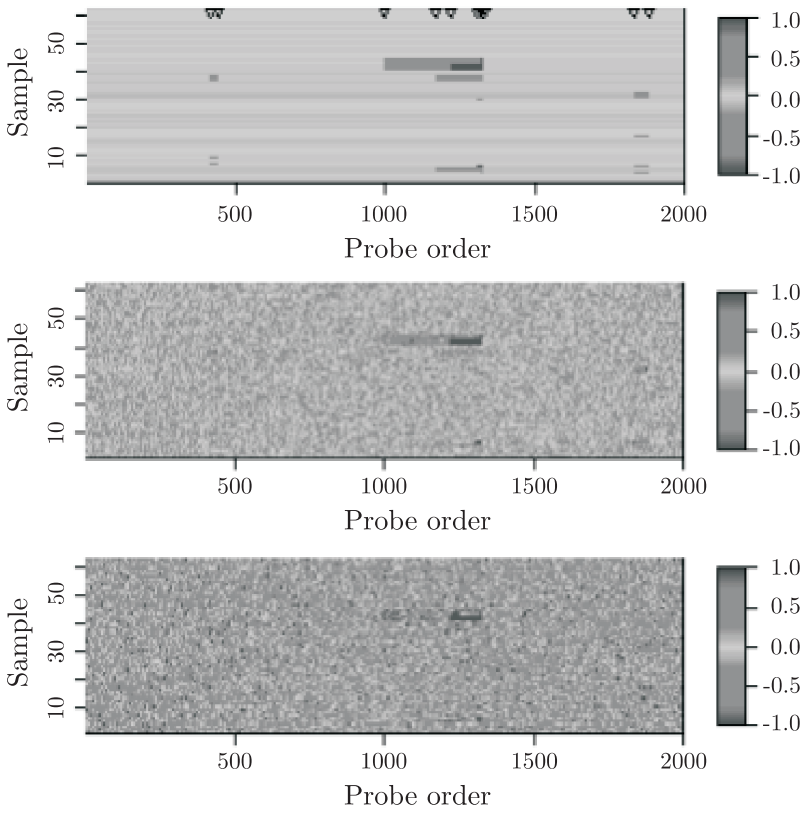

Figure 2. Examples of simulated multi-sequence data sets. The uppermost plot shows the signal matrix with no added noise. The center and bottom plots show the simulation data sets obtained by adding noise with standard deviations 0.2 and 0.4, respectively.

implementation of this method. We first examine the accuracy of the MSCBS-MBIC using simulations. Instead of simulating data from scratch, for more objectivity we used a segmented version of the experimental data described below in Section 7.2 to obtain a matrix of signals. Then, we perturbed that matrix by adding independent Gaussian noise, with varying intensity, to obtain simulated data with varying levels of difficulty. Figure 2 shows the signal matrix and simulated data obtained by adding to the signal matrix independent Gaussian noise with standard deviations 0.2 and 0.4, respectively. The error standard deviation for actual DNA copy number data is typically between 0.1 and 0.3. We experimented with noise standard deviations from 0.05 to 0.4. At 0.05, almost all of the change-points (and carriers) are visibly obvious, whereas we see from Figure 2 that at 0.4 only the strongest signals can be visually identified.

We are interested in the accuracy of both the detection of change-points and

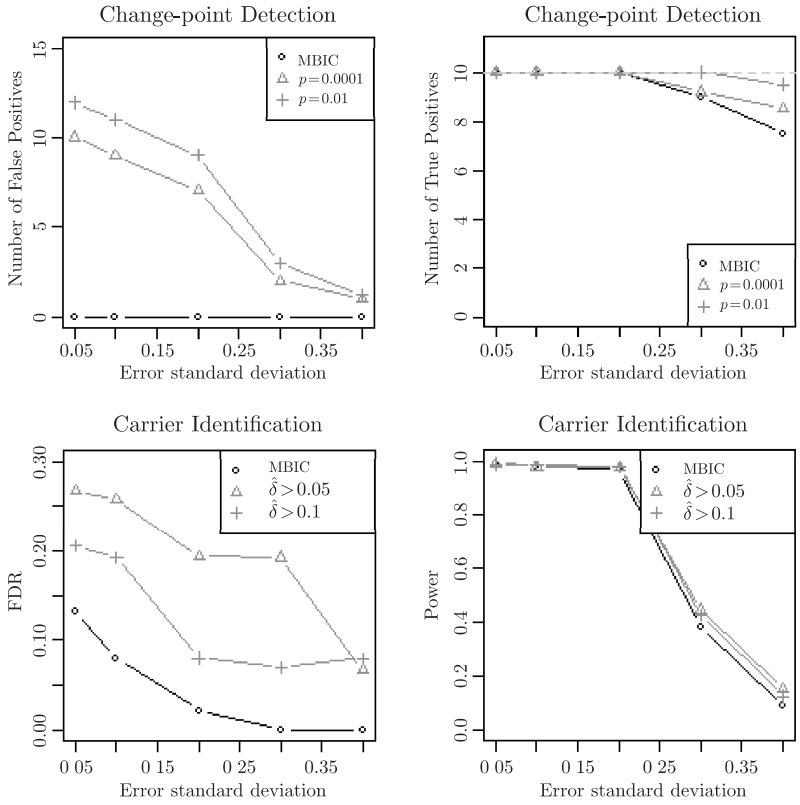

Figure 3. Accuracy of change-point detection and carrier identification on simulated multi-sequence simulation data. The top left and right plots show, respectively, the number of false positives and the number of true positives among the detected change-points. The bottom left and right plots show, respectively, the FDR and power of carrier identification.

the identification of carriers. The signal matrix contains $m = 10$ change-points shown as inverted black triangles in the top plot of Figure 2. For estimated change-points $\mathbf{t} = (t_1, \ldots, t_{\hat{m}})$, we let $\hat{\tau}_i = t_{i^*}$, where $i^* = \operatorname{argmin}_{1 \le j \le \hat{m}} |t_j - \tau_i|$. For each i, the true change-point τ_i was considered detected if $|\tau_i - \hat{\tau}_i| \le h$. We used a threshold of $h = 5$ because we found the estimation of change-point boundaries to be fairly accurate for both methods, and usually within 5 of the true change-point. Once we have a mapping between the estimated and true change-points, we have an estimated carrier set $\hat{J}_i$ for the true carrier set J_i of every change-point $i = 1, \ldots, m$. We can then define true and false positives for change-point detection and carrier identification following standard definitions.

The only comparable method for simultaneous change-point estimation for data of this size and complexity is our previous algorithm (acronym MSCBS,

Zhang et al. (2010)), which uses a recursive circular binary segmentation approach with p-value based stopping criterion and thresholding-based carrier identification. That method relies on, and is quite sensitive to, several user specified parameters: the p-value threshold for stopping the recursion and thresholds for the absolute difference in estimated mean values and related χ^2 test statistics for calling the carriers. Thus, we evaluated the accuracy of MSCBS-MBIC, which is almost hands free, to MSCBS for different threshold values.

Figure 3 shows the results of the simulations. For change-point detection, results for MSCBS-MBIC and MSCBS with p-value stopping criterion at thresholds 0.01 and 0.0001 are shown. We see that when the noise variance is low or moderate (between 0.05 and 0.25), the p-value stopping criterion often oversegments, whereas the modified BIC criterion is able to estimate exactly the number of change-points and produces no false positives. As a trade-off, when the error variance is relatively large (at 0.25-0.35), BIC incurs a slight loss in power. For estimation of the carrier set, results for the MBIC and two different thresholds for the absolute difference in estimated mean values ($|\hat{\delta}_{ij}| > 0.05$ and $|\hat{\delta}_{ij}| > 0.1$) are shown. The trend is the same: the modified BIC criterion significantly reduces the number of false positives with only a slight loss of power when the data are noisy.

7. Cross-sample Summary of DNA Copy Number Data

7.1. Background of application

The DNA copy number of a genomic location is defined as the number of copies of the DNA in that location within the genome of the sample. Advances in microarray and high throughput sequencing technologies within the last decade have enabled the genome-wide fine scale profiling of DNA copy number in high throughput experiments (Pinkel et al. (1988); Pollack et al. (1999); Snijders et al. (2001); Bignell et al. (2004); Peiffer et al. (2006)), leading to systematic studies aimed at using copy number polymorphisms to track distant relationships in population genetics and at understanding the possible role of DNA copy number changes in human disease. For each biological sample, these experiments produce a sequence of intensity measurements for probes that can be mapped to unique locations along the chromosomes. The intensities quantify the average copy number of the corresponding probe's target DNA over the cells in the sample, which can then be compared to that from a control sample. In our model, y_{it} is the normalized intensity for the t-th probe in the i-th sample. See Bengtsson et al. (2008), Peiffer et al. (2006), and Zhang et al. (2010) for normalization methods involved in the pre-processing of raw DNA copy number data.

Changes in DNA copy number can be either inherited or somatic. Inherited changes, often called copy number variants (CNV), comprise a large category

of genetic polymorphism in the human population. Carriers of a given CNV in a population often share the same change-points in their measured intensity profiles, presumably because they are carrying a mutation that occurred in the ancestral history of their DNA. Many inherited CNV are short and spaced far apart, so the base-line model described in Section 5.2 is appropriate. However, some inherited CNV reside in highly mutable regions of the genome, e.g., regions containing segmental duplications (Sharp et al. (2005)). In these cases different breakpoints are often observed among carriers of CNV in the region. In Section 7.2 we analyze one such complex region, that contains different overlapping or nested copy number variants. Other methods have been proposed for multisample segmentation of copy number data (Shah et al. (2007); Lipson et al. (2006)), but unlike our methods they do not allow nested changes or variants that change in opposite directions (i.e., simultaneous gains and deletions) across the cohort.

We also analyze a region from a set of 44 pediatric leukemia samples studied in Schiffman et al. (2009). In tumor samples, most copy number changes are due to somatic mutation events that occur during the growth of the tumor. Since copy number aberrations from different biological samples are due to different somatic mutations, we do not expect a priori to find shared breakpoints across samples. However, there is empirical evidence that copy number aberrations occur at unevenly distributed "hot spots," where they often re-use the same breakpoint junctions (Korbel et al. (2007)). In the analysis of tumor samples, it is of interest to find common deleted or amplified regions across samples, as these regions may harbor genes related to the growth of the tumor. Most of the existing approaches for cross-sample modeling of tumors focus on finding significant "overlaps" between samples after segmentation (Newton et al. (1998); Newton and Lee (2000); Diskin et al. (2006); Beroukhim et al. (2007); Guttman et al. (2007); Rouveirol et al. (2006)).

Thus, in the analysis of both inherited and somatic copy number changes, it is useful to obtain a sparse cross-sample summary of the data as a dimension reduction device for downstream analyses.

7.2. Complex inherited structural variation on chromosome 22

As a first example, we analyze the region on cytoband 11 on the q-arm of chromosome 22. Although small, this region makes a good test case because of its complexity. As documented in the Database of Genomic Variants by Iafrate et al. (2004), the region harbors nested deletions with varying break-points between individuals in the population. Our data come from a set of 62 Illumina 550K Beadchips described in Zhang et al. (2010). We focus on a 2,000 marker segment of the data covering the region of interest, shown in the top panel of Figure 4.

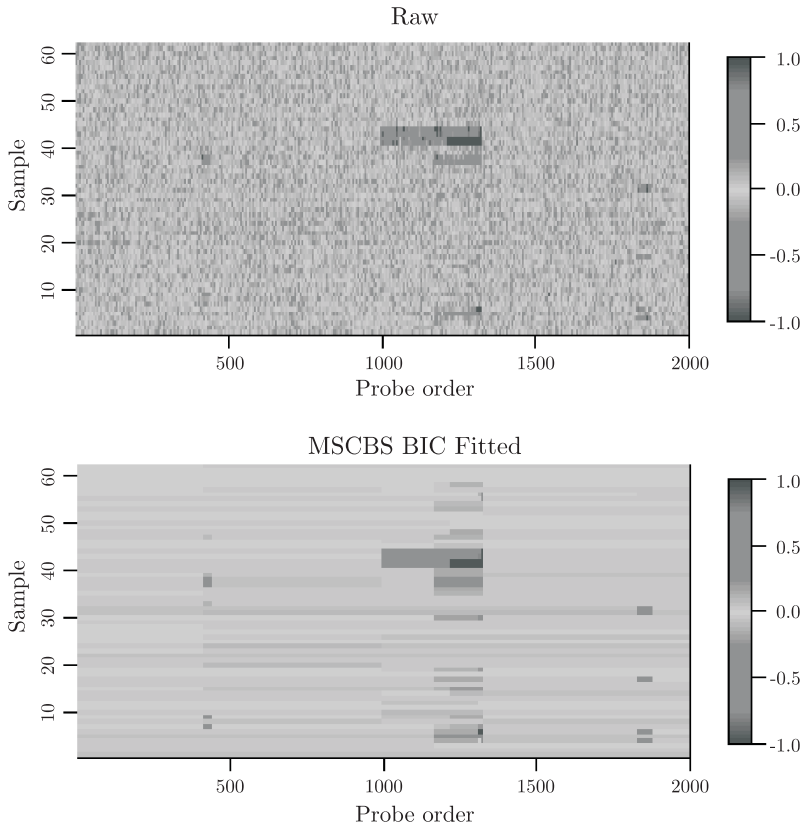

Figure 4. Region of chromosome 22 containing a complex structural variation.

The MSCBS-MBIC algorithm described above was applied to yield the summary shown in the bottom panel of Figure 4.

The segmentation in Figure 4 captures all of the obvious variant regions, with no obvious false positives, as all of the break-points match records in the database of genomic variants.

To assess the accuracy in the identification of carriers, we use the method described in Zhang et al. (2010). The 62 samples represent 10 parent-parent-child trios and 16 pairs of technical replicates. As a result we can make comparisons within nuclear families and between technical replicates to assess the concordance of carriers. A break-point in one of the replicate pairs is counted as concordant if it is shared by the other sample in the pair. Similarly, a break-point in a child is counted as concordant if it is carried by one or both of the child's parents. Thus, we can tally the number of concordant calls among the total number of calls made in the child and replicate samples; see Zhang et al. (2010) for more details. Out of the 153 changes in mean identified over all of the samples, 130

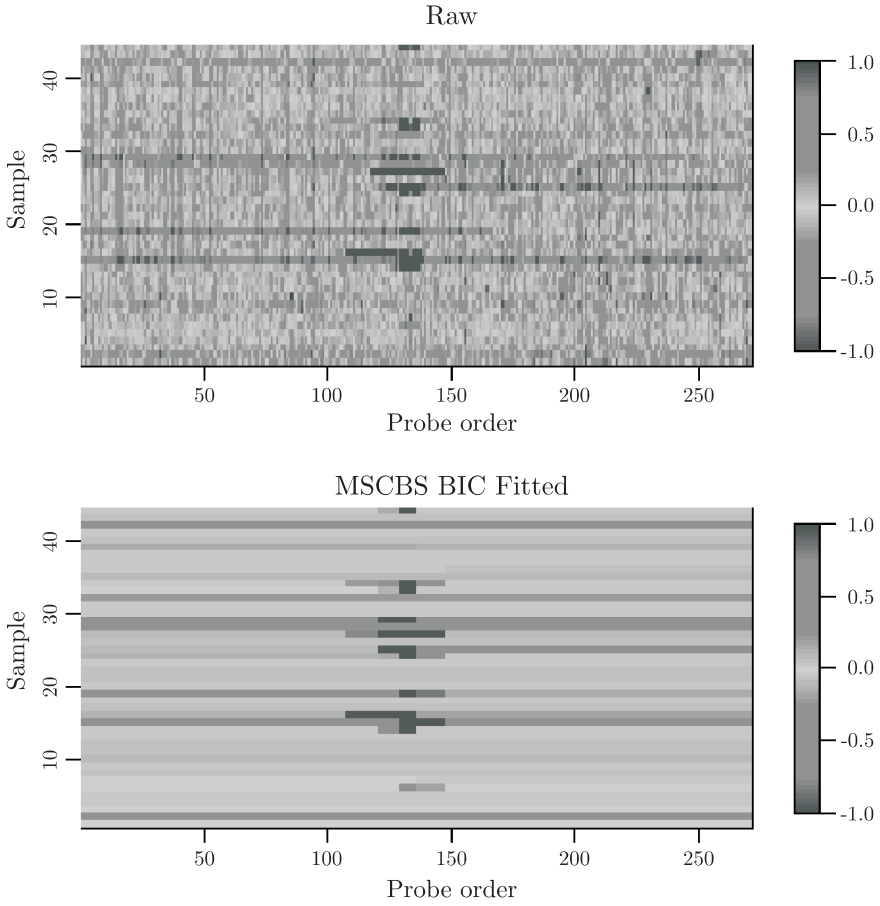

Figure 5. Chromosome 9p in 44 Leukemia samples.

(85%) are concordant. These 153 mean shifts are distributed over 10 change-points. Thus, on average, the carrier proportion at each change-point is 8%. In comparison, the segmentation given in Zhang et al. (2010), which relies on user selected median and p-value thresholds to identify carriers, contains 764 mean shifts distributed over 17 change-points. At most of the change-points, over half of the samples were labeled as carriers. Out of these 764 mean shifts, 533 (70%) are concordant. Hence in comparison with our earlier method, our suggested BIC increases accuracy at the cost of detecting far fewer variant regions.

7.3. Somatic aberrations in chromosome 9p in leukemia

As a second example, we consider the chromosome 9p region in 44 pediatric leukemia samples, that were analyzed using Molecular Inversion Probe technology in Schiffman et al. (2009). Figure 5 shows this region, which was covered by

276 probes in the assay. It is visually obvious that this region contains overlapping deletions, and in some samples there seems to be a sub-region of homozygous deletion nested within a larger hemizygous deletion. Signatures such as this, where multiple samples carry deletions that overlap a small genomic region, often point to key driver genes in the overlapped region. We see that MSCBS-MBIC reconstructs this region quite well, catching both the hemizygous deletion and the nested homozygous deletion. The homozygous deletion centering on probe number 140 encompasses 15,027 bp in length and maps to chr9:21,962,445-21,977,472 (NCBI Build 35.1). This region belongs to the *CDKN2A* locus, which encodes a tumor suppressor gene. The clinical significance of *CDKN2A* deletion in childhood leukemia has been reported previously (Okuda (1995), Heyman et al. (1996)) and its detection in multiple studies of childhood leukemia argues to its importance. Many of the hemizygous and homozygous deletions of CDKN2A found in these data have been validated by RT-PCR. See Schiffman et al. (2009) for a complete analysis of these data, including downstream analyses involving disease subtype and age of onset.

8. Discussion

The original BIC method of Schwarz (1978) has both a philosophical aspect and a technical calculation. In the approximation to the (log) posterior probability suggested by Schwarz, terms that depend on the data dominate terms that depend on the prior, and thus the latter are ignored in the BIC. This makes the BIC attractive from both frequentist and Bayesian points of views. The method has come to be viewed as a penalized likelihood method with a particular penalty function. This viewpoint suggests that the BIC might be applied to a wider range of problems, which may not satisfy the original assumptions of Schwarz, with the possibility that these applications will give disappointing or misleading results.

In Zhang and Siegmund (2007) we observed that Schwarz's assumptions are not satisfied for change-point models; and we calculated an appropriate BIC criterion while keeping the dimension of the model bounded as the number of observations increased. In this paper we have derived a BIC criterion for high-dimensional change point models that may also involve multiple sequences. In this new setting the role of the prior cannot be neglected entirely, since a large model space means that some models inevitably have small prior probability. However, we showed that some features of the prior can still be neglected, and some prior parameters can be estimated by empirical Bayes methods.

In the BIC for high-dimensional, multi-sequence change point models, many terms can be understood as penalties associated with certain parameters. This viewpoint is helpful for interpreting the criterion. It is important to remember

that they are derived from the posterior model probability and thus are not user-tunable regularization terms as appear in *ad hoc* penalties, except to the limited extent that the penalties vary with the prior distribution used.

In Zhang et al. (2010) and Siegmund, Yakir, and Zhang (2011), we treated multi-sample segmentation as a hypothesis testing problem, and proposed p-value approximations which can also be used to determine the number of change-points in the data. In those papers we focused on genome-wide scans for inherited CNVs, where almost all of the data are null. In this case, concepts of "null distribution" and "false positive rate" are useful. However, the two examples we analyzed in Section 7 involve complex, nested signals that make up a large proportion of the entire data set. In this case, a global false positive rate is difficult to interpret, and the problem falls more naturally into an estimation framework.

In a direct comparison with p-value based stopping rules (Olshen et al. (2004); Zhang et al. (2010); Siegmund, Yakir, and Zhang (2011)), the modified BIC method is virtually always more conservative in the number of change-points detected. Whether this is a virtue may depend on the goals of a specific study. It is worth noting, however, that as originally observed by Olshen et al. (2004), copy number data tend to have "local trends," which are oscilations that even elaborate pre-processing fails to remove completely. These local trends show up in the data as intervals of change in level, although biologically they are not intervals of CNV.

In this paper, we focus on detecting changes in mean with independent Gaussian errors. This simple set up allows us to focus on the fundamental issues in high-dimensional change-point models and in multi-sequence change-point detection. The framework we develop can be generalized to more complicated change-point models, and other types multi-sequence signal detection schemes.

Appendix

A.1. Derivation of (3.20) and (3.21)

Let $U(\mathbf{t}) = \Sigma(\mathbf{t})^{-1/2} X(\mathbf{t})$, so $U(\mathbf{t})$ is an m-dimensional random process indexed by $\mathcal{D}$ with elements

$$U_i(\mathbf{t}) = \frac{t_i S_{t_{i+1}}/t_{i+1} - S_{t_i}}{[t_i(1 - t_i/t_{i+1})]^{1/2}}, \ i = 1, \ldots, m. \tag{A.1}$$

Let $Z(\boldsymbol{r}) = U(\boldsymbol{\tau})'[U(\boldsymbol{\tau} + \boldsymbol{r}T) - U(\boldsymbol{\tau})]$. Conditional on $U(\boldsymbol{\tau})$, $Z(\boldsymbol{r})$ is a Gaussian process with mean

$$m(\boldsymbol{r}) = -U(\boldsymbol{\tau})'[I - B(\boldsymbol{r}, \mathbf{0})]U(\boldsymbol{\tau}), \quad \boldsymbol{r}T \in \mathcal{D} \tag{A.2}$$

and covariance

$$\sigma(\boldsymbol{r}, \mathbf{s}) = U(\boldsymbol{\tau})'[B(\boldsymbol{r}, \mathbf{s}) - B(\boldsymbol{r}, \mathbf{0})B(\mathbf{s}, \mathbf{0})']U(\boldsymbol{\tau}), \quad \boldsymbol{r}T, \mathbf{s}T \in \mathcal{D}, \tag{A.3}$$

where $B(\boldsymbol{r},\mathbf{s}) = \mathrm{Cov}[U(\boldsymbol{\tau}+T\boldsymbol{r}), U(\boldsymbol{\tau}+T\mathbf{s})]$ is an $m \times m$ covariance matrix. Let $\boldsymbol{\rho} = \boldsymbol{\tau}/T$. To describe $B(\boldsymbol{r},\mathbf{s})$, for notational simplicity we define $a_i = \rho_i + r_i$, $b_j = \rho_j + s_j$. The entries $B_{i,j}(\boldsymbol{r},\mathbf{s})$ are 0 if $(a_i, a_{i+1}) \cap (b_j, b_{j+1}) = \emptyset$, otherwise they take values

$$(a_i \wedge b_j)\left(1 - \frac{a_i \vee b_j}{a_{i+1} \wedge b_{j+1}}\right)\left[a_i b_j \left(1 - \frac{a_i}{a_{i+1}}\right)\left(1 - \frac{b_j}{b_{j+1}}\right)\right]^{-1/2}.$$

Let

$$\epsilon = \min_i \frac{(\rho_{i+1} - \rho_i)}{2}. \tag{A.4}$$

We can verify that $B(\boldsymbol{r},\mathbf{s})$ satisfies the following properties:

(i) For $\boldsymbol{r},\mathbf{s}$ satisfying $\max_i r_i < \epsilon$, $\max_i s_i < \epsilon$, $B_{i,j}(\boldsymbol{r},\mathbf{s}) = 0$ for all $|i-j| > 1$.

(ii) For $\boldsymbol{r},\mathbf{s}$ satisfying $\min_i r_i < \epsilon$, $\min_i s_i < \epsilon$, only one of the entries $B_{i-1,i}(\boldsymbol{r},\mathbf{s})$ and $B_{i,i-1}(\boldsymbol{r},\mathbf{s})$ can be non-zero, with value

$$a_i(1 - \frac{b_{i+1}}{a_{i+1}})[a_i b_{i+1}(1 - \frac{a_i}{a_{i+1}})(1 - \frac{b_{i+1}}{b_{i+2}})]^{-1/2} = O(\epsilon).$$

(iii) At $\boldsymbol{r} = 0$, the left- and right- partial derivatives of the entries of $B(\boldsymbol{r}, 0)$ exist, are equal in absolute value, but have opposite signs.

We approximate the terms in $B(\boldsymbol{r},\mathbf{s})$ and $B(\boldsymbol{r}, 0)$ by their Taylor series expansion at $\boldsymbol{r} = \mathbf{s} = \mathbf{0}$. Let the absolute value of the left- and right- partial derivatives of $B_{j,k}(\boldsymbol{r},\mathbf{0})$ with respect to r_i be $M_{i,j,k}$, and let M be the array $\{M_{i,j,k} : \ i,j,k = 1,\ldots m\}$. It helps to think of M as a "matrix of vectors" where each element is the absolute directional gradient of the corresponding element of B with respect to $\boldsymbol{r}$, evaluated at $\boldsymbol{r} = \mathbf{0}$. We denote by M^i the i-th "slice" of M,

$$M^i = \{M_{i,j,k} : \ j,k = 1,\ldots,m\}.$$

We use the following notations for the two types of multiplication of M by a vector,

$$(M\mathbf{v})_{i,j} = \sum_{k=1}^{m} M_{j,i,k} v_k, \ \text{for } \mathbf{v} \in \Re^m,$$

$$(M \cdot \mathbf{u})_{i,j} = \sum_{k=1}^{m} M_{k,i,j} u_k, \ \text{for } \mathbf{u} \in \Re^m.$$

It is worth noting a useful property involving these two types of multiplication:

$$(M \cdot \mathbf{u})\mathbf{v} = (M\mathbf{v}) \cdot \mathbf{u}. \tag{A.5}$$

Now, we can express $B(\boldsymbol{r}, 0)$ and $B(\boldsymbol{r}, \mathbf{s})$ in their Taylor expansions as

$$B(\boldsymbol{r}, \mathbf{0}) = I + M \cdot \boldsymbol{r} + o(\|\boldsymbol{r}\|),$$

$$B(\boldsymbol{r}, \mathbf{s}) = I + M \cdot \boldsymbol{r} + M \cdot \mathbf{s} + o(\|\boldsymbol{r}\| + \|\mathbf{s}\|).$$

Substituting the above in (A.2) and (A.3) and using property (A.5), we have the following first order approximations to the variance and covariance functions of the process $Z(\boldsymbol{r})$ conditioned on $U(\boldsymbol{\tau})$.

$$\begin{aligned} m(\boldsymbol{r}) &= -U(\boldsymbol{\tau})'[M \cdot \boldsymbol{r}]U(\boldsymbol{\tau})[1 + E_1(\boldsymbol{r})] \\ &= -C(\boldsymbol{\tau})'\boldsymbol{r}[1 + E_1(\boldsymbol{r})], \end{aligned} \tag{A.6}$$

$$\begin{aligned} \sigma(\boldsymbol{r}, \boldsymbol{r}) &= 2U(\boldsymbol{\tau})'[M \cdot \boldsymbol{r}]U(\boldsymbol{\tau})[1 + E_2(\boldsymbol{r})] \\ &= 2C(\boldsymbol{\tau})'(\mathbf{t} - \boldsymbol{\tau})[1 + E_2(\boldsymbol{r})], \end{aligned} \tag{A.7}$$

$$\begin{aligned} \sigma(\boldsymbol{r}, \mathbf{s}) &= 2U(\boldsymbol{\tau})'[M \cdot (\boldsymbol{r} - \mathbf{s})]U(\boldsymbol{\tau})[1 + E_3(\boldsymbol{r}, \mathbf{s})] \\ &= 2C(\boldsymbol{\tau})'(\boldsymbol{r} - \mathbf{s})[1 + E_3(\boldsymbol{r} \wedge \mathbf{s})], \end{aligned} \tag{A.8}$$

where $C(\boldsymbol{\tau})$ is an m dimensional vector with entries

$$\begin{aligned} C_i(\boldsymbol{\tau}) &= [U(\boldsymbol{\tau})'MU(\boldsymbol{\tau})]_i \\ &= U(\boldsymbol{\tau})'M^iU(\boldsymbol{\tau}), \quad i = 1, \ldots, m, \end{aligned}$$

and $E_1(\boldsymbol{r})$, $E_2(\boldsymbol{r})$, and $E_3(\boldsymbol{r})$ are random variables that satisfy

$$\lim_{\|\boldsymbol{r}\| \to 0} E_i(\boldsymbol{r}) = 0 \quad i = 1, 2, 3.$$

Using properties (i) and (ii) of $B(\boldsymbol{r}, \mathbf{s})$, we see that $M^i(j, k)$ is only non-zero in the two by two sub-matrix $i - 1 \le j, k \le i$, with values

$$M^i(i, i) = T^{-1}[2\tau_i(1 - \frac{\tau_i}{\tau_{i+1}})]^{-1}, \tag{A.9}$$

$$M^i(i-1, i-1) = T^{-1}(\frac{\tau_{i-1}}{\tau_i})^2[2\tau_{i-1}(1 - \frac{\tau_{i-1}}{\tau_i})]^{-1} \tag{A.10}$$

and, between $M^i(i-1, i)$ and $M^i(i, i-1)$, only one can be non-zero with value

$$T^{-1}(\frac{\tau_{i-1}}{\tau_i})[\tau_{i-1}\tau_i(1 - \frac{\tau_{i-1}}{\tau_i})(1 - \frac{\tau_i}{\tau_{i+1}})]^{-1/2}. \tag{A.11}$$

This yields

$$C_i = \frac{\hat{\delta}_i^2}{2}, \quad i = 1, \ldots, m, \tag{A.12}$$

where $\hat{\delta}_i$ is defined in (3.8) and (3.9). Thus, for ϵ satisfying (A.4), within a small ϵ neighborhood of 0 the process $Z(\boldsymbol{r})$, conditioned on $U(\boldsymbol{\tau})$, has the same mean and covariance functions as the process

$$\{\hat{\delta}_i W_{i,r_i} - \hat{\delta}_i^2 \frac{|r_i|}{2},\ i = 1, \ldots, m\}, \tag{A.13}$$

where for $i = 1, \ldots, m$,

$$\{W_{i,t} :\ -\infty < t < \infty\} \tag{A.14}$$

are independent Brownian motions.

We look at continuous versions of the cumulative sums process

$$\{S_i :\ i = 1, \ldots, T\}$$

which, for large T, can be embedded into a Brownian motion on the positive real line. Let

$$E_{\boldsymbol{\tau},\mathbf{t}} = \frac{[U(\boldsymbol{\tau}) - U(\boldsymbol{t})]}{[2U(\boldsymbol{\tau})]}.$$

Let $\{c_T\}$ be a sequence satisfying $c_T \to 0$, $Tc_T \to \infty$, and

$$\lim_{T\to\infty} \frac{\min_i(\tau_{i+1} - \tau_i)}{Tc_T} > 1. \tag{A.15}$$

Note that, since $U(\boldsymbol{\tau}) = O(T)$, for $\|\mathbf{t} - \boldsymbol{\tau}\| < Tc_T$, $\|E_{\boldsymbol{\tau},\mathbf{t}}\|/c_T = O(1)$. Then, using the distribution of (A.13) as an approximation to the distribution of $Z(\boldsymbol{r})$, we have

$$\begin{aligned}
&\int_{\|\mathbf{t}-\boldsymbol{\tau}\|<Tc_T} e^{U(\boldsymbol{\tau})'U(\boldsymbol{\tau}) - U(\boldsymbol{t})'U(\boldsymbol{t})} d\mathbf{t} \\
&\quad = \int_{\|\mathbf{t}-\boldsymbol{\tau}\|<Tc_T} e^{U(\boldsymbol{\tau})'[U(\mathbf{t})-U(\boldsymbol{\tau})](1+E_{\boldsymbol{\tau},\mathbf{t}})} d\mathbf{t} \\
&\quad =_d (1 + e_1) \left[\prod_{i=1}^m \hat{\delta}_i^2\right] \int_{\|\boldsymbol{r}\|<c_T} \exp\left[\sum_{i=1}^m (W_{i,r_i} - \frac{|r_i|}{2})\right] dr_1 \ldots dr_m,
\end{aligned} \tag{A.16}$$

where e_1 is $O(c_T)$. Since, as $\|\mathbf{t} - \boldsymbol{\tau}\| \to \infty$,

$$U(\boldsymbol{\tau})'U(\boldsymbol{\tau}) - U(\boldsymbol{t})'U(\boldsymbol{t}) = O_p(\|\mathbf{t} - \boldsymbol{\tau}\|)$$

and

$$P[U(\boldsymbol{\tau})'U(\boldsymbol{\tau}) - U(\boldsymbol{t})'U(\boldsymbol{t}) < 0] \to 1,$$

we have, for the sequence c_T defined above,

$$\lim_{T\to\infty} \sum_{\|\mathbf{t}-\boldsymbol{\tau}\|>Tc_T} e^{U(\boldsymbol{\tau})'U(\boldsymbol{\tau}) - U(\boldsymbol{t})'U(\boldsymbol{t})} = 0. \tag{A.17}$$

Combining (A.16) and (A.17) gives

$$\log\left[\int_{D(m,T)} e^{U(\boldsymbol{\tau})'U(\boldsymbol{\tau})-U(\boldsymbol{t})'U(\boldsymbol{t})}dt\right] =_d \sum_{i=1}^m \log\hat{\delta}_i^2 + \sum_{i=1}^m f(W_i) + o_p(1), \quad \text{(A.18)}$$

where

$$f(W_i) = \log\left(\int_{-\infty}^{\infty} e^{W_{i,t}-|t|/2}dt\right).$$

By the Law of Large Numbers,

$$\sum_{i=1}^m f(W_i) = m\kappa_2 + 0_p(m). \quad \text{(A.19)}$$

This proves (3.21). For the proof of (3.20), observe that, with probability converging to 1,

$$\begin{aligned}\frac{1}{2}\max_{\mathbf{t}\in D}[U(\boldsymbol{\tau})'U(\boldsymbol{\tau}) - U(\mathbf{t})'U(\mathbf{t})] &= \frac{1}{2}\max_{\|\mathbf{t}-\boldsymbol{\tau}\|<Tc_T}[U(\boldsymbol{\tau})'U(\boldsymbol{\tau}) - U(\mathbf{t})'U(\mathbf{t})] \\ &= \max_{\|\mathbf{t}-\boldsymbol{\tau}\|<Tc_T} U(\boldsymbol{\tau})'[U(\boldsymbol{\tau}) - U(\mathbf{t})][1 + E_{\mathbf{t},\boldsymbol{\tau}}] \\ &=_d \sum_{i=1}^m \max_{r_i}(\hat{\delta}_i W_{i,r_i} - \hat{\delta}_i^{\,2}\frac{|r_i|}{2}) + o_p(1) \\ &= \sum_{i=1}^m \max_{r_i}(W_{i,r_i} - \frac{|r_i|}{2}) + o_p(1), \quad \text{(A.20)}\end{aligned}$$

so (3.20) follows by the Law of Large Numbers.

Remarks about (A.18)–(A.20)**.** Let $Y_1, Y_2, \ldots$ be independent standard normal and put $V_n = Y_1 + \ldots + Y_n$ for $n = 0, 1, \ldots$. In (A.18) and (A.20) we have simplified some computations by introducing functionals of Brownian motion in the place of functionals of discrete time random walks. The connecting link is the weak convergence as $\delta \to 0$ of

$$\log\Big\{\sum_n \exp[\delta V_n - \delta^2\frac{|n|}{2}]\delta^2\Big\} \quad \text{(A.21)}$$

and

$$\max_n[\delta V_n - \delta^2\frac{|n|}{2}] \quad \text{(A.22)}$$

to

$$\log\Big\{\int_{-\infty}^{\infty} \exp[W_t - \frac{|t|}{2}]dt\Big\} \quad \text{(A.23)}$$

and

$$\max_t [W_t - \frac{|t|}{2}], \tag{A.24}$$

respectively. Ultimately we are interested in the expectations of these quantities, which are easily computed.

Simulations, which could be used in place of the approximations based on Brownian motion, indicate that the approximation of the expectation of (A.21) by that of (A.23) is very accurate for small δ and reasonably accurate for $|\delta|$ less than about 2.75, where the relative error reaches 10%. Standard calculations show that the expected value of (A.24) is 3/2. The approximation of the expectation of (A.22) by that of (A.24) can be substantially improved for δ different from 0, by appealing to Siegmund (1979), which suggests using $2\nu(\delta) - \nu^2(\delta)/2$, where the function ν is easily evaluated numerically and, for small δ, is approximately $\exp(-0.583\delta)$ Siegmund (1985). Even with this improved approximation, the relative error reaches 10% at about $\delta = 1.4$, where the true value is about 0.71.

A.2. Derivation of results for multi-sequence models

The derivations of (5.4) and (5.8) for the multi-sequence models follow the same logic as for a single sequence. Here, we outline the places where the derivations differ. For each model, let M be the total number of shifts in mean. In the unanimous change-point model $M = mN$, and in the mixture models, M is defined in (5.1). For all three models, the expansion of the Bayes factor follows steps (3.13)-(3.17), except with m replaced by M in all places. In the terms $A(\boldsymbol{\tau})$ and $B(\boldsymbol{\tau})$, the definition of the process $X(t)$ changes to (5.2) or (5.6) under each appropriate model.

In the proofs of (3.20) and (3.21), the definitions of the processes $U(\mathbf{t})$ and $Z(\boldsymbol{r})$ do not need to be changed, as long as the definition for $X(\mathbf{t})$ is adjusted appropriately. We can show using the same, albeit more technically cumbersome steps, that $Z(\boldsymbol{r})$, conditioned on $U(\boldsymbol{\tau})$, has the same mean and covariance functions as

$$\{\Delta_i W_{i,r_i} - \Delta_i^2 \frac{|r_i|}{2},\ i = 1, \ldots, m\}, \tag{A.25}$$

where

$$\Delta_i^2 = \sum_{j \in J_i} \hat{\delta}_{i,j}^2. \tag{A.26}$$

The proof of (3.20) then follows steps (A.16)-(A.18), with $\hat{\delta}_i$ replaced by Δ_i. For the proof of (3.21), note that for each i, $\max_{r_i}(\Delta_i W_{i,r_i} - \Delta_i^2 |r_i|/2)$ has the same distribution as $\max_{r_i}(W_{i,r_i} - |r_i|/2)$, and so (A.20) does not change.

References

Bengtsson, H., Irizarry, R., Carvalho, B. and Speed, T. (2008). Estimation and assessment of raw copy numbers at the single locus level. *Bioinformatics* **24**, 759-767.

Berger, J. O., Ghosh, J. K. and Mukhopadhyay, N. (2003). Approximations and consistency of bayes factors as model dimension grows. *J. Statist. Plann. Inference* **112**, 241-258.

Beroukhim, R., Getz, G., Nghiemphu, L., Barretina, J., Hsueh, T., Lin- hart, D., Vivanco, I., Lee, J. C., Huang, J. H., Alexander, S., Du, J., Kau, T., Thomas, R. K., Shah, K., Soto, H., Perner, S., Prensner, J., Debiasi, R. M., Demichelis, F., Hatton, C., Rubin, M. A., Garraway, L. A., Nelson, S. F., Liau, L., Mischel, Cloughesy, T. F., Meyer- son, M., Golub, T. A., Lander, E. S., Mellinghoff, I. K. and Sellers, W. R. (2007). Assessing the significance of chromosomal aberrations in cancer: Methodology and application to glioma. *Proc. Nat. Acad. Sci.*, 0710052104+.

Bignell, G. R., Huang, J., Greshock, J., Watt, S., Butler, A., West, S., Grigorova, M., Jones, K. W., Wei, W., Stratton, M. R., Futreal, P. A., Weber, B., Shapero, M. H. and Wooster, R. (2004). High-resolution analysis of dna copy number using oligonucleotide microarrays. *Genome Research* **14**, 287-295.

Chen, J. and Chen, Z. (2008). Extended bayesian information criteria for model selection with large model spaces. *Biometrika* **95**, 759-771.

Diskin, S. J., Eck, T., Greshock, J., Mosse, Y. P., Naylor, T., Stoeck- ert Jr., C. J., Weber, B. L., Maris, J. M. and Grant, G. R. (2006). Stac: A method for testing the significance of dna copy number aberrations across multiple array-cgh experiments. *Genome Research* **16**, 1149-1158.

George, E. and Foster, D. (2000). Calibration and empirical bayes variable selection. *Biometrika* **87**, 731-747.

Guttman, M., Mies, C., Dudycz-Sulicz, K., Diskin, S. J., Baldwin, D. A., Stoeckert, C. J. and Grant, G. R. (2007). Assessing the significance of conserved genomic aberrations using high resolution genomic microarrays. *PLoS Genetics* **3**, e143+.

Heyman, M., Rasool, O., Borgonovo, B. L., Liu, Y., Grander, D., Soderhall, S., Gustavsson, G. and Einhorn, S. (1996). Prognostic importance of p15INK4B and p16INK4 gene inactivation in childhood acute lymphocytic leukemia. *J. Clin Oncol.* **14**, (1996), 1512-1520.

Iafrate, J. A., Feuk, L., Rivera, M. N., Listewnik, M. L., Donahoe, P. K., Qi, Y., Scherer, S. W. and Lee, C. (2004). Detection of large-scale variation in the human genome. *Nature Genetics* **36**, 949-951.

Korbel, J. O., Urban, A. E., Affourtit, J. P., Godwin, B., Grubert, F., Simons, J. F., Kim, P. M., Palejev, D., Carriero, N. J., Du, L., Taillon, B. E., Chen, Z., Tanzer, A., Saunders, A. C., Chi, J., Yang, F., Carter, N. P., Hurles, M. E., Weissman, S. M., Harkins, T. T., Gerstein, M. B., Egholm, M. and Snyder, M. (2007). Paired-end mapping reveals extensive structural variation in the human genome. *Science* **318**, 420-426.

Liang, F., Paulo, R., German, M., Clyde, M. A. and Berger, J. O. (2008). Mixtures of g priors for Bayesian variable selection. *J. Amer. Statist. Assoc.* **103**, 410-423.

Lipson, D., Aumann, Y., Ben-Dor, A., Linial, N. and Yakhini, Z. (2006). Efficient calculation of interval scores for dna copy number data analysis. *J. Computational Biology* **13**, 215-228.

Newton, M., Gould, M., Reznikoff, C. and Haag, J. (1998). On the statistical analysis of allelic-loss data. *Statist. Medicine* **17**, 1425-1445.

Newton, M. and Lee, Y. (2000). Inferring the location and effect of tumor suppressor genes by instability-selection modeling of allelic-loss data. *Biometrics* **56**, 1088-1097.

Ninomiya, Y. (2004). Information criterion for gaussian change-point model. *Statist. Probab. Lett.* **72**, 237-247.

Okuda T, Shurtleff, S. A., Valentine, M. B., Raimondi, S. C., Head, D. R., Behm, F., Curcio-Brint, A. M., Liu, Q., Pui, C. H. and Sherr, C. J. (1995). Frequent deletion of p16INK4a/MTS1 and p15INK4b/MTS2 in pediatric acute lymphoblastic leukemia. *Blood* **85**, 2321-2330.

Olshen, A. B., Venkatraman, E. S., Lucito, R. and Wigler, M. (2004). Circular binary segmentation for the analysis of array-based dna copy number data. *Biostatistics* **5**, 557-572.

Peiffer, D. A., Le, J. M., Steemers, F. J., Chang, W., Jenniges, T., Garcia, F., Haden, K., Li, J., Shaw, C. A., Belmont, J., Cheung, S. W., Shen, R. M., Barker, D. L. and Gunderson, K. L. (2006). High-resolution genomic profiling of chromosomal aberrations using infinium whole-genome genotyping. *Genome Research* **16**, 1136-1148.

Pinkel, D., Segraves, R., Sudar, D., Clark, S., Poole, I., Kowbel, D., Collins, C., Kuo, W. L., Chen, C., Zhai, Y., Dairkee, S. H., Ljung, B. M., Gray, J. W. and Albertson, D. G. (1998). High resolution analysis of dna copy number variation using comparative genomic hybridization to microarrays. *Nature Genetics* **20**, 207-211.

Pollack, J., Perou, C., Alizadeh, A., Eisen, M., Pergamenschikov, A., Williams, C., Jeffrey, S., Botstein, D. and Brown, P. (1999). Genome-wide analysis of dna copy-number changes using cdna microarrays. *Nature Genetics* **23**, 41-46.

Pollak, M. and Siegmund, D. (1985). A diffusion process and its applications to detecting a change in the drift of Brownian motion. *Biometrika* **72**, 267-280.

Rouveirol, C., Stransky, N., Hupé, P., La Rosa, P., Viara, E., Barillot, E. and Radvanyi, F. (2006). Computation of recurrent minimal genomic alterations from array-cgh data. *Bioinformatics* **22**, 849-856.

Schiffman, J. D., Wang, Y., Mcpherson, L. A., Welch, K., Zhang, N., Davis, R., Lacayo, N. J., Dahl, G. V., Faham, M. and Ford, J. M. (2009). Molecular inversion probes reveal patterns of 9p21 deletion and copy number aberrations in childhood leukemia. *Cancer Genetics and Cytogenetics* **193**, 9-18.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 461-464.

Shah, S. P., Lam, W. L., Ng, R. T. and Murphy, K. P. (2007). Modeling recurrent dna copy number alterations in array cgh data. *Bioinformatics* **23**, 450-458.

Sharp, A. J., Locke, D. P., McGrath, S. D., Cheng, Z., Bailey, J. A., Vallente, R. U., Pertz, L. M., Clark, R. A., Schwartz, S., Segraves, R., Oseroff, V. V., Albertson, D. G., Pinkel, D. and Eichler, E. E. (2005). Segmental duplications and copy-number variation in the human genome. *Amer. J. Human Genetics* **77**, 78-88.

Siegmund, D. O. (1979). Corrected diffusion approximations in certain random walk problems. *Adv. Appl. Probab.* **11**, 701-719.

Siegmund, D. O. (1985). *Sequential Analysis: Tests and Confidence Intervals*. Springer-Verlag, New York, Heidelberg, Berlin.

Siegmund, D., Yakir, B. and Zhang, N. (2011). Detecting simultaneous variant intervals in aligned sequences. *Ann. Appl. Statist.* **5**, 645-668.

Snijders, A. M., Nowak, N., Segraves, R., Blackwood, S., Brown, N., Conroy, J., Hamilton, G., Hindle, A. K., Huey, B., Kimura, K., Law, S., Myambo, K., Palmer, J., Ylstra, B., Yue, J. P., Gray, J. W., Jain, A. N., Pinkel, D. and Albertson, D. G. (2001). Assembly of microarrays for genome-wide measurement of dna copy number. *Nature Genetics* **29**, 263-264.

Vostrikova, L. (1981). Detecting disorder in multidimensional random process. *Soviet Mathematics Doklady* **24**, 55-59.

Wang, H., Veldink, J. H., Ophoff, R. A. and Sabatti, C. (2008). Markov models for inferring copy number variations from genotype data on illumina platforms. Technical report, Dept. of Statistics, University of California at Los Angeles.

Yao, Y.-C. (1988). Estimating the number of change-point via schwarz' criterion. *Statist. Probab. Lett.* **6**, 181-189.

Zellner, A. (1986). On assessing prior distributions and bayesian regression analysis with g-prior distributions. In *Bayesian Inference and Decision Techniques*: Essays in Honor of Bruno de Finetti, 233-243. North-Holland/Elsevier.

Zhang, N. and Siegmund, D. (2007). A modified bayes information criterion with applications to the analysis of comparative genomic hybridization data. *Biometrics* **63**, 22-32.

Zhang, N., Siegmund, D., Ji, H. and Li, J. Z. (2010). Detecting simultaneous change-points in multiple sequences. *Biometrika* **97**, 631-645.

Department of Statistics, Stanford University, Sequoia Hall, 390 Serra Mall, Stanford, CA 94305-4065, U.S.A.

E-mail: nzhang@stanford.edu

Department of Statistics, Stanford University, Sequoia Hall, 390 Serra Mall, Stanford, CA 94305-4065, U.S.A.

E-mail: dos@stat.stanford.edu

(Received November 2010; accepted September 2011)