Supplementary Material for Model Selection for High Dimensional, Multi-sequence Change-point Problems

Nancy R. Zhang and David O. Siegmund

Department of Statistics, Stanford University

* The starred equation numbers refer to equations in the manuscript.

1 Simulations

1.1 Effect of prior assumptions on the shifts in mean

The first simulation compares the performance of the two criterion BIC_1 and BIC_2 derived under the two prior assumptions for $\boldsymbol{\delta}$. Specifically, the data were generated as follows:

Model 1 The change-points $\boldsymbol{\tau}$ were first sampled from a Poisson process with rate λ on $\{1,\dots,T\}$. Conditioned on $\boldsymbol{\tau}$, $\boldsymbol{\delta}$ was sampled from $N(0, w^{-1}\Sigma^{-1}(\boldsymbol{\tau}))$. Given $\boldsymbol{\tau}$ and $\boldsymbol{\delta}$, y was sampled from (1*).

Model 2 This model is similar, but conditional on $\boldsymbol{\tau}$, $\boldsymbol{\delta}$ is sampled from $N(0, w^{-1}I)$.

The parameters of the models are T, w, and λ. Figure 1 shows the results for $T = 1000$, $\lambda = 0.01$, and a range of values for w. The left plot is for data simulated from Model 1, the right plot for data simulated from Model 2. Let $\hat{m}$ be the number of change-points estimated using either BIC_1 or BIC_2. The mean error $\hat{m} - m$ over 100 simulations is plotted against w^{-1}. Although BIC_1 seems to be slightly preferable to BIC_2 under Model 1, and vice versa for Model 2, as expected, the differences are so small that it seems safe to conclude that the two criteria perform similarly and are relatively insensitive to the covariance matrix of the prior distribution.

1.2 Effect of the number of change-points

We also compared BIC_1 and its variance unknown version to the criterion (2*) from Zhang & Siegmund (2007). In this case we generated sequences from (1*), with change-point locations sampled according to uniform order statistics in $[0, T]$. The variance parameter σ^2 was set to 1, and the mean parameters $\{\mu_i : \ i = 1,\dots,m+1\}$ were sampled independently from $N(0, \sigma_\mu^2)$. Larger σ_μ gave on average larger changes in mean and thus stronger signals for each change-point.

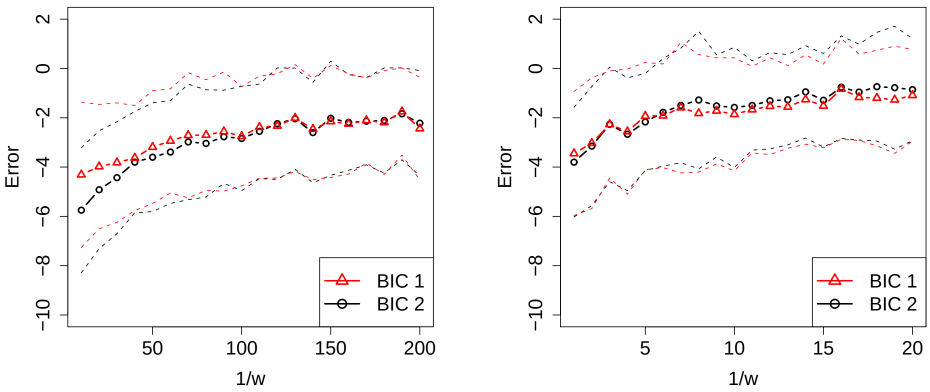

Figure 1: w^{-1} versus $\hat{m} - m$ for data from simulation Model 1 (left plot) and from simulation Model 2 (right plot). The red lines are for BIC_1, black lines are for BIC_2. In each case, the bold solid line shows the mean of $\hat{m} - m$ over 100 simulations, with the dashed lines being the mean ± 1 standard deviation.

We examined two settings: $T = 10000, m = 99$ and $T = 1000, m = 9$. In the first setting, $m = T^{1/2}$ is much larger than $\log T$, and thus we expect BIC_1 to be more accurate than (2*), but not so in the second setting where m is relatively small. For each of $\sigma_\mu = 1, 2 \ldots, 7$, we simulated 100 sequences under both settings and estimated the number of change-points using (2*), BIC_1, and its unknown-variance version as described at the end of Section 3 of the manuscript. Figure 2 shows the average value of $\hat{m}$ in 100 simulations at each value of σ_μ. For small σ_μ, the power for finding the change-points is low, and the estimated number of change-points is usually smaller than the true number of change-points, whereas the converse is true for large σ_μ. As shown by Figure 2, the new approximation BIC_1 achieves more sensitivity when the signal is weak and higher specificity when the signal is strong, at both $T = 10000$ and $T = 1000$. The results also show that the performance of BIC_1 and its variance unknown version are very similar. Thus, not knowing the variance, even for this relatively difficult data set, does not compromise the accuracy of estimating m.

2 Multi-sequence Segmentation Algorithm

In applications of these ideas to high throughput experimental settings such as DNA copy number detection, the length of the sequences can reach over a million, and N can be in the thousands. The size of the space of possible models makes systematic comparison of all models impractical.

In our earlier work (Zhang et al., 2010), we adopted a recursive hypothesis testing approach

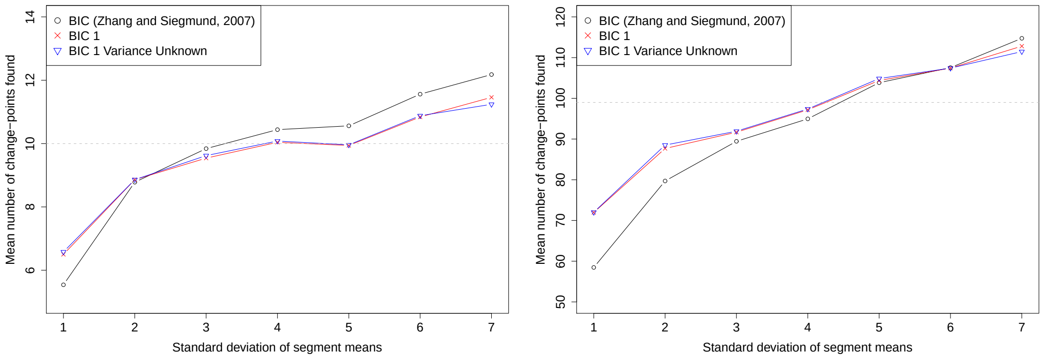

Figure 2: The mean of $\widehat{m}$ versus σ_μ for $T = 1000, m = 9$ (left plot) and $T = 10000, m = 99$ (right plot).

that generalized the Circular Binary Segmentation algorithm (CBS) (Vostrikova, 1981; Olshen et al., 2004) designed for a single sequence. Termination of the algorithm was based on a p-value criterion, and selection of carriers was based on an ad hoc thresholding rule inside each iteration. The BIC procedures described above eliminates the ad hoc aspect of our earlier algorithm and hence gives us a more rational method for carrier selection. We now combine BIC carrier selection with a version of our earlier search algorithm; but in the spirit of trying to eliminate arbitrary features of the algorithm, we no longer use a p-value based criterion for continuation/termination.

Intuitively, the procedure starts by scanning the region $1, \ldots, T$ for a candidate changed interval (s^*, t^*) by computing a "score" for every interval, and then maximizing the score. The candidate interval splits $[1, T]$ into three sub-regions: $(0, s^*]$, $(s^*, t^*]$, and $(t^*, T]$. The next iteration scans each sub-region for a candidate changed interval, and splits at the best candidate among the sub-regions. The BIC criterion is computed at each split.

For each sub-interval $(a, b]$ nested within an interval $(s, t]$ we consider the score

$$Z_{s,t}(a,b) = \max_J \left[\frac{1}{2} \sum_{j \in J} U^2_{s,t,j}(a,b) - |J|/2 + \log P_\pi(J) \right], \tag{1}$$

where for each j

$$U^2_{s,t,j}(a,b) = \frac{[S_{b,j} - S_{a,j} - (b-a)(S_{t,j} - S_{s,j})/(t-s)]^2}{(b-a)[1-(b-a)/(t-s)]}$$

is the χ^2 statistic for testing a square wave change at $(a, b]$ within $(s, t]$. Computing $Z_{s,t}(a, b)$ requires estimating J. Since the maximizing subset must be the sequences with the highest χ^2 values, for any given π, J can be determined by sorting the χ^2 values, which is an $O(N \log N)$ operation.

Below is a detailed description of the algorithm, which we call MSCBS-MBIC for **M**ulti-**s**ample **CBS** with **M**odified **BIC** model selection. In our notation, S, $\mathcal{Z}$, and $\mathcal{R}$ are ordered arrays of elements. At any iteration, S contains the increasing list of change-points, $\mathcal{Z}$ contains the maximized scores (1) for the next split between adjacent pairs of change-points, and $\mathcal{R}$ holds the best splitting locations.

Initialize: M, π

Set $k \leftarrow 0$, $S^{(k)} \leftarrow S \leftarrow \{0, T\}$,

$$Z\text{max} = \max_{0<s<t<T} Z_{0,T}(s,t), \quad (s^*, t^*) = \text{argmax}_{0<s<t<T} Z_{0,T}(s,t),$$

Set $\mathcal{Z} \leftarrow \{Z\text{max}\}$, $\mathcal{R} \leftarrow \{(s^*, t^*)\}$, $BIC(0) \leftarrow 0$.

While $|S| - 2 < M$ **repeat:**

1. Let $l^* \leftarrow \text{argmax}_l \mathcal{Z}(l)$, $(s^*, t^*) \leftarrow \mathcal{R}(l^*)$,

   $$s \leftarrow \max\{s' \in S, s' < s^*\}, \quad t \leftarrow \min\{t' \in S, t' > t^*\}.$$

   For each of $(c, d] \in \{(s, s^*], (s^*, t^*], (t^*, t]\}$, compute

   $$Z\text{max} = \max_{c<a<b<d} Z_{c,d}(a,b), \quad (s^*, t^*) = \text{argmax}_{c<a<b<d} Z_{c,d}(a,b).$$

   Let Z_L, Z_C, and Z_R be respectively the value of $Z\text{max}$ computed for the left segment $(s, s^*]$, the center segment $(s^*, t^*]$, and the right segment $(t^*, t]$. Similarly, let R_L, R_C, R_R be respectively the maximizer for the left, center, and right regions.

2. Let $L = |\mathcal{Z}|$, Set:

   $$\begin{gathered} k \leftarrow k+1, \\ S \leftarrow \{S[1 : l^* - 1], s^*, t^*, S[l^* + 1, L + 1]\}, \\ \mathcal{Z} \leftarrow \{\mathcal{Z}[1 : l^* - 1], Z_L, Z_C, Z_R, \mathcal{Z}[l^* + 1, L]\}, \\ \mathcal{R} \leftarrow \{\mathcal{R}[1 : l^* - 1], R_L, R_C, R_R, \mathcal{R}[l^* + 1, L]\}. \end{gathered}$$

   Given S, compute $\text{BIC}^{\hat{\pi}}(k, \hat{J})$ by optimizing over J and π. Store the current set of change-points in $S^{(k)} \leftarrow S$.

Finally, let $k^* = \text{argmax}_{0 \le k \le M} \text{BIC}^{\hat{\pi}(k,\hat{J})}$. Return the change-points $S^{(k^*)}$ and the estimated carrier sets $\hat{J}$.

The only user specified parameters for this algorithm are M, the maximum number of splits to allow, and π, the initial value of the carrier proportion for the scan. The maximum value of M is usually set to a large number (depending on computational resources). The BIC is computed for every split, and the best model up to M is chosen. Thus, as long as M is not too small (and if m is maximized at M we can increase the value of M), the algorithm is not very sensitive to this parameter. We fix π initially to speed up the scanning process, which is the computational bottleneck of the algorithm. Note that after identification of putative change-points the value of π and the carrier sets are re-estimated by maximizing the BIC, so the reported carrier sets are based on a data determined value of π.

3 Physical location of estimated change-points in Stanford Quality Control Data

The change-points detected on chromosome 22 of the Stanford Quality Control Data in Section 7.2 of the manuscript are listed in the following table.

SNP number	Physical location
417	17,252,341
443	17,382,662
997	20,706,322
1168	21,005,866
1218	21,116,954
1310	21,381,016
1322	21,452,139
1330	21,573,594
1831	23,988,962
1881	24,235,221

4 Derivation of Result for Variance Unknown Case

The derivation for the variance unknown case follows the derivation of Proposition 1, but with an additional integral over σ. With slight abuse of notation, we let $\mathbf{X} = (X_1, \ldots, X_{T-1})$, where

$$X_t = t)S_T/T - S_t, \quad t = 1, \ldots, T.$$

Let $\Sigma = \mathrm{Cov}(\mathbf{X})$. Then, $SS_{\text{all}} = \mathbf{X}'\Sigma^{-1}\mathbf{X}$ and by the same argument as (4*),

$$P_{\mathbf{t},\boldsymbol{\delta},\sigma}(\mathbf{X}) = (2\pi)^{-1/2}\sigma^{-T+1}|\Sigma|^{-1/2} \exp\left\{-\frac{1}{2\sigma^2}[\boldsymbol{\delta}'\Sigma(\mathbf{t})\boldsymbol{\delta} - 2\boldsymbol{\delta}'\mathbf{X}(\mathbf{t}) + SS_{\text{all}}]\right\}.$$

Let $\phi = \sigma^{-2}$, and let $\pi_\phi(\delta)$ be the prior

$$\pi(\boldsymbol{\delta}) = \int_0^\infty g(w) w^{m/2} (2\pi)^{-m/2} |\phi\Gamma(\mathbf{t})|^{1/2} e^{-\frac{w}{2}\boldsymbol{\delta}'\phi\Gamma(\mathbf{t})\boldsymbol{\delta}} dw, \tag{2}$$

where $\Gamma(\mathbf{t}) = \Sigma(\mathbf{t})$ for the g-prior and equals the identity for the independence prior. We compute the posterior probability of the model with m change-points,

$$\begin{aligned} P(M_m|\mathbf{y}) &= |\mathcal{D}|^{-1} \sum_{\mathbf{t}\in D} \int_0^\infty \int_0^\infty \int_{\Re^m} P(\mathbf{t}, \boldsymbol{\delta}, \phi)\pi_\phi(\boldsymbol{\delta}) d\boldsymbol{\delta} d\sigma d\pi \\ &= \sum_{\mathbf{t}} \int_0^\infty g(w) \left(\frac{w}{1+w}\right)^{m/2} [2^{-1}(SS_{\text{all}} - SS_{\text{bg}}(\mathbf{t}))]^{-(T+1)/2} dw \\ &\quad \times |D(m,T)|^{-1} |\Sigma|^{-1/2} (2\pi)^{(T-1)/2} \Gamma[(T+1)/2]. \end{aligned} \tag{3}$$

The posterior probability for the model with no change-points is

$$P(M_0|y) = \Gamma[(T+1)/2](2\pi)^{(T-1)/2}|\Sigma|^{-1/2}[2^{-1}(SS_{\text{all}})]^{-(T+1)/2}. \tag{4}$$

Dividing (3) by (4) and using the identity (suppressing notation in $\mathbf{t}$)

$$\begin{aligned} &\left[\frac{SS_{\text{all}} - SS_{\text{bg}}(1+w)^{-1}}{SS_{\text{all}}}\right]^{-(T+1)/2} \\ &= \left[1 + \frac{SS_{\text{bg}}}{SS_{\text{wg}}}\right]^{(T+1)/2} \left[1 + \left(\frac{w}{1+w}\right)\frac{SS_{\text{bg}}}{SS_{\text{wg}}}\right]^{-(T+1)/2}, \end{aligned}$$

we have

$$\frac{P(M_m|\mathbf{y})}{P(M_0|\mathbf{y})} = |\mathcal{D}|^{-1}\sum_{\mathbf{t}}\left[1 + \frac{SS_{\text{bg}}(\mathbf{t})}{SS_{\text{wg}}(\mathbf{t})}\right]^{(T+1)/2}\int_0^\infty g(w)f(\eta)dw, \tag{5}$$

where $\eta = w(1+w)^{-1}$ and

$$f(\eta) = \eta^{m/2}(1 + \eta SS_{\text{bg}}(\mathbf{t})/SS_{\text{wg}}(\mathbf{t}))^{-(T+1)/2}.$$

Repeating steps (16*)-(19*), but with this new definition of $f(\eta)$, gives

$$\begin{aligned} \frac{P(M_m|\mathbf{y})}{P(M_0|\mathbf{y})} &= |\mathcal{D}|^{-1}\sum_{\mathbf{t}}\left[1 + \frac{SS_{\text{bg}}(\mathbf{t})}{SS_{\text{wg}}(\mathbf{t})}\right]^{(T+1)/2} \\ &\quad \times \exp\left[-\frac{m}{2}\log\left(\frac{(T-m+1)SS_{\text{bg}}(\mathbf{t})}{mSS_{\text{wg}}(\mathbf{t})}\right) - \frac{m}{2} + O_p(1)\right]. \end{aligned} \tag{6}$$

As $T \to \infty$, by Assumption III,

$$\left[1 + \frac{SS_{\text{bg}}(\mathbf{t})}{SS_{\text{wg}}(\mathbf{t})}\right]^{(T+1)/2} = \exp[C(\mathbf{t})SS_{\text{bg}}(\mathbf{t})/(2\hat{\sigma})], \tag{7}$$

with $C(\mathbf{t}) = O_p(1)$. Also by Assumption III,

$$\frac{T+1}{2}\log\left[1 + \frac{SS_{\text{bg}}(\mathbf{t})}{SS_{\text{wg}}(\mathbf{t})}\right] = \frac{T-1}{2}\log\left[1 + \frac{SS_{\text{bg}}(\mathbf{t})}{SS_{\text{wg}}(\mathbf{t})}\right] + O_p(1). \tag{8}$$

With $\ell(\mathbf{t})$ defined as in (28*),

$$l(\hat{\boldsymbol{\tau}}) = C(\hat{\boldsymbol{\tau}})2^{-1}SS_{\text{bg}}(\hat{\boldsymbol{\tau}})/[SS_{\text{wg}}(\hat{\boldsymbol{\tau}})/(T-m+1)].$$

The rest of the derivation follows the variance known case, with the process $U(\mathbf{t})$ replaced by the process $\hat{\sigma}^{-1}U(\mathbf{t})$.

Bibliography

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

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

Zhang, N. & 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. & Li, J. Z. (2010). Detecting simultaneous change- points in multiple sequences. *Biometrika* **97**, 631–645.

Department of Statistics, Stanford University

E-mail: (nzhang@stanford.edu)

Department of Statistics, Stanford University

E-mail: (dos@stat.stanford.edu)