

# A Modified Bayes Information Criterion with Applications to the Analysis of Comparative Genomic Hybridization Data

Nancy R. Zhang\* and David O. Siegmund\*\*

Department of Statistics, Stanford University, Stanford, California 94305, U.S.A.

\* *email:* nzhang@stanford.edu

\*\* *email:* dos@stat.stanford.edu

**SUMMARY.** In the analysis of data generated by change-point processes, one critical challenge is to determine the number of change-points. The classic Bayes information criterion (BIC) statistic does not work well here because of irregularities in the likelihood function. By asymptotic approximation of the Bayes factor, we derive a modified BIC for the model of Brownian motion with changing drift. The modified BIC is similar to the classic BIC in the sense that the first term consists of the log likelihood, but it differs in the terms that penalize for model dimension. As an example of application, this new statistic is used to analyze array-based comparative genomic hybridization (array-CGH) data. Array-CGH measures the number of chromosome copies at each genome location of a cell sample, and is useful for finding the regions of genome deletion and amplification in tumor cells. The modified BIC performs well compared to existing methods in accurately choosing the number of regions of changed copy number. Unlike existing methods, it does not rely on tuning parameters or intensive computing. Thus it is impartial and easier to understand and to use.

**KEY WORDS:** Bayes information criterion; Change-point; Comparative genomic hybridization; Model selection.

## 1. Introduction

The Bayes information criterion (BIC) proposed by Schwarz (1978) is a popular method for model selection. However, its usage is not theoretically justified for change-point problems, where the likelihood functions do not satisfy the required regularity conditions. An aspect of model selection in change-point problems is to determine the number of change-points. In this article, we will give a modified version of the BIC for data consisting of independent normally distributed observations with constant variance and piecewise constant means.

There does not exist a standard method for model selection in change-point problems. Although lacking in theoretical justification, the BIC has previously been applied to models of the form treated in this article (e.g., see Yao, 1988; Li, 2001). The BIC was derived from large-sample estimates of posterior model probabilities that do not depend on the prior probability distributions of models nor of parameters. More recent model selection methods have taken a computationally based Bayesian approach (e.g., George and McCulloch, 1993; Green, 1995). They use Monte Carlo methods to estimate posterior model probabilities, and hence involve substantial computation and subjective choice of prior.

Recently, Siegmund (2004) suggested a new approach toward obtaining a large-sample approximation to the Bayes factor that bypasses the Taylor expansion used in the derivation of the BIC. Using this approach, he derived a new BIC-like statistic for use in mapping quantitative trait loci in genetic linkage studies, a problem which is related to finding change-points in the mean of an Ornstein–Uhlenbeck process.

The approach that we take here is similar to that suggested by Siegmund, but we will treat problems that have a more natural theoretical formulation and reduce in continuous time to finding change-points in the drift of Brownian motion.

Both the original and the modified BIC can be interpreted as penalized likelihood methods, although it should be emphasized that the “penalty” arises as a natural consequence of the computation and is not a regularization device to avoid overfitting. A brief comparison of the modified BIC with other penalized likelihood methods is given in Section 2.

As an example, we will apply the new version of the BIC to the analysis of array-based comparative genomic hybridization (array-CGH) data, which quantitatively measures the DNA copy number at thousands of locations linearly ordered along the genome. The goal in array-CGH data analysis is to detect accurately regions of DNA deletion or amplification. More details of array-CGH studies and comparison with the methods of Olshen et al. (2004) and Fridlyand et al. (2004) are given in Sections 4 and 5.

## 2. A Modified BIC for Gaussian Change-Point Models

Consider a sequence of observations  $\mathbf{y} = (y_1, y_2, \dots, y_T)$ , where  $y_i$  are independently distributed Gaussian random variables:

$$y_i \sim N(\mu_j, \sigma^2), \quad \text{for } i = \tau_j + 1, \dots, \tau_{j+1}, \quad j = 0, \dots, m. \quad (1)$$

Here we will refer to the  $\tau$ 's as the change-points of this process. Of course the change-points are constrained to lie in the set

$$\mathcal{D}_m = \{(t_1, \dots, t_m) : 0 \leq t_1 \leq t_2 \leq \dots \leq t_m \leq T\}.$$

For obvious reasons we add the restriction that  $\mu_j \neq \mu_{j+1}$ . We will assume that for large sample sizes, the change-points are far enough from each other in that

$$\lim_{T \rightarrow \infty} \tau_i/T \rightarrow r_i \quad \text{for } 1 \leq i \leq m, \quad (2)$$

where  $0 < r_1 < \dots < r_m < 1$ . To make exposition easier, we will also define  $\tau_0 = 0$  and  $\tau_{m+1} = T$ .

If the number of change-points,  $m$ , is known, then procedures based on maximum likelihood for finding and testing for the change-points have been derived (the simple one-dimensional case is given in James, James, and Siegmund, 1987). Our goal will be to determine  $m$ . From a model selection perspective, for  $0 \leq m < \infty$  we define  $\mathcal{M}_m$  to be the Gaussian model described above with  $m$  change-points. Thus,  $\mathcal{M}_0$  is the simplest model with no change and the dimension of  $\mathcal{M}_m$  increases linearly with  $m$ . The fit of model  $\mathcal{M}_m$  to the observed data will of course increase monotonically with increase in  $m$ . In our model selection procedure, we will be guided by the principle of finding the most parsimonious model that gives a “good” explanation of the data.

As in Schwarz (1978) and Siegmund (2004), we will approach this problem by deriving an asymptotic approximation of the Bayes factor. It is important to note here that although the following theorems are proved for specific prior distributions, the results hold for a very large class of priors. The prior on  $\boldsymbol{\tau}$  must be of the form  $f(\boldsymbol{\tau}) = g(\boldsymbol{\tau})T^{-m}$ , with  $C_1 < \max_{\boldsymbol{\tau}} g(\boldsymbol{\tau}) < C_2$ , where  $C_1, C_2$  are constants that do not depend on  $m$ . The density of the prior on the means and variance (when the variance is unknown) must have a support that covers an open set containing the maximum likelihood estimator. Then, as the sample size  $T$  increases, the likelihood term dominates in the Bayes factor, and the prior recedes into a negligible remainder term.

For simplicity, we first assume that the variance  $\sigma^2$  is known, and without loss of generality, let it be 1. After the reparameterization  $\mu_j = \mu_0 + \sum_{l=1}^j \delta_l$ , the parameters of the model  $\mathcal{M}_m$  are  $\boldsymbol{\theta} \triangleq (\mu_0, \tau_j, \delta_j : j = 1, \dots, m)$  and belong to the subparameter space

$$\Theta_m \triangleq \{(\mu_0, \tau_1, \dots, \tau_m, \delta_1, \dots, \delta_m) : \mu_0 \in \mathfrak{R}, \boldsymbol{\tau} \in \mathcal{D}_m, \boldsymbol{\delta} \in \mathfrak{R}^m\}. \quad (3)$$

The next theorem gives the large-sample approximation of the Bayes factor for this case.

**THEOREM 1:** *Let  $\mathcal{M}_m$  be the model defined in equation (1). Let  $\sigma = 1$  and assume that  $(m, \boldsymbol{\theta})$  follow a uniform prior over  $\mathcal{Z}^+ \times \Theta_m$ , then*

$$\begin{aligned} \log \frac{P(\mathbf{y} | \mathcal{M}_m)}{P(\mathbf{y} | \mathcal{M}_0)} &= \frac{1}{2} \sum_{i=1}^{m+1} n_i(\hat{\boldsymbol{t}}) [\bar{y}_i(\hat{\boldsymbol{t}}) - \bar{y}]^2 \\ &\quad - \frac{1}{2} \sum_{i=1}^{m+1} \log n_i(\hat{\boldsymbol{t}}) + \left(\frac{1}{2} - m\right) \log(T) + O_p(1), \end{aligned} \quad (4)$$

where

$$\begin{aligned} n_i(\hat{\boldsymbol{t}}) &= \hat{t}_i - \hat{t}_{i-1}, \\ \bar{y}_i(\hat{\boldsymbol{t}}) &= \frac{1}{n_i(\hat{\boldsymbol{t}})} \sum_{j=\hat{t}_{i-1}}^{\hat{t}_i-1} y_j, \quad \bar{y} = \frac{1}{T} \sum_{j=1}^T y_j \\ \hat{\boldsymbol{t}} &= (\hat{t}_1, \dots, \hat{t}_m) \\ &= \arg \max_{0 < \hat{t}_1 < \dots < \hat{t}_m < T} \sum_{i=1}^{m+1} n_i(\hat{\boldsymbol{t}}) [\bar{y}_i(\hat{\boldsymbol{t}}) - \bar{y}]^2. \end{aligned}$$

A brief sketch of the proof of this theorem will be given in the supporting Web Appendix. Asymptotically, selecting the model with the largest posterior probability is equivalent to selecting the model with the largest Bayes factor. We will call the right-hand side of equation (4), in which we neglect the  $O_p(1)$  remainder term, the “modified BIC.” We propose as a model selection procedure to select the model that maximizes the modified BIC.

It is easy to see that the first term of the modified BIC is the maximized log likelihood under  $\mathcal{M}_m$ . One can interpret the (negative of the) remaining terms on the right-hand side of equation (4) as a “penalty” for the complexity of the model. The classic BIC procedure can be similarly interpreted. In comparison to the classical BIC and penalized likelihood methods (e.g., Tibshirani, 1996; Birgé and Massart, 2001; Gu and Wang, 2003; Lavielle, 2005), the modified BIC differs in the penalty term. All penalized likelihood methods choose the model that maximizes a criterion of the form

$$l_m(\hat{\boldsymbol{\theta}}_m) - p_m(\hat{\boldsymbol{\theta}}_m),$$

where  $l_m$  is the log-likelihood function of  $\mathcal{M}_m$ ,  $\hat{\boldsymbol{\theta}}_m$  is the maximum likelihood parameter estimate, and  $p_m$  is the penalty function. In the classic BIC,  $p_m = d_m \log T/2$ , where  $d_m$  is the dimension of the model  $\mathcal{M}_m$ . In some other methods (e.g., Tibshirani, 1996; Birgé and Massart, 2001; Gu and Wang, 2003),  $p_m$  depends on a shrinkage parameter  $\beta$  that must be chosen by the user, usually through cross-validation. Lavielle (2005) proposed  $p_m = \beta m \log T$  and gave a data-adaptive way of choosing  $\beta$ .

Like the classic BIC, the modified BIC gives a precise formula for  $p_m$ , without the need for a user-chosen shrinkage parameter. However, while in the BIC each parameter contributes one dimension to the model, for the modified BIC the penalty term no longer follows a simple formula. If we reparameterize  $(\hat{t}_0 = 0, \hat{t}_1, \dots, \hat{t}_{m+1} = T) = (Tr_0, Tr_1, \dots, Tr_{m+1})$ , where  $0 = r_0 < r_1 < \dots < r_{m+1} = 1$ , then the penalty terms from equation (4) can be written as

$$-\frac{1}{2} \left[ 3m \log(T) + \sum_{i=1}^{m+1} \log(r_i - r_{i-1}) \right]. \quad (5)$$

The second term in equation (5) is maximized when the change-points are evenly spaced on  $(0, T)$ , and minimized when they are as close together as possible (e.g.,  $\hat{t}_1 = 1, \hat{t}_2 = 2, \dots, \hat{t}_m = m$ ):

$$\max_{\mathbf{r}} \sum_{i=1}^{m+1} \log(r_i - r_{i-1}) = -(m+1) \log(m+1), \quad (6)$$

$$\min_r \sum_{i=1}^{m+1} \log(r_i - r_{i-1}) = \log[(1/T)^m (T - m)/T] \approx -m \log(T). \quad (7)$$

Because of assumption (2), equation (7) is not possible under our model, and thus a rough approximation for the penalty term of the modified BIC is  $p_m \approx 3m \log(T)/2$ . However, because of the slow-growing nature of  $\log(T)$ , even for large data set sizes, the second term in equation (5) can still make a significant contribution to the penalty. In equation (5), we will attribute one dimension to each “jump” parameter  $\delta$ . Then, we contribute between one and two dimensions to each change-point location parameter, with the exact contribution depending on their relative locations to each other. This is consistent with practical experience on the dimension of change-point models (e.g., Birgé and Massart, 2001; Lavielle, 2005).

In applications we will need a version of the modified BIC for the case of unknown variance, which is given in the next theorem.

**THEOREM 2:** *Under uniform priors for  $\{\mu, \tau_i, \delta_i : i = 1, \dots, m\}$ , and the noninformative prior  $1/\sigma$  for  $\sigma$ ,*

$$\begin{aligned} \log \frac{P(\mathcal{M}_m | \mathbf{y})}{P(\mathcal{M}_0 | \mathbf{y})} &= \left( \frac{T - m + 1}{2} \right) \log \left[ 1 + \frac{SS_{\text{bg}}(\hat{\mathbf{t}})}{SS_{\text{wg}}(\hat{\mathbf{t}})} \right] \\ &+ \log \left[ \frac{\Gamma \left( \frac{T - m + 1}{2} \right)}{\Gamma \left( \frac{T + 1}{2} \right)} \right] + \frac{m}{2} \log(SS_{\text{all}}) \\ &- \frac{1}{2} \sum_{i=1}^{m+1} \log n_i(\hat{\mathbf{t}}) + \left( \frac{1}{2} - m \right) \log(T) + O_p(1), \end{aligned} \quad (8)$$

where

$$\begin{aligned} SS_{\text{bg}} &= \sum_{i=1}^{m+1} n_i(\hat{\mathbf{t}}) [\bar{y}_i(\hat{\mathbf{t}}) - \bar{y}]^2, \quad SS_{\text{all}} = \sum_{j=1}^T (y_j - \bar{y})^2, \\ SS_{\text{wg}} &= SS_{\text{all}} - SS_{\text{bg}}, \\ \hat{\mathbf{t}} &= \arg \max_{\mathbf{t} \in \mathcal{D}_m} \frac{SS_{\text{bg}}(\mathbf{t})}{SS_{\text{wg}}(\mathbf{t})}, \\ \Gamma(t) &= \int_0^\infty x^{t-1} e^{-x} dx, \end{aligned} \quad (9)$$

and  $n_i(\hat{\mathbf{t}})$ ,  $\bar{y}_i(\hat{\mathbf{t}})$ , and  $\bar{y}$  are as defined in Theorem 1.

The form of the approximation in the second theorem is similar to that in Theorem 1. The first term is still a maximized log-likelihood ratio, where with the variance unknown the density is Student's  $t$  instead of Gaussian. The penalty terms,  $-\frac{1}{2} \sum_{i=0}^m \log(\hat{t}_{i+1} - \hat{t}_i) + (\frac{1}{2} - m) \log(T)$ , remain the same. The other terms in the second line of the approximation, all of order  $m \log(T)$ , come from integrating out the nuisance parameter  $\sigma$ . For the proof of this theorem, see Zhang (2005).

### 3. Details of Implementation

Here we describe in detail how the modified BIC can be applied to the data  $\{y_1, \dots, y_T\}$ . Assume the variance is un-

known. If for each fixed  $m$  up to some prechosen maximal value  $M$ , maximum likelihood estimates  $\hat{\mathbf{t}}^{(m)} = \{\hat{t}_1, \dots, \hat{t}_m\}$  of the putative change-point locations have been calculated, then the formula from Theorem 2 can be directly applied to evaluate modified BIC( $m$ ) for  $m = 1, \dots, M$ . Then we choose the  $m^*$  that maximizes modified BIC( $m$ ). The segmentation that our method returns is  $\hat{\mathbf{t}}^{(m^*)}$ . The maximum value  $M$  is an algorithmic convenience to avoid unnecessary computing. The modified BIC is consistent as long as  $M$  is greater than the true number of change-points. In our experience BIC( $m$ ) increases to a maximum and then decreases. We try to set  $M$ , perhaps iteratively, to allow us to see this pattern clearly.

The remaining question is how to calculate  $\hat{\mathbf{t}}^{(m)}$  for a given  $m$ . While one would like in principle to consider all possible partitions of  $\{1, \dots, T\}$  by points  $t_1 < t_2 < \dots < t_m$ , this would be computationally infeasible for all but very small values of  $m$ . Because the recursive method of Olshen et al. (2004) is very fast and from their examples appears to detect true change-points (and perhaps also false change-points) very effectively, we have used the following modification of their circular binary segmentation (CBS) algorithm (and have had a similar experience). Some definitions and technical details are given in the supporting Web Appendix.

1. Let  $\hat{\mathbf{t}} = \{0, T\}$ .
2. While  $|\hat{\mathbf{t}}| < m + 2$ 
  - (a) For each  $i \in \{0, \dots, |\hat{\mathbf{t}}|\}$ , do the following:
    - If  $|\hat{\mathbf{t}}| < m + 1$ : Let  $X'_i$  be the maximum likelihood statistic for a square wave change (equation (3) in the Web Appendix) in  $\{y_{t_{i+1}}, \dots, y_{t_{i+1}}\}$ , and let  $s_i, s'_i$  be the maximum likelihood change-point locations. Then,
      - if  $s_i$  is not significant as a change-point in  $\{y_{t_{i+1}}, \dots, y_{s'_i}\}$ , let  $r_i = s'_i$  and  $X_i$  be the likelihood ratio of a single change at  $r_i$ ;
      - else if  $s'_i$  is not significant as a change-point in  $\{y_{s_i}, \dots, y_{t_{i+1}}\}$ , let  $r_i = s_i$  and  $X_i$  be the likelihood ratio of a single change at  $r_i$ ;
      - otherwise let  $r_i = \{s_i, s'_i\}$  and  $X_i = X'_i$ .
    - Else, when  $|\hat{\mathbf{t}}| = m + 1$  let  $X_i$  be the log-likelihood ratio statistic for a single change (equation (2) in the Web Appendix) in  $\{y_{t_{i+1}}, \dots, y_{t_{i+1}}\}$ , and let  $r_i$  be the maximum likelihood change-point location.
  - (b) Pick the  $i$  with the largest  $X_i$ , and add  $r_i$  to the ordered list  $\hat{\mathbf{t}}$ .
3. Discard 0,  $T$  from  $\hat{\mathbf{t}}$ .

In Step 2(a) above, the significance of the putative change-points  $s_i$  and  $s'_i$  can be evaluated by a fixed p-value threshold, which we set to 0.05. The performance of the algorithm is robust to this threshold. Please refer to the Web Appendix for descriptions of these likelihood ratio statistics and their significance calculations. The above algorithm differs from the original CBS algorithm in that the recursion is stopped when a predetermined number  $m$  of change-points have been found, not when no more “significant” change-points can be detected. Also, as in Olshen et al. (2004), we correct for edge effects in Step 2(a) in cases where either  $s_i$  or  $s'_i$  is near the edge

of the sequence. Although this algorithm is not guaranteed to be optimal, in our experience it is accurate and very fast.

#### 4. Simulation Study

We first use simulated data to test the reasonableness of the modified BIC as a model selection criterion. To assess performance we will use as benchmarks the classic BIC and the methods from Olshen et al. (2004) and Fridlyand et al. (2004). One reason for comparison against the latter two methods is that they have both been designed with the analysis of array-CGH data in mind, a problem that we will examine in Section 5. We begin by describing briefly how these two methods work.

Olshen et al. (2004) developed a modification of the binary segmentation algorithm called CBS for estimating the change-point locations. CBS is almost exactly the same as the algorithm given in Section 3, except that it uses a different stopping criterion for the recursion: the search is stopped if none of the segments contain a significant change, based on some prechosen p-value threshold  $\alpha$ . However, they found that this stopping criterion has a tendency to overestimate the number of change-points, and thus added a pruning step at the end of the CBS algorithm. The pruning step depends on a threshold parameter  $\gamma$ . Since varying  $\gamma$  can limit the number of change-points reported to any value, and for a single study there is no clear impartial way of choosing  $\gamma$ , we will consider only the basic CBS algorithm with the threshold stopping rule for choosing  $m$ .

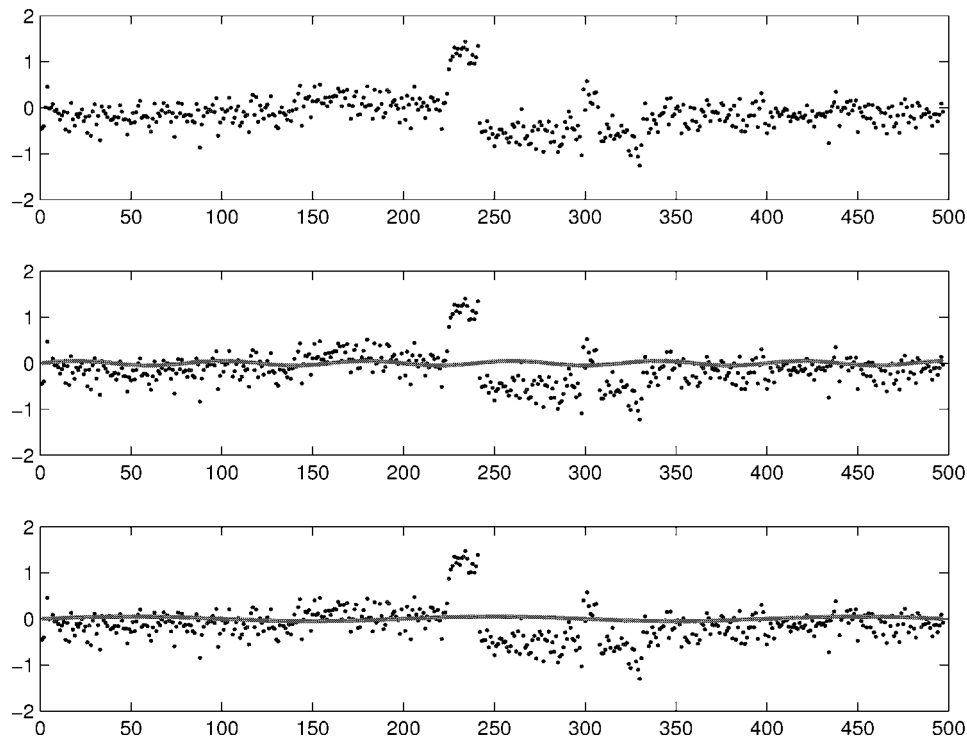
Fridlyand et al. (2004) use hidden Markov models (HMM) for segmentation of array-CGH data. In this approach, the un-

known mean values at each time-point are the hidden states, with the change-points being transitions between different states. Finding the change-points requires first training the HMM on the data and then calculating the most probable state sequence. One crucial parameter in this method is the number of states in the HMM, which is selected using the Akaike information criterion (AIC). For array-CGH data analysis, Fridlyand et al. found it advisable to add a merging step on top of the AIC to reduce the number of states further. Like the pruning step in Olshen et al. (2004), this state merging step can be tuned to make the method arbitrarily conservative, and there are no clear criteria for choosing the pruning parameters. Thus, we will use the HMM method with the AIC for model selection.

The simulation data set that we used is generated from a model created by Olshen et al. (2004). The model is as follows:

$$y_i \sim \mu_i + 0.25\sigma \sin(a\pi i) + \epsilon_i, \quad i = 1, \dots, T, \quad (10)$$

where  $\epsilon_i \sim N(0, 0.04)$ ,  $T = 497$ , and  $\mu_i$  changes six times with  $\tau = (138, 225, 242, 299, 308, 332)$ ,  $\mu_0 = -0.18$ , and  $\delta = (0.26, 0.99, 1.6, 0.69, 0.85, 0.53)$ . The second term of equation (10) is a sinusoid trend component intended to make this data set more challenging, and was originally added by Olshen et al. (2004) to mimic the periodic trends that are found to exist in array-CGH data. As in their paper, we will experiment with the trend parameters  $a \in \{0, 0.01, 0.025\}$  corresponding, respectively, to no trend and local trend with long and short periods. An example of a data series generated using this model with each of the trend parameters used is shown in Figure 1.



**Figure 1.** Example of one Monte Carlo sample of simulation model (10). The trend parameters are  $a = 0, 0.025$ , and  $0.01$ , respectively, for the top, middle, and bottom figures. The line depicts the trend added to the mean.

**Table 1**

Test results on simulation data for different values of the trend parameter. The columns list the number of times out of 100 simulations each method found 5, 6, 7, or >8 change-points. The correct number is 6.

Trend	Method	5	6	7	8+
None	Traditional BIC	0	77	14	9
None	Olshen et al. (2004)	0	91	6	3
None	Fridlyand et al. (2004)	0	83	13	4
None	Modified BIC (Theorem 1)	0	95	5	0
None	Modified BIC (Theorem 2)	0	97	3	0
Short	Traditional BIC	0	58	30	12
Short	Olshen et al. (2004)	0	66	23	11
Short	Fridlyand et al. (2004)	0	70	27	3
Short	Modified BIC (Theorem 1)	0	88	11	1
Short	Modified BIC (Theorem 2)	0	98	2	0
Long	Traditional BIC	0	38	35	27
Long	Olshen et al. (2004)	0	68	11	21
Long	Fridlyand et al. (2004)	1	80	13	6
Long	Modified BIC (Theorem 1)	0	78	19	3
Long	Modified BIC (Theorem 2)	2	94	4	0

Table 1 shows the test results for each of the different methods on simulation data generated from model (10). The table lists the percentage of time in 100 Monte Carlo replicates that each method finds  $m$  change-points, with  $m = 5, 6, 7, 8$ . It is clear from these results that the modified BIC is more specific than the traditional BIC, while not sacrificing sensitivity. This is expected, since the classic BIC has the wrong penalty terms that do not penalize enough for the change-point parameters. The results also show that the unknown variance version of the modified BIC is more conservative than the known variance version. This is due to the dampening effect of the log function in the first term of equation (8), thus giving the penalty terms a relatively larger weight than the variance-known version of the BIC. Comparing the first terms of equations (4) and (8), note that by the law of large numbers,  $SS_{bg}/\sigma^2 \rightarrow Tx$  for some  $x$  that depends on  $\delta$  and  $\mathbf{r}$ . Therefore,  $(T - m - 1) \log(1 + SS_{bg}/SS_{wg}) \rightarrow T \log(1 + x)$ , and thus the first term of equation (8) is dampened with respect to the first term of equation (4) by the amount of  $\log(1 + x)$ . This effect is more pronounced for more statistically significant changes.

These simulation results show that the modified BIC performs better than the basic CBS algorithm without the additional pruning step, with the improvement being more pronounced when the sinusoidal trend is added. On this data set, the modified BIC also seems to be preferable to the basic HMM method by Fridlyand et al. (2004). However, our simulation study shows that the results of the HMM method depend on initialization of model parameters (see supporting Web Appendix). We chose not to compare to the pruned CBS or state-merged HMM method, because there is no systematic or obvious way to choose the pruning (or state merging) thresholds.

## 5. Analysis of Array-CGH Data

For a homogeneous cell line, the DNA copy number at a given genome location is the number of copies of genomic DNA that each cell has for that location. In human and other

diploid organisms, the DNA copy number is normally two for autosomes. However, the genomes of tumor cells have been observed to exhibit regions with amplified or reduced copy numbers. The accurate identification of these regions is important for studying tumor progression and may also be useful in tumor classification (Albertson et al., 2003).

The recently developed array-CGH technology provides the means to quantitatively measure DNA copy number at thousands of locations on the genome simultaneously (Pinkel et al., 1998; Pollack et al., 1999). The technology behind array-CGH experiments is similar to that behind cDNA gene expression experiments. Typically, the test genomic DNA pool (e.g., genomic DNA from tumor cell samples) and a diploid reference genomic DNA pool are differentially labeled with dyes (e.g., Cy3 and Cy5). These two dye-labeled samples are mixed and hybridized to a microarray chip which is spotted with thousands of genomic targets, each mapping to a known genome location. The hybridized microarray chip is then scanned, and image analysis software is used to calculate the ratio of the test and reference fluorescence intensities for each genomic target. The ratio of the intensities of the dyes is a surrogate for the ratio of the abundance of the DNA sample labeled with the dyes. For our current data analysis purposes, we can view array-CGH experiments as producing, for each cell sample, an ordered sequence of  $(t, y_t)$  pairs, where  $t$  represents the location in the genome and  $y_t$  represents the  $\log_2$  ratio of the test versus reference spot intensities for the genomic target from that location.

The first step in the analysis of array-CGH data is to determine accurately the regions of changed copy number. There are two major aspects to this problem: the accurate estimation of the locations of the aberrations, and the determination of the number of aberrations. Based on their performance on the Coriel data set of Snijders et al. (2001), existing methods (e.g., Fridlyand et al., 2004; Olshen et al., 2004; Hsu et al., 2005; Wang et al., 2005) seem to more or less agree on the locations of the aberrations, once the number of aberrations is carefully chosen. However, the estimation of the number

of segments involves more subtle aspects of modeling, and is a major source of disagreement between current methods. The study by Picard et al. (2005) compares several different model selection criteria for array-CGH data, and shows the importance of this step to the overall segmentation result. We propose the modified BIC as a model selection procedure for array-CGH data analysis. Although the algorithm given in Section 3 pairs the modified BIC with the CBS algorithm, it can potentially be used as a model selection procedure to complement most segmentation algorithms.

Because of the linear ordering of the probes by genome location and the discontinuous nature of chromosome aberrations, the copy number estimation problem for array-CGH data fits naturally into the change-point detection framework. We will assume the following high-level model for the data

$$y_t = f(c_t) + \epsilon_t, \quad (11)$$

where  $f$  is a monotone increasing function of the discrete copy number  $c_t$  at location  $t$ , and  $\epsilon_t$  are independent identically distributed Gaussian noise with common variance  $\sigma^2$ . The reason for introducing  $f$  is that the mean of the log fluorescence intensity ratio increases with, but does not equal, the log of the copy number ratio, possibly due to contamination or mosaicism (Pollack et al., 1999). This model is similar to the one used by Olshen et al. (2004). The motivation for assuming a common variance can be found in Picard et al. (2005), although our method can also accommodate the slightly more general case that the variances differ but the ratio of any two variances is known.

Like the previous studies, we first use the Coriel data set featured in Snijders et al. (2001) to test the modified BIC. This data set consists of single experiments on 15 fibroblast cell lines (of which we will use the nine that were used by Olshen et al., 2004). Because this data set consists of pure diploid cell lines with chromosome copy number changes that have been previously characterized by spectral karyotyping and are easily detectable by eye, it is used mainly as a proof of principle. The results of our method, in terms of the number of false positives and false negatives, are listed in column 3 of Table 2 and compared to the performance on the same data

set by the methods of Olshen et al. (2004) and Fridlyand et al. (2004). Please note that the pruning step in Olshen et al. (2004) is not applied. The fitted means of the models selected by each of the listed methods are plotted against the data for cell lines GM03563 and GM01750 in Figure 2. From these results, we can see that the modified BIC performs reasonably on this simple data set, being more specific than the method in Olshen et al. (2004) while not compromising sensitivity. The HMM algorithm of Fridlyand et al. (2004) also performs well, but it involves user selection of training parameters and thresholds, and so might not be a fair comparison. Careful inspection of Figure 2 suggests that many of the false positives in Olshen et al. (2004) and our results are caused by curved local trends present in the data. Olshen et al. suggest that these local trends might have a biological origin.

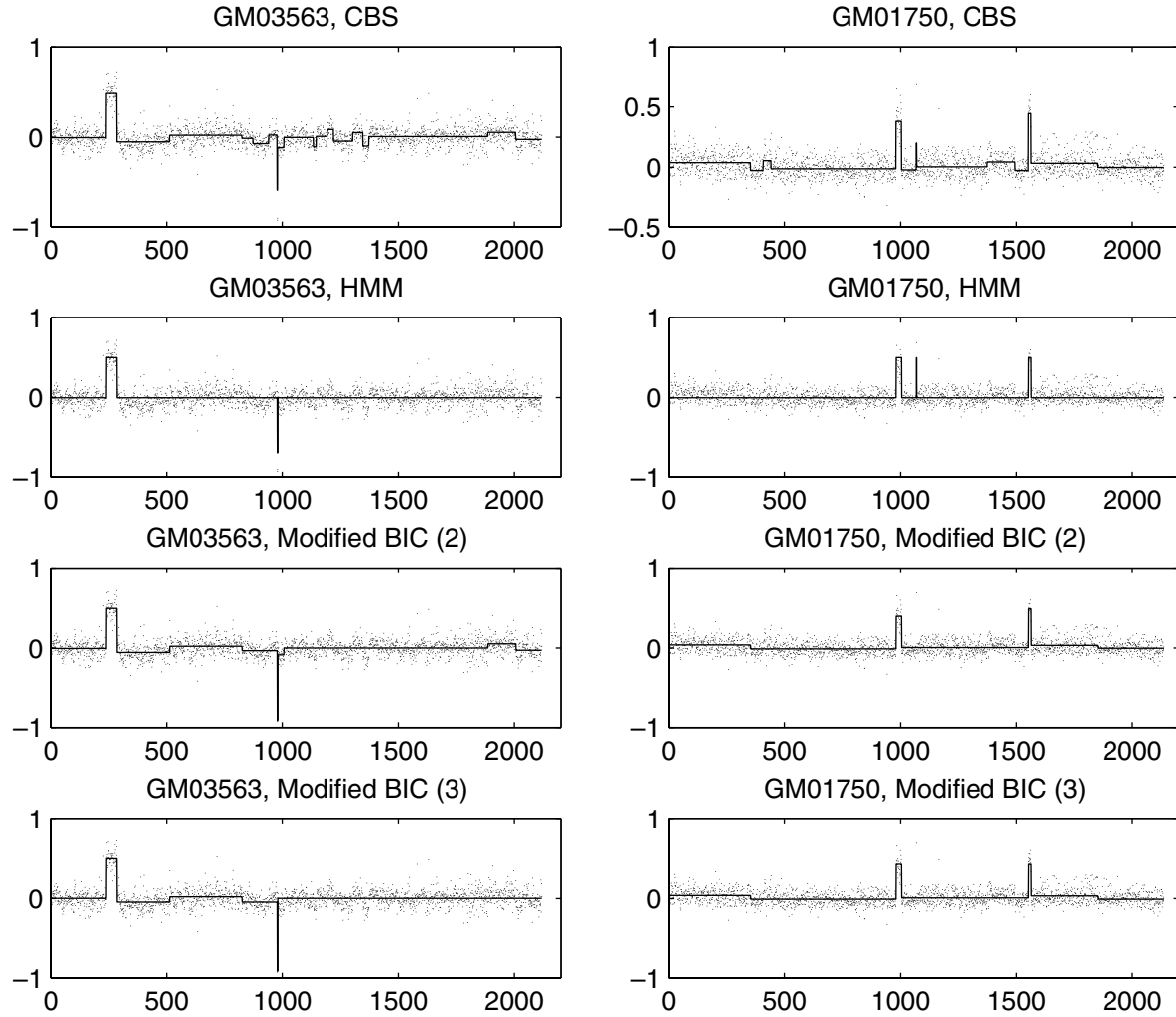
To further test the modified BIC, we use the more challenging data set from Snijders et al. (2003). The experiment behind this data set examines how defects in mismatch repair (MMR) genes affect the chromosome aberration profile of tumors. For our testing, we will use the BT474 and MCF7 cell lines, both of which have complex aberration profiles. The true chromosome copy number is not known for this experiment, and thus there is no direct way to verify the results. Our goal in testing the modified BIC on this data set is twofold: first, to verify by eye that the modified BIC performs reasonably on a more complicated data set, and second, to compare the behavior of the modified BIC with that of the HMM method in a more challenging scenario.

This more complex data set shows our procedure to be a more conservative change-point detection procedure than HMM. For the BT474 cell line, the modified BIC peaks at 39 change-points while with HMM, 103 change-points are found if the state merging step is not taken. After the merging of states, the complete procedure from Fridlyand et al. (2004) reports 69 change-points for this data series. For the MCF7 cell line, HMM without merging found 102 change-points, which are reduced to 95 after the merging step, while the modified BIC peaks at 55. These two data series and their mean values estimated by each method are shown in Figures 3 and 4. (The figures plot the results from the next section, which are very

**Table 2**

*Test results on nine cell lines from Snijders et al. (2001) for the CBS algorithm of Olshen et al. (2004) (threshold  $p$ -value = 0.01, no pruning), the HMM algorithm of Fridlyand et al. (2004; state merging threshold 0.35), the modified BIC of Theorem 2, and the modified BIC for a limited-levels model of Theorem 3*

Cell line	Olshen et al.		Fridlyand et al.		Mod BIC (Theorem 2)		Mod BIC (Theorem 3)	
	FP	FN	FP	FN	FP	FN	FP	FN
GM03563	14	0	0	0	5	0	2	0
GM05296	11	0	4	0	4	0	4	0
GM01750	8	0	2	0	2	0	2	0
GM03134	12	0	4	0	6	0	6	0
GM13330	17	0	0	0	5	0	2	0
GM01535	6	0	2	0	2	0	2	0
GM07081	6	1	2	1	2	1	2	1
GM13031	7	0	2	0	7	0	1	0
GM01524	9	0	0	0	2	0	0	0



**Figure 2.** Comparison of fitted means for GM03563 and GM01750 cell lines in the Coriel data set. Row 1: CBS algorithm without pruning. Row 2: HMM algorithm with state merging threshold set to 0.35. Row 3: modified BIC of Theorem 2. Row 4: modified BIC of Theorem 3 selecting for both number of change-points and number of levels.

similar to the results of this section in terms of the number of change-points found.) Since the true copy numbers are not known, we cannot judge from these results which method is more “accurate.” However, by careful inspection of the figures, it is clear that the segmentation produced by the three methods is very different. This shows that, while it is easy for segmentation methods to agree on a simple data set such as the Coriel cell line, on a more complex data set the choice of statistical method is crucial.

## 6. Restricting the Magnitude of Change

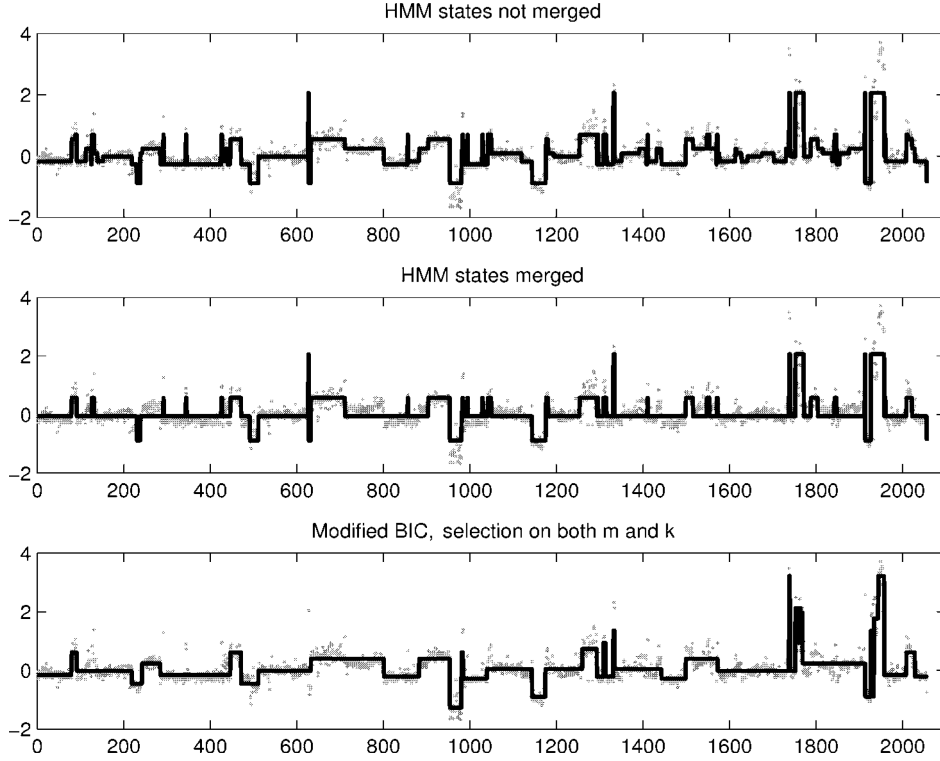
In array-CGH data, the underlying chromosome copy numbers change only by whole integers. Hence, for many data sets, the underlying segment means of  $\{y_t\}$  should take on a limited number of values. Different segments of the genome that have the same underlying copy number composition should also receive the same estimated mean, instead of a slightly different estimated mean for each segment. The modified BICs in Theorems 1 and 2 do not take these limitations into account. By model (1), at each change-point the

process mean can jump to any value. Thus, if there are  $m$  change-points, there can be, and surely will be by maximum likelihood,  $m + 1$  different levels for the estimated means.

To create a more specific model for array-CGH data, we would like to restrict the jumps in mean. Here, we propose the limited-levels model  $\mathcal{M}_{m,k}$ , in which at each of the  $m$  change-points the mean must jump to one of only  $k$  levels:

$$y_t \sim N(\mu + \delta_{a_j}, \sigma^2) \text{ for } t = \tau_j + 1, \tau_j + 2, \dots, \tau_{j+1}, \quad j = 0, \dots, m, \quad (12)$$

where  $a_j \in \{0, \dots, k\}$  is the level assignment of segment  $i$  between change-points  $t_{i-1}$  and  $t_i$ . We enforce that  $a_0 = 0$  and  $\delta_0 = 0$ , so that before the first change-point the mean is at the baseline level  $\mu$ . Of course, the total number of levels,  $k + 1$ , must be no greater than the total number of sections,  $m + 1$ . When  $k = m$  we have our previous unrestricted-jump model. To maximize the likelihood in this new problem, we must not only search over the possible change-points but also



**Figure 3.** Comparison of fitted means for the BT474 cell line in MMR data set. Top panel: means estimated using the HMM algorithm proposed in Fridlyand et al. (2004), without the state merging step. Middle panel: same as top panel, but with the merging step with threshold parameter set to 0.35. Bottom panel: the statistic in Theorem 3 is used to select both the number of change-points and levels; change-points found using the CBS algorithm.

over the possible assignments of segments to levels  $a$ . Given  $\hat{\mathbf{t}}$  where  $m, k$ , the domain of possible values for  $a$  is

$$\mathcal{A}_{m,k} = \{(a_1, \dots, a_{m+1}) : a_i \in \{1, \dots, k\}; \\ a_i \neq a_{i+1}; \forall 1 \leq j \leq k \exists i, a_i = k\}. \quad (13)$$

Since the dimension of the model increases with both  $m$  and  $k$ , we now have a two-dimensional model selection problem. The following theorem gives a modified version of the BIC for simultaneous model selection on both  $m$  and  $k$  when the variance is unknown.

**THEOREM 3:** *Let  $\mathcal{M}_{m,k}$  be the model where there are  $m$  change-points and  $k$  distinct levels. Then,*

$$\log \frac{P(\mathcal{M}_{m,k} | \mathbf{y})}{P(\mathcal{M}_0 | \mathbf{y})} = \left( \frac{T-k-1}{2} \right) \log \left[ 1 + \frac{SS_{\text{bg}}(\hat{\mathbf{t}}, \hat{\mathbf{a}})}{SS_{\text{wg}}(\hat{\mathbf{t}}, \hat{\mathbf{a}})} \right] \\ + \log \left[ \frac{\Gamma \left( \frac{T-k-1}{2} \right)}{\Gamma \left( \frac{T-1}{2} \right)} \right] + \frac{k}{2} \log(SS_{\text{all}}) \\ - \frac{1}{2} \sum_{j=1}^k \log n_j(\hat{\mathbf{t}}, \hat{\mathbf{a}}) - (m-1/2) \log(T) \\ + O_p(1), \quad (14)$$

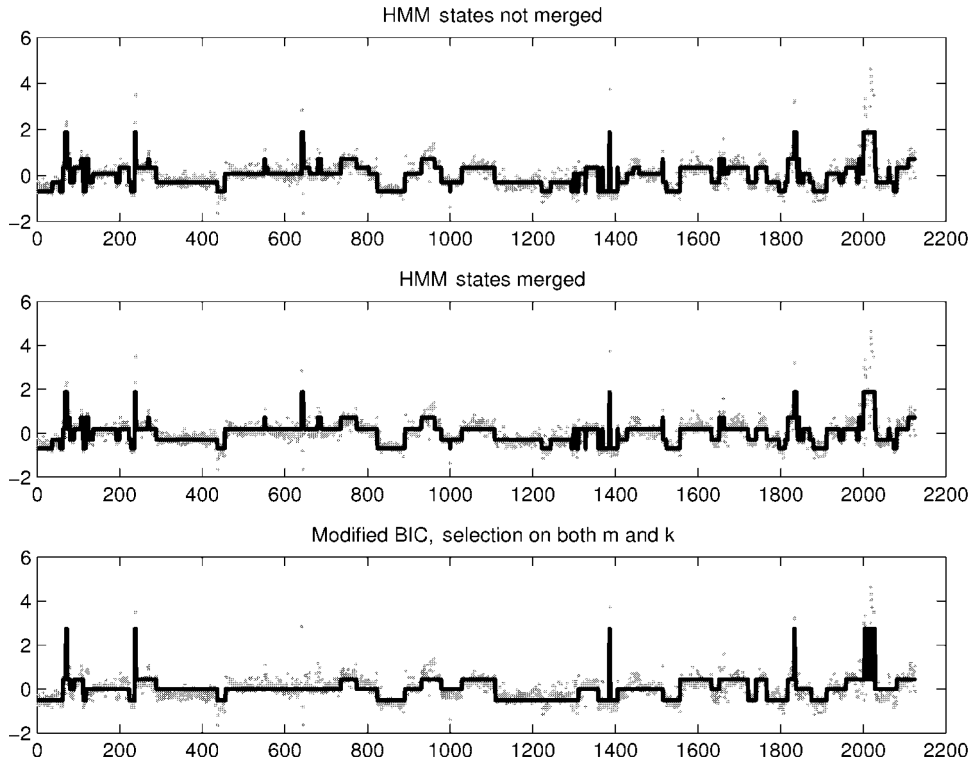
$$SS_{\text{bg}} = \sum_{i=1}^k n_i(\hat{\mathbf{t}}, \hat{\mathbf{a}}) [\bar{y}_i(\hat{\mathbf{t}}, \hat{\mathbf{a}}) - \bar{y}]^2,$$

$$n_i(\hat{\mathbf{t}}, \hat{\mathbf{a}}) = |\{j : y_j \in \text{level } i \text{ under change-points } \hat{\mathbf{t}} \\ \text{and assignment } \hat{\mathbf{a}}\}|,$$

$$\bar{y}_i(\hat{\mathbf{t}}, \hat{\mathbf{a}}) = \sum_{j: y_j \in \text{level } i \text{ under } \hat{\mathbf{t}}, \hat{\mathbf{a}}} y_j / n_i(\hat{\mathbf{t}}, \hat{\mathbf{a}}),$$

$$(\hat{\mathbf{t}}, \hat{\mathbf{a}}) = \arg \max_{\mathbf{t} \in \mathcal{D}_m, \mathbf{a} \in \mathcal{A}_{m,k}} \sum_{i=1}^k n_i(\hat{\mathbf{t}}, \hat{\mathbf{a}}) [\bar{y}_i(\hat{\mathbf{t}}, \hat{\mathbf{a}}) - \bar{y}]^2.$$

The proof of this theorem is very similar to the proof of Theorem 2 and will be omitted. The structure of equation (14), which we call the modified BIC for  $m$  change-points and  $k$  levels, is almost exactly the same as the structure of the modified BIC from Theorem 2. The key change is that while in Theorem 2 each segment between two adjacent change-points comprises its own group in the calculation of  $SS_{\text{bg}}$  and  $SS_{\text{wg}}$ , here all segments that are assigned to the same level by  $\hat{\mathbf{a}}$  belong to the same group. In some parts of the remaining terms of equation (14),  $m$  is replaced by  $k$  because  $\delta$  is now  $k$ -dimensional. In particular, the penalty term  $\frac{1}{2} \sum_{j=1}^k \log n_j(\hat{\mathbf{t}}, \hat{\mathbf{a}})$  is now of order  $k \log(T)$ . Thus, the penalty term grows with both the number of change-points and the number of levels.



**Figure 4.** Comparison of fitted means for the MC57 cell line in MMR data set. Top panel: means estimated using the HMM algorithm proposed in Fridlyand et al. (2004), without the state merging step. Middle panel: same as top panel, but with the merging step with threshold parameter set to 0.35. Bottom panel: the statistic in Theorem 3 is used to select both the number of change-points and levels; change-points found using the CBS algorithm.

The modified BIC for the limited-levels model is tested on the Coriel and MMR data sets. Like before, the CBS algorithm is used to find the change-points, then the segment means are clustered into  $k$  groups using  $k$ -means. However, with restrictions on the jump sizes CBS frequently leads to suboptimal choices for  $\hat{t}$ , but is used here for its speed and simplicity. The performance of this new model is listed alongside others in Table 2 and Figures 1–3. For the Coriel data set, the limited-levels model performs better than the unrestricted-jump model of Theorem 2, but does not completely resolve the problem of local trends. For the MMR data set, the limited-levels model does not drastically reduce the number of change-points detected, but substantially reduces the number of levels.

## 7. Discussion

Test results on both simulated and array-CGH data show that the modified BIC is a reasonable statistic for model selection in the types of problems discussed here. The simulated data that we used were taken from Olshen et al. (2004), and thus offer an unbiased assessment of our method. In the Coriel data set, it is seen that the modified BIC is more specific than the simple threshold method from Olshen et al. (2004), being more robust to the local trends that seem to be a problem in array-CGH data sets. In the more complex MMR data set, the modified BIC is clearly more conservative than the method of Fridlyand et al. and gives a smoother fit to the data. One of the advantages of the modified BIC over the other methods for deciding the number of aberrations in array-CGH data is

its impartiality, since like the classic BIC, it does not require a specific prior or tuning parameters. Thus, the modified BIC is a natural model selection criterion to be used in conjunction with a modification of the CBS algorithm of Olshen et al. (2004) for array-CGH data analysis.

A class of general model selection methods we have not considered is Monte Carlo methods to estimate the posterior probability of a model (e.g., Green, 1995). It would be quite reasonable to conduct a comparative study. When it was invented, the BIC had a computational advantage, because lack of computing power made the determination of p-values and posterior probabilities an arduous task. Now computation is less of an issue, but these Monte Carlo based methods involve choices of prior and tuning parameters, as well as questions of convergence of the Monte Carlo estimates. Hence they are less transparent, and without a systematic robustness study one might reasonably be concerned that differences between them and the modified BIC are a reflection of what may be arbitrary choices, not information in the data.

There are three critical modeling assumptions in our analysis, that the data are normally distributed, homoscedastic, and uncorrelated. As indicated above, the assumption of homoscedasticity can be slightly weakened, and this may have advantages for data where the variance changes systematically with some covariate. Serial correlation could be a serious problem, but correcting for it appears to be reasonably straightforward. This may become important for the next generation of array-CGH data. The natural setting of our analysis is the exponential family of distributions, so the assumption of

normality can be replaced by other specific families of distributions. The question of what (if anything) to do about moderate departures from normality deserves study. The array-CGH data appear to have tails somewhat longer than normal, although a slightly incorrect segmentation might exaggerate the tails of the residuals. To the extent that the likelihood ratio involves sums of observations, one might expect the central limit theorem to make the normality hypothesis reasonable provided segments of constant mean are reasonably long; but one cannot (especially for array-CGH data) exclude the possibility of quite short segments.

In Section 6, we gave a version of the modified BIC that selects the number of levels as well as the number of change-points. In Theorem 3, the level assignment of each segment between change-points is unknown, but the result is easily adapted to the simpler case where this information is known. For example, we may wish to restrict the changes so that the process always returns to a baseline mean level after it jumps to a changed state. The modified BIC for this case would be the same as equation (14), but  $k$  would be a function of  $m$  and the level assignments of the “baseline” segments would be fixed. In this way, the modified BIC is very flexible in adapting to model restrictions on the magnitude of change. However, for these restricted change models, the CBS algorithm would no longer be adequate as a search method for finding the optimal change-points, since it does not take into account the level assignments.

One difference between the modified BIC and the original BIC is that the remainder term for the modified BIC is a bounded random variable, while for the original BIC it is a bounded constant. For this reason, the modified BIC is less stable than the original BIC. As in Schwarz (1978) we left out of the modified BIC all terms in the approximation of the Bayes factor that do not grow with the sample size. However, in some cases, it may be beneficial to include terms of order  $m \log k$ . Inclusion of such terms involves arbitrariness and introduces a certain degree of bias. Which terms to include and under what circumstances they can be justified is an interesting topic for future investigation.

Finally, we would like to emphasize that the modified BIC is a general model selection method that is not limited to the analysis of array-CGH data.

## 8. Supplementary Materials

Matlab code for the procedure in Section 3 can be obtained by email from the authors. The supporting Web Appendix is available at <http://www.tibs.org/biometrics>.

## ACKNOWLEDGEMENTS

The research of the first author was supported by a National Defense Science and Engineering Graduate Fellowship. The research of the second author was supported by grants from the NSF and NIH.

## REFERENCES

- Albertson, D. J., Collins, C., McCormick, F., and Gray, J. W. (2003). Chromosome aberrations in solid tumors. *Nature Genetics* **34**, 369–376.
- Birgé, L. and Massart, P. (2001). Gaussian model selection. *Journal of the European Mathematical Society* **3**, 203–268.
- Fridlyand, J., Snijders, A., Pinkel, D., Albertson, D. G., and Jain, A. N. (2004). Application of hidden Markov models to the analysis of the array-CGH data. *Special Genomic Issue of the Journal of Multivariate Analysis* **90**, 132–153.
- George, E. I. and McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association* **88**, 881–889.
- Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**, 711–732.
- Gu, C. and Wang, J. (2003). Penalized likelihood density estimation: Direct cross-validation and scalable approximation. *Statistica Sinica* **13**, 811–826.
- Hsu, L., Self, S. G., Grove, D., Randolph, T., Wang, K., Delrow, J. J., Loo, L., and Porter, P. (2005). Denoising array-based comparative genomic hybridization data using wavelets. *Biostatistics* **6**, 211–226.
- James, B., James, K. L., and Siegmund, D. (1987). Tests for a change-point. *Biometrika* **74**, 71–84.
- Lavielle, M. (2005). Using penalized contrasts for the change-point problem. *Signal Processing* **85**, 1501–1510.
- Li, W. (2001). DNA segmentation as a model selection process. In *Proceedings of the Fifth International Conference on Computational Biology*, T. Lengauer, D. Sankoff, S. Istrail, P. Pevzner, and M. Waterman (eds), 204–210. New York: Association for Computing Machinery.
- 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.
- Picard, F., Robin, S., Lavielle, M., Vaisse, C., and Daudin, J. (2005). A statistical approach for array CGH data analysis. *BMC Bioinformatics* **6**, 27.
- Pinkel, D., Seagraves, R., Sudar, D., et al. (1998). High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nature Genetics* **20**, 207–211.
- Pollack, J. R., Perou, C. M., Alizadeh, A. A., Eisen, M. B., Pergamenschikov, A., Williams, C. F., Jeffrey, S. S., Botstein, D., and Brown, P. O. (1999). Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nature Genetics* **23**, 41–46.
- Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* **6**, 461–464.
- Siegmund, D. (1992). Tail approximations for maxima of random fields. In *Probability Theory: Proceedings of the 1989 Singapore Probability Conference*, L. H. Y. Chen, K. P. Choi, K. Yu, and J.-H. Lou (eds), 147–158. Berlin: de Gruyter.
- Siegmund, D. (2004). Model selection in irregular problems: Applications to mapping quantitative trait loci. *Biometrika* **92**, 785–800.
- Snijders, A. M., Nowak, N., Seagraves, R., et al. (2001). Assembly of microarrays for genome-wide measurement of DNA copy number. *Nature Genetics* **29**, 263–264.

- Snijders, A. M., Fridlyand, J., Mans, D. A., Segraves, R., Jain, A. N., Pinkel, D., and Albertson, D. G. (2003). Shaping of tumor and drug-resistant genomes by instability and selection. *Oncogene* **22**, 4370–4379.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B* **58**, 267–288.
- Wang, P., Kim, Y., Pollack, J., Narasimhan, B., and Tibshirani, R. (2005). A method for calling gains and losses in array-CGH data. *Biostatistics* **6**, 45–58.
- Yao, Y. C. (1988). Estimating the number of change-points via Schwarz criterion. *Statistics and Probability Letters* **6**, 181–189.
- Zhang, N. R. (2005). Change-point detection and sequence alignment: Statistical problems of genomics. Ph.D. Thesis, Statistics Department, Stanford University, Stanford, California.

*Received July 2005. Revised May 2006.*

*Accepted May 2006.*