Stepwise Signal Extraction via Marginal Likelihood

Chao Du, Chu-Lan Michael Kao, and S. C. Kou

This article studies the estimation of a stepwise signal. To determine the number and locations of change-points of the stepwise signal, we formulate a maximum marginal likelihood estimator, which can be computed with a quadratic cost using dynamic programming. We carry out an extensive investigation on the choice of the prior distribution and study the asymptotic properties of the maximum marginal likelihood estimator. We propose to treat each possible set of change-points equally and adopt an empirical Bayes approach to specify the prior distribution of segment parameters. A detailed simulation study is performed to compare the effectiveness of this method with other existing methods. We demonstrate our method on single-molecule enzyme reaction data and on DNA array comparative genomic hybridization (CGH) data. Our study shows that this method is applicable to a wide range of models and offers appealing results in practice. Supplementary materials for this article are available online.

KEY WORDS: Array comparative genomic hybridization; Asymptotic consistency; Change-points; Choice of prior; Dynamic programming; Single-molecule experiment.

1. INTRODUCTION

In signals measured at successive times or locations, abrupt changes are often present. Such changes reflect the evolution of the underlying system and effectively break the signal into segments. Within a segment, the observations are homogeneously distributed; between adjacent segments, the signal has distinct characteristics. One notable case is the stepwise signal, characterized by, but not limited to, the shift of means between successive segments, as often encountered in modern biophysical experiments. For example, in the study of single-molecule enzymology, the fluorescence intensity of an enzyme molecule fluctuates in a stepwise fashion in response to the conformational change of molecule over time (Lu, Xun, and Xie 1998; English et al. 2006; Kou 2008). In eukaryotic cells, kinesin, a molecular motor protein, moves along a micro-tube in discrete steps; its movement displays a stepwise pattern (Yildiz et al. 2004). Similar examples also occur in various other disciplines, such as organizing DNA sequences into homogeneous segments (Braun and Müller 1998), detecting chromosomal aberration in the DNA array data (Lai et al. 2005), studying neuromuscular activation patterns and control in electromyography (Johnson, Elashoff, and Harkema 2003), probing the layered rocks from nuclear-magnetic response in oil drilling (O Ruanaidh and Fitzgerald 1996), estimating the market structural changes in equity markets (Bai and Perron 1998, 2003; Bekaert, Harvey, and Lumsdaine 2002), and analyzing the changes of coal mine disasters in Britain (Jarrett 1979).

In statistics literature, the time points or the locations where the abrupt changes take place are often referred to as change-points. The term "change" in the broad sense not only refers to the change of distributional parameters, but also includes all other possible variations of the underlying model, such as the changes of explanatory variables in linear regression (Bai and Perron 1998, 2003) and the parameter changes in hidden Markov models (Fuh 2003, 2004). Our study in change-point problems is mainly motivated by the aforementioned biophysical applications, where extracting information about the length of segments from experimental data is often the key and first step to understand the underlying complex biological system. Therefore, our discussion will focus on estimating the number and locations of change-points in stepwise signals.

The early study on change-point problems started with the assumption of at most one change-point. The change of means in a series of normally distributed variables was studied by Chernoff and Zacks (1964). The location of a single change-point can be inferred using frequentist maximum likelihood estimation (MLE) (Hinkley 1970) or Bayesian posterior distribution (Smith 1975; Carlin, Gelfand, and Smith 1992). Hypothesis testing on the existence of the change-point was developed as well (Bhattacharya 1994).

In multiple-change-point problems, most frequentist methods draw inference through optimizing a fitness or cost function, such as log-likelihood or sum of squared errors over all possible segmentations. As the total number of change-points is usually unknown, model selection tools, often in the form of a penalty function, are implemented alongside the optimization procedure to avoid overfitting. For normally distributed data with constant variance, the Bayesian information criterion (BIC) (Yao 1988; Yao and Au 1989) or its modification (Zhang and Siegmund 2007) can be used. Braun, Braun, and Müller (2000) generalized the BIC approach to include a broader distribution family in which the variance is proportional to a function of the mean. Under the equal-variance assumption, the sum of squares with an appropriate penalty term works as well (Boysen et al. 2009). In the fused lasso method (Tibshirani et al. 2005; Tibshirani and Wang 2008), an L_1 penalty which penalizes differences between successive segment means is added to the least-square term. Recently, Frick, Munk, and Sieling (2014) estimated the unknown step function through minimizing the number of change-points

Chao Du is Assistant Professor of Statistics, University of Virginia, Charlottesville, VA 22904 (E-mail: *cd2wb@virginia.edu*). Chu-Lan Michael Kao is Postdoctoral Fellow, Research Center of Adaptive Data Analysis, National Central University, Taoyuan County 32001, Taiwan (E-mail: *chulankao@gmail.com*). S. C. Kou, the corresponding author, is Professor of Statistics, Harvard University, Cambridge, MA 02138 (E-mail: *kou@stat.harvard.edu*). S. C. Kou's research is supported in part by NIH/NIGMS grant R01GM090202. The authors thank Professor Alex Munk and Professor Cheng-Der Fuh for valuable discussions, the Xie group at the Department of Chemistry and Chemical Biology of Harvard University for sharing the experimental data, and the Ministry of Science of Technology, R.O.C for the support of a fellowship for CLK.

© 2016 American Statistical Association
Journal of the American Statistical Association
March 2016, Vol. 111, No. 513, Theory and Methods
DOI: 10.1080/01621459.2015.1006365

over the acceptance region of a multiscale test. The formidable task of searching over an astronomical number of change-point configurations can be handled by heuristic methods such as binary segmentation (Scott and Knott 1974) and circular binary segmentation (Olshen et al. 2004; Venkatraman and Olshen 2007) methods, genetic algorithms (Davis, Lee, and Rodriguez-Yam 2006), or exact dynamic programming algorithms (Bellman and Roth 1969; Bement and Waterman 1977; Auger and Lawrence 1989; Jackson et al. 2005; Killick, Fearnhead, and Eckley 2012).

From a Bayesian perspective, imposing priors over the locations of change-points and segment parameters would automatically penalize unnecessary model complexity. The inference can then be handled by Markov chain Monte Carlo (MCMC) sampling. Gibbs sampling is commonly employed as the conditional distributions of the target density are generally well defined (Barry and Hartigan 1993; Chib 1998). Other sampling techniques include reversible jump MCMC (Green 1995), sequential importance sampling and particle filtering (Koop and Potter 2007; Fearnhead and Liu 2007), and the adaptive Metropolis–Hastings algorithm (Giordani and Kohn 2008). Still, for problems with many change-points, MCMC algorithms may fail due to nonconvergence or the strong dependence between the samples. Marginal likelihood can be used to integrate out the segment parameters to reduce the sampling dimension, as in the binary segmentation procedure of Yang and Kuo (2001) and the MCMC algorithm of Wyse and Friel (2010). Fearnhead (2005, 2006) proposed an algorithm based on recursive computation for sampling from the exact posterior distribution of change-points. A similar scheme can also be applied to deduce the exact posterior means of segment parameters (Lai and Xing 2011).

Theoretical investigations of the estimation of change-points are mainly conducted from the frequentist perspective. The consistency of frequentist change-point estimators can often be established if the added penalty terms meet certain criteria. A detailed theoretical study on the penalty criteria, the asymptotic consistency and rates of convergence of the estimators, can be found in Boysen et al. (2009). These asymptotic results typically require that the observed data follow certain distribution family. In practice, cross-validation is often necessary in choosing penalty terms, but the use of cross-validation presents additional theoretical challenges, such as the effect of adaptation.

In contrast, Bayesian approaches using prior distributions are more flexible and can be applied to broad classes of models. The theoretical properties of Bayesian methods, however, have received less investigation. In this article, we take an (empirical) Bayesian perspective. We formulate a maximum marginal likelihood estimator for stepwise signal estimation. As discussed in the preceding paragraphs, various aspects have been considered in the literature. In this article, we carry out a comprehensive investigation of the marginal likelihood method.

1. We conduct a detailed investigation on the choice of prior. For the prior distribution of change-point locations, we impartially treat each possible configuration of change-points as an individual model to minimize the influence of the prior. For the prior distribution of segment parameters, we study its impact on the estimator and propose an effective empirical Bayesian method for the prior specification.
2. The asymptotic properties of the estimator, including the asymptotic consistency, are established.
3. We study efficient computation, outlining fast dynamic programming methods.
4. The finite-sample performance of the estimator is studied in depth, which offers guidelines for the practical use of the method.
5. We address the problem of evaluating different change-point estimators, which has not been well studied because there is no simple one-to-one mapping between the estimated change-points and the true change-points. We propose criteria for this evaluation and compare our method to a number of existing methods.

This article is organized as follows. In Section 2, we define notations used throughout this article and outline our method. Asymptotic theory of our method is discussed in Section 3. In Section 4, we discuss the role of prior in our method and provide guidelines for choosing the prior, especially when the data follow a normal or Poisson distribution. We then perform extensive simulation experiments to compare our method to six existing methods in Section 5. In Section 6, we apply our method to real examples, including the array comparative genomic hybridization (CGH) data and the fluorescence intensity trajectory of a single enzyme molecule. This article ends in Section 7 with a summary and concluding remarks. The R and Matlab packages that implement our marginal likelihood method can be downloaded at *http://www.people.fas.harvard.edu/~skou/publication.htm*.

2. THE MARGINAL LIKELIHOOD METHOD FOR STEPWISE SIGNALS

2.1 Basic Notations

Assume that we have a dataset $\boldsymbol{x} = \{x_1, x_2, \ldots, x_n\}$, measured at successive times $\boldsymbol{t} = \{t_1, t_2, \ldots, t_n\}$, $t_1 < t_2 < \cdots < t_n$. Such data are not necessarily limited to time series since they can also represent a spatial sequence, but for simplicity we will use terms and notations usually associated with temporal dimension throughout this article. Also, there is no restriction on the dimension of x_i.

The underlying parameter $\theta \in \Theta$ determines the distribution of $\boldsymbol{x} = \{x_i\}_{i=1}^n$ through a family of densities $f(x|\theta)$. There is no restriction on the dimension of parameter θ either. We assume that θ is a step function of time whose transitions are determined by $m-1$ change-points $\boldsymbol{\tau}_{1:(m-1)} = \{\tau_1, \ldots, \tau_{m-1}\}$:

$$\theta(t) = \theta_j \quad \text{if } t \in (\tau_{j-1}, \tau_j], \tag{2.1}$$

where $\tau_j \in [t_1, t_n]$ for $j \in \{1, \ldots, m-1\}$. The $m-1$ change-points split the signal into m segments. We refer to $\boldsymbol{\theta}_{1:m} = \{\theta_j\}_{j=1}^m$ as the segment parameters. We also assume that the adjacent θ_j's are distinguishable through $f(\cdot|\cdot)$, that is, the measure of the set $\{x : f(x|\theta_j) \neq f(x|\theta_{j+1})\}$ is greater than 0 for all $1 \leq j < m$. Given the change-points $\boldsymbol{\tau}_{1:(m-1)}$ and the associated segment parameters $\boldsymbol{\theta}_{1:m}$, the observations are assumed to be independently distributed:

$$P(\boldsymbol{x}|\boldsymbol{\tau}_{1:(m-1)}, \boldsymbol{\theta}_{1:m}) = \prod_{j=1}^{m} \prod_{t_i \in (\tau_{j-1}, \tau_j]} f(x_i|\theta_j). \tag{2.2}$$

The observations up to time τ_1 have density $f(\cdot|\theta_1)$; the observations after time τ_1 but up to τ_2 have density $f(\cdot|\theta_2)$; ...; the observations after time τ_{m-1} are characterized by the parameter θ_m. Please note that we set $\tau_0 \equiv 0$ and $\tau_m \equiv t_n$ for notational ease.

Although it is not necessary that the change-points can only take discrete values from the set $\{t_i\}_{i=1}^n$, it is often pointless to work on a higher resolution without further model assumption. Hence, unless specifically stated otherwise, we will assume that $\tau_j \in \{t_1, \ldots, t_{n-1}\}$ for $j \in \{1, \ldots, m-1\}$.

Assume that only $\{x_i\}_{i=1}^n$ and $\{t_i\}_{i=1}^n$ are available and that the parametric form of $f(x|\theta)$ is known. The estimation goal is to determine the number m and positions $\{\tau_j\}_1^{m-1}$ of the change-points along with the segment parameters $\boldsymbol{\theta}_{1:m}$ from the observations.

2.2 The Maximum Marginal Likelihood Estimator

We approach the problem by using the marginal likelihood in which $\boldsymbol{\theta}_{1:m}$ are integrated out. We assume that, given the set of change-points $\boldsymbol{\tau}_{1:(m-1)}$, $\boldsymbol{\theta}_{1:m}$ are independently and identically drawn from a prior distribution $\pi(\cdot|\alpha)$. In the literature, the hyperparameter(s) α can be either modeled as constant (Chib 1998; Fearnhead 2005, 2006), or with a hyperprior distribution (Carlin, Gelfand, and Smith 1992; Barry and Hartigan 1993; Pesaran, Pettenuzzo, and Timmermann 2006; Koop and Potter 2007). The latter approach, though can be potentially handled by MCMC sampling, introduces dependence between the segments and, thus, undermines the possibility of applying recursive algorithms to accelerate the computation (Fearnhead 2005, 2006; Lai and Xing 2011). For this reason, we model α as preset constants; the choice of α will be discussed in Section 4.

We can express the marginal likelihood, given the set of change-points, as

$$\begin{aligned} P(\boldsymbol{x}|\boldsymbol{\tau}_{1:(m-1)}) &= \prod_{j=1}^{m} \int_{\theta_j} \prod_{t_i \in (\tau_{j-1}, \tau_j]} f(x_i|\theta_j)\pi(\theta_j|\alpha)d\theta_j \\ &= \prod_{j=1}^{m} D(\boldsymbol{x}_{(\tau_{j-1},\tau_j]}|\alpha), \end{aligned} \tag{2.3}$$

where, in general, $D(\boldsymbol{x}_{(a,b]}|\alpha)$ denotes the probability of obtaining the observations during the period $(a, b]$ with no change-point in between. A closed form of $D(\boldsymbol{x}_{(a,b]}|\alpha)$ can be obtained if conjugate priors are used. Otherwise, we may estimate $D(\boldsymbol{x}_{(a,b]}|\alpha)$ through numerical methods or approximation schemes such as Laplace's method.

We take each distinctive set of change-points $\boldsymbol{\tau}_{1:(m-1)}$ as a specific model, and estimate the set of change-points as the maximizer of $P(\boldsymbol{x}|\boldsymbol{\tau}_{1:(m-1)})$ over all feasible combinations of change-points, restricted by an upper bound $M \leq n$ on the number of segments. Such an upper bound often arises in biological data; for instance, in chemical experiments, reaction rate considerations typically limit the number of reaction cycles in a given time window. In the extreme case of $M = n$, every observation t_i can be a segment itself.

Note that if we assign a uniform prior $P(\boldsymbol{\tau}_{1:(m-1)}) \propto 1$, $(m \leq M)$ on the set of change-points, which is proper since the total number of change-points is bounded, then

$$P(\boldsymbol{\tau}_{1:(m-1)}|\boldsymbol{x}) \propto P(\boldsymbol{x}|\boldsymbol{\tau}_{1:(m-1)})P(\boldsymbol{\tau}_{1:(m-1)}) = P(\boldsymbol{x}|\boldsymbol{\tau}_{1:(m-1)}).$$

Thus, our approach of maximizing the marginal likelihood is equivalent to finding the maximum a posteriori estimate of the change-points with a uniform prior.

The prior we impose here can be viewed as "noninformative" in the sense that it implies that the prior probability of observing a change-point at t_i is the same for all $i \in \{1, \ldots, n-1\}$ (less than M/n). We choose this prior to minimize the impact of prior on the segmentation of signal. In contrast, priors employed in various Bayesian methods tend to impose certain transition structures or favor certain locations. For example, Chib (1998) modeled the change-point locations through a one-way hidden Markov chain, which implied a near geometric distribution on the segment length. Koop and Potter (2009) also pointed out that such prior favors change-points near the end of the time window and suggested a noninformative prior so that the conditional prior distribution $p(\tau_j|\tau_{j-1})$ is uniformly over a finite support of constant length. This prior, however, is still not entirely "uniform", as there is a slightly larger chance of observing a change-point at t_2 than at t_1, for example. To construct a prior free of specific transition structures, it seems natural to assign a prior, flat or not, over the total number of change-points and then treat all the configurations with the same number of change-points equally likely, as in Fearnhead (2005) and Moreno, Javier Girón, and García-Ferrer (2013). The prior probability of a specific configuration with m change-points would then equal the prior probability of having m total change-points divided by the total number of distributing m change-points. Consequently, for $m_1 > m_2$, a configuration with m_1 change-points would receive a much heavier penalty than a configuration with m_2 change-points (as long as $m_1 < n/2$). In this regard, a configuration with more change-points is penalized not directly because it represents a more complicated model, but because of the existence of many more models with the same number of change-points (this echoes what physicists call "entropic effect"). This is the rationale behind our choice of the prior. It must be admitted, though, the notion of "noninformative" in the setting of a multiple-change-point problem is ambiguous at best, and our choice is only one in many ways to interpret it. Nonetheless, we will demonstrate in the subsequent sections that such construction does yield proper estimation of the change-points.

2.3 Fast Computation Through Dynamic Programming

In our formulation of a change-point model, $P(\boldsymbol{x}|\boldsymbol{\tau}_{1:(m-1)})$ can be expressed as a product of nonoverlapping $D(\boldsymbol{x}_{(a,b]}|\alpha)$. Thus, dynamic programming (Bellman and Roth 1969; Bement and Waterman 1977; Auger and Lawrence 1989) can be applied. Suppose that $M \leq n$ is an upper bound for the number of segments, we suggest the following algorithm:

$$\begin{aligned} H(x_1, \ldots, x_i|m) = &\max_{\boldsymbol{\tau}_{1:(m-1)} \subseteq \{t_1, \ldots, t_{i-1}\}} P(x_1, \ldots, x_i|\boldsymbol{\tau}_{1:(m-1)}), \\ &m = 1, \ldots, M. \end{aligned}$$

Step 1. For $1 \leq i \leq n$: $H(x_1, \ldots, x_i|1) = D(x_1, \ldots, x_i|\alpha)$

Step m. For $m \leq i \leq n$: $H(x_1, \ldots, x_i|m) = \max_{m-1 \leq j \leq i-1} H(x_1, \ldots, x_j|m-1)D(x_{j+1}, \ldots, x_i|\alpha)$

Step M. For $M \leq i \leq n$: $H(x_1, \ldots, x_i|M) = \max_{M-1 \leq j \leq i-1} H(x_1, \ldots, x_j|M-1)D(x_{j+1}, \ldots, x_i|\alpha)$

Using the above recursive functions, we can obtain the following estimators with computational cost $O(n^2M)$ and storage $O(nM)$:

- the maximum marginal likelihood estimator $\hat{\boldsymbol{\tau}}_{1:(m-1)}$ with exactly m segments ($m \le M$)

$$\hat{\boldsymbol{\tau}}_{1:(m-1)} = \underset{\boldsymbol{\tau}_{1:(m-1)}}{\arg\max}\ P\left(\boldsymbol{x}|\boldsymbol{\tau}_{1:(m-1)}\right), \tag{2.4}$$

- the maximum marginal likelihood estimator $\hat{\boldsymbol{\tau}}_M$ with up to M segments

$$\hat{\boldsymbol{\tau}}_M = \underset{\boldsymbol{\tau}_{1:(m-1)},\, 1\le m\le M}{\arg\max}\ P\left(\boldsymbol{x}|\boldsymbol{\tau}_{1:(m-1)}\right). \tag{2.5}$$

Jackson et al. (2005) developed a more efficient but less flexible algorithm in which the maximum marginal likelihood estimator $\hat{\boldsymbol{\tau}}$ (with up to n segments) can be computed with computational cost $O(n^2)$ and storage $O(n)$. This algorithm is based on the following recursive functions:

$$G(x_1,\ldots,x_i) = \max_{\boldsymbol{\tau}\subseteq\{t_1,\ldots,t_{i-1}\}} P(x_1,\ldots,x_i|\boldsymbol{\tau}),\ i=1,\ldots,n.$$

Step 1. $G(x_1) = D(x_1|\alpha)$
Step i. $G(x_1,\ldots,x_i) = \max_{1\le j\le i-1} G(x_1,\ldots,x_j)D(x_{j+1},\ldots,x_i|\alpha)$
Step n. $G(x_1,\ldots,x_n) = \max_{1\le j\le n-1} G(x_1,\ldots,x_j)D(x_{j+1},\ldots,x_n|\alpha)$

Generally speaking, for a large value of M, we expect that $\hat{\boldsymbol{\tau}}_M$ from the first algorithm is identical to $\hat{\boldsymbol{\tau}}$ from the second algorithm. Thus, the second algorithm is the algorithm of choice for large M. On the other hand, if there is a strong restriction on the number of segments M or one needs to compare models with different numbers of change-points, the first algorithm should be used.

It is possible to further speed up the dynamic programming algorithms. One possibility is to reduce the computation by imposing restrictions on the potential change-point sequence. For example, we could put a lower or an upper bound to the size of segments. Such restriction can be easily adapted into dynamic programming and may speed up the computation without sacrificing much accuracy. Another possibility is to try to eliminate unnecessary steps in the algorithm. Killick, Fearnhead, and Eckley (2012) proposed a pruned exact linear time (PELT) method in which the computational cost could be improved up to $O(n)$. However, to apply a PELT method in our setting, there must be a positive constant C such that $\frac{D(x_i,\ldots x_j|\alpha)D(x_{j+1},\ldots x_k|\alpha)}{D(x_i,\ldots x_k|\alpha)} \ge C$ for all $1 \le i < j < k \le n$. Unfortunately, it is impossible to find a general C for this inequality in the case of marginal likelihood. Still, we would like to examine this possibility in our future work.

3. ASYMPTOTIC STUDY OF THE MARGINAL LIKELIHOOD METHOD

Before we start a rigorous theoretical investigation, we would like to present an intuitive explanation of why the estimator based on marginal likelihood would not overestimate the number of change-points. Suppose that there is no change-point for the sequence $(x_1,\ldots,x_n)$, then based on the consistency of the maximum a posteriori estimator $\hat{\theta}$, the associated marginal likelihood can be approximated by the Laplace method as

$$\begin{aligned}\log D(x_1\ldots,x_n|\alpha) &\approx \sum_i l(x_i|\hat{\theta})) + \log\pi(\hat{\theta}|\alpha) + \frac{1}{2}\log(2\pi/n)\\ &\quad - \frac{1}{2}\log\left(\frac{1}{n}|\sum_i l''(x_i|\hat{\theta}) + (\log\pi(\hat{\theta}|\alpha))''|\right)\\ &\approx \text{Const} + \sum_i l(x_i|\hat{\theta}) - \frac{1}{2}\log n,\end{aligned}$$

where $l(x|\theta) = \log f(x|\theta)$. The logarithm of the marginal likelihood can thus be perceived as the logarithm of maximum likelihood with an additional term $-\frac{1}{2}\log n$. This term in effect places a heavy penalty on models with unnecessary change-points since $\log n_1 + \log n_2 > \log(n_1+n_2)$. This is why we do not have to explicitly put a penalty term in our estimation.

Similarly, if the sequence $(x_1,\ldots,x_n)$ can be divided into m segments and the ith segment contains n_i observations, omitting the terms that correspond to the prior, the logarithm of marginal likelihood can be approximated by the logarithm of maximum likelihood plus a penalty $-1/2\sum_{i=1}^m \log n_i$. Following the parameterization used in Zhang and Siegmund (2007), this penalty can be rewritten as

$$-\frac{1}{2}\sum_{i=1}^m \log n_i = -\frac{1}{2}m\log n - \frac{1}{2}\sum_{i=1}^m \log k_i,$$

where $k_i = n_i/n$. In this expression, the first term is identical to the classical BIC penalty, while the second term is minimized when the change-points are evenly spaced and maximized when the change-points are placed as close as possible. Thus, in light of this penalty function, our maximum marginal likelihood method modifies the BIC by penalizing not only too many change-points, but also the placement of change-points. A similar penalty term was proposed by Zhang and Siegmund (2007) in the form of $-\frac{3}{2}(m-1)\log n - \frac{1}{2}\sum_{i=1}^m \log k_i$ in which more weight is placed on the BIC penalty.

The rest of this section is devoted to a rigorous study of the asymptotic properties of the maximum marginal likelihood estimator. We shall prove that, under suitable conditions, the set of estimated change-points would converge to the set of true change-points in probability.

Without loss of generality, we assume that all observations are made within the time interval (0, 1]: $0 < t_i \le 1$. We assume that there are m_0 segments in total, which are defined by the $m_0 - 1$ true change-points $0 < \tau_1^0 < \tau_2^0 < \cdots < \tau_{m_0-1}^0 < 1$. For technical reason, we will also treat $\tau_0^0 = 0$ and $\tau_{m_0}^0 = 1$ as change-points. We denote $\boldsymbol{\tau}^0 = \{\tau_j^0\}_{j=0}^{m_0}$. The true segment parameters will be denoted as $\{\theta_j\}_{j=1}^{m_0}$. In the case of no change-point in (0, 1), according to our definition, the true set of change-points would be $\boldsymbol{\tau}^0 = \{0, 1\}$ and the associated parameter is θ_1. The prior density of θ_j is represented by $\pi(\cdot|\alpha)$, where α is the hyperparameter. We denote the number of observations within a given interval as $n_{(a,b]} = \#\{i : a < t_i \le b, 1 \le i \le n\}$. In addition, we also define the shortest distance between two consecutive change-points in change-point sequence $\{\tau_j\}_{j=0}^m$ as

$$\Delta(\{\tau_j\}_{j=0}^m) := \min\{\tau_j - \tau_{j-1} : j = 1,\ldots,m\}.$$

We let our maximum marginal likelihood estimator be

$$\hat{\boldsymbol{\tau}} = \{\hat{\tau}_j\}_0^{\hat{m}} = \underset{\substack{\{\tau_j\}_1^{m-1}\subseteq\{t_i\}_1^n,\tau_0=0,\tau_m=1 \\ 1\le m\le n \\ \Delta\left(\{\tau_j\}_{j=0}^m\right)\ge\overline{\Delta}(n)}}{\arg\max} P\left(\boldsymbol{x}|\{\tau_j\}_{j=0}^m\right),$$

where $\overline{\Delta}(n) > 0$ serves as a lower bound for the time lag between two consecutive change-points.

We shall prove that, under the following regularity conditions, the estimated change-points $\{\hat{\tau}_j\}_0^{\hat{m}}$ converge to the true change-points $\{\tau_j^0\}_{j=0}^{m_0}$ in location and in total number:

1. The prior density $\pi(\theta|\alpha)$ is continuous and positive at all θ_j $(1 \leqslant j \leqslant m_0)$.
2. For any adjacent θ_j and θ_{j+1} $(1 \leqslant j \leqslant m_0 - 1)$, there exists a neighborhood $N_j(\delta) = \{\theta : \left\|\theta - \theta_j\right\| < \delta\}$ of θ_j and a neighborhood $N_{j+1}(\delta) = \{\theta : \left\|\theta - \theta_{j+1}\right\| < \delta\}$ of θ_{j+1} such that $N_j(\delta) \bigcap N_{j+1}(\delta) = \emptyset$.
3. Given any interval $(a, b]$ $(0 < a < b \leqslant 1)$, $n_{(a,b]}/n \to C_{(a,b]} > 0$ as $n \to \infty$, where the constant $C_{(a,b]}$ depends on a and b. Moreover, $\inf\{C_{(a,b]}/(b-a) : 0 < a < b < 1\} > 0$.
4. The segment parameters θ_j and the density function $f(\cdot|\cdot)$ satisfy conditions (A1)–(A5) and (B1)–(B4) listed in the supplementary material.
5. $\overline{\Delta}(n) \to 0$ and $n\overline{\Delta}(n) \to \infty$, as $n \to \infty$.

The first regularity condition ensures the proper behavior of the prior around the true parameter values. The second regularity condition ensures that adjacent parameters are distinguishable. The third regularity condition defines what we mean by “asymptotics” for a stepwise signal: The number of observations within any interval should approach infinity as the total number of observations goes to infinity. However, there is no requirement for the observational density to be uniform. The fourth group of regular conditions are to ensure the usual asymptotic consistency and normality of the MLE of θ_j. The last condition ensures that the estimated number of change-points would converge.

Under these conditions, we have the following asymptotic results.

Lemma 3.1. Assume regularity conditions (1)–(4). If the true set of change-points $\{\tau_j^0\}_0^{m_0} = \{0, 1\}$, that is, there is only one segment and no real change-point, then as $n \to \infty$, for any given set of change-points $\{\tau_j\}_0^m \neq \{0, 1\}$, we have the ratio of marginal likelihood

$$\frac{P\left(\boldsymbol{x}|\{\tau_j\}_0^m\right)}{P\left(\boldsymbol{x}|\{0, 1\}\right)} = O_p\left(1/\sqrt{n\Delta_\tau}\right),$$

where $\Delta_\tau = \Delta(\{\tau_j\}_0^m)$.

Lemma 3.2. Assume regularity conditions (1)–(4). If the true set of change-points $\{\tau_j^0\}_0^{m_0} \neq \{0, 1\}$, that is, there is at least one real change-point, then as $n \to \infty$, we have the ratio of marginal likelihood

$$\frac{P\left(\boldsymbol{x}|\{0, 1\}\right)}{P\left(\boldsymbol{x}|\{\tau_j^0\}_0^{m_0}\right)} = O_p\left(\sqrt{n\Delta_0}\exp(-cn\Delta_0)\right),$$

where $\Delta_0 = \Delta(\{\tau_j^0\}_0^{m_0})$, and $c > 0$ is a constant.

Lemma 3.1 and Lemma 3.2 indicate that maximizing the marginal likelihood would not result in overfitting or underfitting. Using them, the consistency of our estimator $\hat{\boldsymbol{\tau}}$ is established in the next theorem. The proofs are deferred to the supplementary material.

Theorem 3.1. Assume regularity conditions (1)–(5). Let $\hat{\boldsymbol{\tau}}$ be the estimated set of change-points, and $\hat{\boldsymbol{\tau}}^{\mathbf{0}}$ be the true set of change-points. Then, as $n \to \infty$, $\hat{m} \xrightarrow{P} m_0$, and

$$\sup_{\tau^0\in\hat{\boldsymbol{\tau}}^{\mathbf{0}}} \inf_{\hat{\tau}\in\hat{\boldsymbol{\tau}}} \left|\hat{\tau} - \tau^0\right| \xrightarrow{P} 0. \tag{3.1}$$

The theorem assumes fixed hyperparameter(s) α. In fact, it can be strengthened to cover empirical Bayes maximum marginal likelihood estimators in which the hyperparameter(s) are themselves estimated from the data.

Corollary 3.1. Assume regularity conditions (1)–(5). Let $\hat{\alpha}_n$ be a sequence of hyperparameter estimators. Suppose $\hat{\alpha}_n \xrightarrow{P} \alpha^*$, and $\pi(\theta|\alpha)$ is continuous at α^*. Then, all the results in Theorem 3.1 hold when $\pi(\theta|\alpha)$ are replaced by $\pi(\theta|\hat{\alpha}_n)$.

Remark 1. The restriction on the time lag $\overline{\Delta}$ is a relatively minor condition. A similar condition can also be found in Frick, Munk, and Sieling (2014). To see why such a condition is necessary, simply consider the case where t_i are an ordered iid sequence from Unif(0, 1). Then, the independence of x_i implies that there would be a positive probability that the algorithm set a change-point on t_1. The $\overline{\Delta}$ restriction is to asymptotically avoid this kind of situation. In practice, one could simply assume that $\overline{\Delta}$ is smaller than $\min\{t_{j+1} - t_j : j = 0, \ldots, n-1\}$, which essentially removes the condition.

4. THE CHOICE OF PRIOR DISTRIBUTION

Section 3 shows that, given the regularity conditions, the asymptotic consistency of the maximum marginal likelihood estimator does not depend on the specific choice of the prior. For a finite sample, however, the choice of prior or the choice of hyperparameters could affect the final estimation. In this section, we will discuss the role of the prior distribution and the choice of hyperparameters.

Intuitively speaking, in the marginal likelihood approach, the prior serves as a ruler in deciding whether two adjacent segments of data are similar enough to be regarded as from the same distribution. Consider the following example as an illustration: Two observations x_1 and x_2 follow normal distributions with variance 1 and unknown means. Let the prior of the unknown mean be $N(0, \sigma_0^2)$. Then, for a relative large σ_0^2, the ratio of marginal likelihood is approximately

$$\frac{D\left(x_1|\sigma_0^2\right) D\left(x_2|\sigma_0^2\right)}{D\left(x_1, x_2|\sigma_0^2\right)} \approx \frac{\sqrt{2}}{\sigma_0} \exp\left((x_1 - x_2)^2/4\right).$$

For fixed x_1 and x_2, a strong prior (i.e., a small σ_0^2) leads to a large ratio, favoring the model with one change-point over the model without. More importantly, this ratio is determined by both σ_0^2 and $x_1 - x_2$: The strength of the prior is always relative to information contained in the data.

The same phenomenon can also be observed in more sophisticated and realistic examples. Figure 1 shows a stepwise

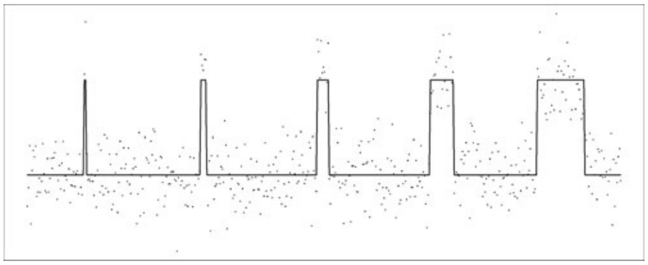

Figure 1. A stepwise signal with five high-level segments.

signal, the solid line, which contains five high-level segments (with value 1) and baseline segments (with value 0). This example is motivated by the array CGH data, where the high-level segments correspond to abnormal regions in a genomic DNA sequence. We will discuss this example in greater details soon. Here we use this example to investigate the effect of prior on the estimation. The dots in Figure 1 are 500 simulated observations $x_i|(\mu_j, \sigma_j^2) \sim N(\mu_j, \sigma_j^2)$, where $\sigma_j^2 = 0.25^2$ and μ_j is either 0 or 1. Note that the shortest high-level segment contains only two observations, while the longest contains 40 observations.

Suppose one uses the conjugate prior: $\sigma_j^2|m \sim$ scaled Inv-$\chi^2(\nu_0, \sigma_0^2)$, $\mu_j|(\sigma_j^2, m) \sim N(\mu_0, \sigma_j^2/\kappa_0)$. Then, the four hyperparameters $\mu_0, \sigma_0^2, \kappa_0$, and ν_0 will affect the final estimation. To study their impact, we first fix $\kappa_0 = 1/2$, $\nu_0 = 3$, and $\mu_0 = \bar{\boldsymbol{x}}$, where $\bar{\boldsymbol{x}}$ denotes the sample average, and then take $\sigma_0^2 = l_0\hat{\sigma}^2$, where $\hat{\sigma}^2$ represents the sample variance. We let l_0 vary, which changes the spread and the relative strength of the prior. Figure 2 shows the estimated change-points and the corresponding signal for 15 different values of l_0. As revealed in Figure 2, the estimated number of change-points decreases as the value of l_0 increases, that is, as the prior becomes weaker. When the value of l_0 is smaller than 1, although the five high-level segments are correctly identified, the estimated step function contains many false spikes since the strong priors tend to direct the estimator to treat any observation with a large deviation from the main sequence as a separate segment. When the value of l_0 is between 1 and 5, the estimators match the truth well. As the value of l_0 grows beyond 5, the estimator starts to miss the high-level segments. The segment with the shortest length is the first to be missed, since it contains least information; the longest high-level segment is more resilient to the change of priors and is missed only under extremely high values of l_0.

In our next investigation, instead of changing the spread of the prior, we simply shift it. We fix $\kappa_0 = 1/2$, $\nu_0 = 3$, and $\sigma_0^2 = \hat{\sigma}^2$ but let μ_0 take 15 different values, as shown in Figure 3. It is seen that when the value of μ_0 is close to the sample average—between -1.5 and 2—the estimated step functions match the truth well. However, as the value of μ_0 shifts away from the sample average, the relative strength of prior grows weaker and the estimator starts to miss the high-level segments. Similar to the previous picture, the long segments are more resilient.

In summary, the behavior of the maximum marginal likelihood estimator depends on the relative strength of the prior to the data. A relatively strong prior tends to overfit the data, yielding too many change-points, while a relatively weak prior tends to underfit the data, missing the real change-points. Therefore, to ensure a proper performance of the maximum marginal likelihood estimator, great care should be taken in choosing the hyperparameters.

For the choice of hyperparameters, Fearnhead (2005) suggested that hyperparameters can be chosen so that the summary statistics based on the prior distribution match the statistics based on the posterior of the preliminary study. However, this approach would require a number of iterations before the summary statistics converge, leading to a significant increase of the computational cost. It is also suggested in the literatures that the hyperparameters are chosen based on expert knowledge so that the prior can be relatively consistent with the data (Chib 1998; Fearnhead 2005, 2006). However, it is often ambiguous on how we should set the prior according to the expert knowledge, and such knowledge may not always be available in practice.

We recommend using an empirical Bayes approach to set the hyperparameters so that the prior could carry appropriate information to effectively function. Since the estimation is relatively robust to the choice of prior within a reasonably wide range, as shown in Figures 2 and 3, there is some flexibility in choosing a good prior. Furthermore, Corollary 3.1 guarantees the asymptotic consistency of the empirical Bayes maximum marginal likelihood estimator. In particular, the following guidelines can be used to choose the hyperparameters:

1. Derive the expectation and variance of a single observation as functions of α, the hyperparameter: $E(x|\alpha)$ and $\text{var}(x|\alpha)$.
2. Set the value of α so that $E(x|\alpha) = \hat{\mu}$, the sample average, and that $\text{var}(x|\alpha)$ is a large multiple of $\hat{\sigma}^2$, the sample variance.

Next, for normal and Poisson data, we recommend the following priors for practical data analysis. We found them worked well in our simulation (Section 5) and real data (Section 6) studies.

Normal data. For normal data $x_i|(\mu_j, \sigma_j^2) \sim N(\mu_j, \sigma_j^2)$, we use the conjugate prior: $\sigma_j^2|m \sim$ scaled Inv-$\chi^2(\nu_0, \sigma_0^2)$, $\mu_j|(\sigma_j^2, m) \sim N(\mu_0, \sigma_j^2/\kappa_0)$.

1. When the variability of the segment means μ_j is low or moderate (e.g., if it is known that the range of μ_j is moderate), we recommend two conjugate priors with hyperparameters:

$$\text{Norm-A}: \ \mu_0 = \bar{x}, \ \sigma_0^2 = \hat{\sigma}^2, \ \kappa_0 = \frac{1}{2}, \ \nu_0 = 3; \quad (4.1)$$

$$\text{Norm-B}: \ \mu_0 = \bar{x}, \ \sigma_0^2 = 2.5\hat{\sigma}^2, \ \kappa_0 = \frac{1}{2}, \ \nu_0 = 3. \quad (4.2)$$

Under the prior Norm-A, $E(x|\mu_0, \kappa_0, \nu_0, \sigma_0^2) = \bar{x}$ and $\text{var}(x|\mu_0, \kappa_0, \nu_0, \sigma_0^2) = 9\hat{\sigma}^2$. Under the prior Norm-B, $E(x|\mu_0, \kappa_0, \nu_0, \sigma_0^2) = \bar{x}$ and $\text{var}(x|\mu_0, \kappa_0, \nu_0, \sigma_0^2) = 22.5\hat{\sigma}^2$. The prior Norm-A (as will be shown in the following section) is good at locating short segments. However, it may overfit the data, giving too many small segments, especially when outliers are common. The Norm-B prior is a more conservative choice. In practice, it is recommended to apply the Norm-A prior first. If the resulting step function appears to be overfitting,

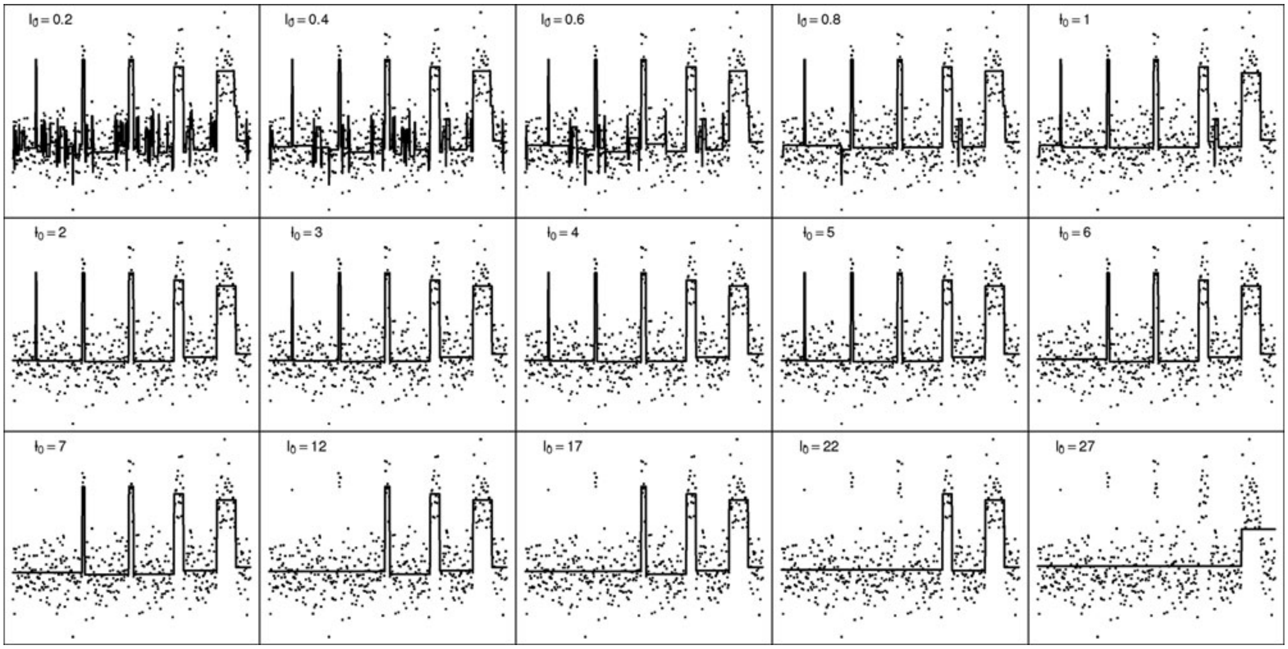

Figure 2. The data and change-points estimated by the maximum marginal likelihood method for 15 different values of l_0. From left to right, top to bottom, the values of l_0 are 0.2, 0.4, 0.6, 0.8, 1, 2, 3, 4, 5, 6, 7, 12, 17, 22, and 27, respectively.

the Norm-B prior can be applied to re-analyze the data. It must be noted, though, because different priors essentially reflect different prior knowledge regarding the nature of the data, the comparison between estimators under different priors often goes beyond pure statistical analysis and requires specific domain scientific knowledge.

2. When the variability of the segment means μ_j is large (e.g., if the range of μ_j is large), we recommend the following conjugate prior:

$$\text{Norm-C: } \mu_0 = \bar{x},\ \sigma_0^2 = \frac{3}{5}\hat{\tau}^2,\ \kappa_0 = \frac{5}{12}\frac{\hat{\tau}^2}{\hat{\sigma}^2},\ \nu_0 = 3, \tag{4.3}$$

where $\hat{\tau}^2$ is the average within-segment sample variance based on the change-point estimator obtained through the prior Norm-A (i.e., $\hat{\tau}^2$ is the average of $\hat{\sigma}_j^2$, where $\hat{\sigma}_j^2$ is the sample variance within the jth segment identified

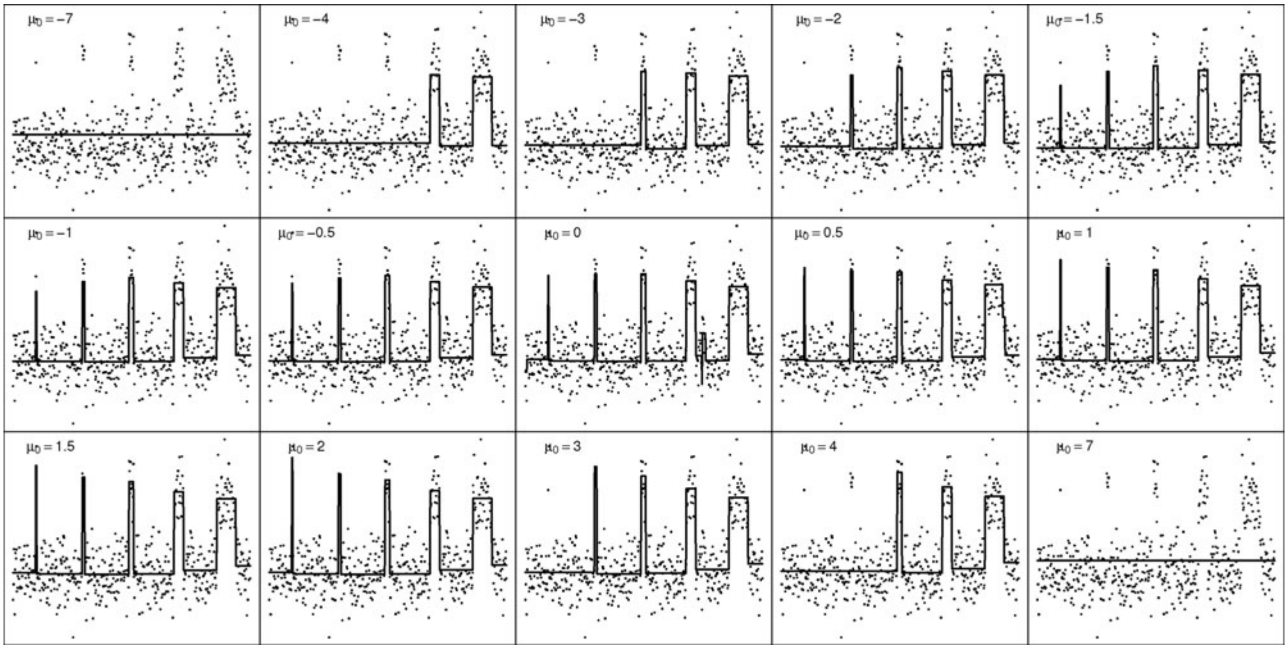

Figure 3. The data and change-points estimated by the maximum marginal likelihood method for 15 different values of μ_0. From left to right, top to bottom, the values of μ_0 are −7, −4 , −3, −2, −1.5, −1, −0.5, 0, 0.5, 1, 1.5, 2, 3, 4, and 7, respectively.

by first applying the prior Norm-A). The rationale behind this prior is that the total variance var(X) can be thought as the sum of var(μ_j) and $E(\sigma_j^2)$, the variance between segments and the variance within segments. In the construction of Norm-A and Norm-B priors, $E(\sigma_j^2|\alpha)$ is taken to be a multiple of $\hat{\sigma}^2$ ($3\hat{\sigma}^2$ and $7.5\hat{\sigma}^2$, respectively). Since $\hat{\sigma}^2$ measures the overall variance var(X) rather than the variance within segments, if var$(\mu_j) \gg E(\sigma_j^2)$ (e.g., when the range of μ_j is large), Norm-A and Norm-B priors then might match $E(\sigma_j^2|\alpha)$ to a considerable large value, resulting in a relatively weak prior which underestimates the number of change-points. Under the Norm-C prior, $E(\sigma_j^2|\alpha)$ is matched to $\hat{\tau}^2$ to avoid this problem, and we have $E(x|\mu_0,\kappa_0,\nu_0,\sigma_0^2) = \bar{x}$, var$(x|\mu_0,\kappa_0,\nu_0,\sigma_0^2) \approx 4\hat{\sigma}^2$, and $E(\sigma_j^2|\mu_0,\kappa_0,\nu_0,\sigma_0^2) \approx 2\hat{\tau}^2$.

Remark 2. Under the conjugate prior, which has density

$$\pi\left(\mu_j,\sigma_j^2|\mu_0,\kappa_0,\nu_0,\sigma_0^2\right) = \frac{\left(\sigma_0^2\nu_0/2\right)^{\nu_0/2}\left(\sigma_j^2\right)^{-(\nu_0/2+1)}}{\Gamma(\nu_0/2)\left(2\pi\sigma_j^2/\kappa_0\right)^{1/2}} \exp\left(-\frac{1}{2\sigma_j^2}\left(\kappa_0(\mu_j-\mu_0)^2+\nu_0\sigma_0^2\right)\right),$$

the marginal likelihood has a closed form

$$D\left(\boldsymbol{x}_{(\tau_{j-1},\tau_j]}|\mu_0,\kappa_0,\nu_0,\sigma_0^2\right) \propto \left(\sigma_0^2\nu_0\right)^{\nu_0/2}\frac{\Gamma\left(\frac{\nu_0+n_j}{2}\right)}{\Gamma\left(\nu_0/2\right)}\sqrt{\frac{\kappa_0}{\kappa_0+n_j}} \times\left(\nu_0\sigma_0^2+\sum x_i^2-\frac{1}{n_j}\left(\sum x_i\right)^2 + \frac{\kappa_0\left(\sum x_i-n_j\mu_0\right)^2}{n_j\left(\kappa_0+n_j\right)}\right)^{-(\nu_0+n_j)/2},$$

where all the sums are over the set $\{i : t_i \in (\tau_{j-1},\tau_j]\}$, and $n_j = n_{(\tau_{j-1},\tau_j]}$.

Poisson data. When the data consist of counts, such as fluorescence or photon counts from biophysical experiments, modeling them as Poisson, $x_i|\lambda_j \sim$ Poisson(λ_j), is more appropriate. We recommend a conjugate prior $\lambda_j|\alpha,\beta \sim \Gamma(\alpha,\beta)$ with hyperparameters:

$$\text{Pois-P:}\quad \alpha = \bar{x}\beta,\ \beta = \frac{1}{2\hat{\sigma}^2}. \tag{4.4}$$

With this prior we have $E(x|\alpha,\beta) = \bar{x}$ and var$(x|\alpha,\beta) = \bar{x}(1+2\hat{\sigma}^2)$. We found this prior to work well in our simulation and real data analysis.

Remark 3. Under the conjugate prior $\lambda_j|\alpha,\beta \sim \Gamma(\alpha,\beta)$, which has density $\pi(\lambda_j|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\lambda_j^{\alpha-1}e^{-\beta\lambda_j}$, the marginal likelihood has a closed form

$$D\left(\boldsymbol{x}_{(\tau_{j-1},\tau_j]}|\alpha,\beta\right) \propto \frac{\Gamma\left(\sum x_i+\alpha\right)}{\Gamma\left(\alpha\right)}\beta^\alpha/\left(n_j+\beta\right)^{\alpha+\sum x_i},$$

where the sums are over $\{i : t_i \in (\tau_{j-1},\tau_j]\}$, and $n_j = n_{(\tau_{j-1},\tau_j]}$.

5. SIMULATION STUDY

In this section, we carry out simulations to compare the maximum marginal likelihood estimator to six other estimators. We set up three testing scenarios to explore different patterns of stepwise signal. In each scenario, 1000 independent datasets are generated and the change-points and stepwise signal are estimated. Several criteria are then employed to assess the performance of different estimators. First, we briefly discuss the other six methods:

1. Fused lasso (Tibshirani et al. 2005; Tibshirani and Wang 2008). Under an equal-segment-variance assumption (i.e., $\sigma_1^2 = \sigma_2^2 = \cdots$), the change-points are estimated by minimizing the sum of squares with two constraints that penalize the L_1-norm of both the means and their successive differences. This method was implemented in the R package "cghFLasso" (*http://www-stat.stanford.edu/∼tibs/cghFLasso.html*). We will use the default setting and use "Lasso" to label this method.
2. The second method is based on Boysen et al. (2009). Under the equal-variance assumption, the change-points are estimated by minimizing a Potts functional, defined as the mean squared error plus a penalty term $\gamma_n J$, where γ_n is a function of sample size and J is the number of change-points. We adopt the recommended penalty term $\gamma_n = 2.5\log n$ and use "Potts-func" to label this method.
3. The third method is based on Yao (1988) and Braun, Braun, and Müller (2000). Yao (1988) discussed the change-point estimation for normal data with equal variance. Braun, Braun, and Müller (2000) generalized this method to cases where the variance can be expressed as a product of an overdispersion parameter σ^2 and a known function of means. The change-points are estimated by minimizing $n\log\hat{\sigma}_R^2 + RC_n$, where $\hat{\sigma}_R^2$ is the MLE of σ^2, R is the number of change-points, and C_n is a function of the sample size. We will use the recommended formulas of C_n. For normal data, $C_n = 0.5\log n$ (Yao 1988). For Poisson data, $C_n = n^\alpha$ (Braun, Braun, and Müller 2000), and $\alpha = 0.42$ (based on cross-validation). We will use "quasi-lik" to label this method.
4. The fourth method is based on Zhang and Siegmund (2007), in which the change-points are estimated using a modified BIC procedure. The penalty function that adds to the likelihood function is of the form $-\frac{3}{2}(m-1)\log n - \frac{1}{2}\sum_{i=1}^m \log(n_i/n)$, where m is the number of segments and n_i is the length of the ith segment. The assumption used in this article is that the data are normally distributed. We will use "mBIC" to label this method.
5. The fifth method is based on Frick, Munk, and Sieling (2014), where the change-points are estimated by minimizing the number of change-points over the acceptance region of a multiscale test. This method was implemented in the R package "stepR" (*http://www.stochastik.math.uni-goettingen.de/index.php?id=189*). Distribution families implemented in this package include Poisson and normal distributions with constant variance. In the simulation, we adopt the default setting for Scenario I, and apply a less conservative setting by choosing significant level $\alpha = 0.9$ for Scenarios II and III. We will use "SMUCE" to label this method.

6. The sixth method is named circular binary segmentation (CBS), proposed in Olshen et al. (2004) and Venkatraman and Olshen (2007). In contrast to the binary segmentation method, CBS can detect a small changed segment buried in the middle of a large segment using a likelihood ratio test. As the corresponding p-value is determined based on the permutation reference distribution, this method does not require any specific distributional assumption. This method was implemented in the R package "PSCBS" (*http://cran.r-project.org/web/packages/PSCBS/index.html*). We will use the default setting and use "CBS" to label this method.

Note that the equal-variance assumption is needed to establish the asymptotic consistency of the first four aforementioned estimators, and no asymptotic result is available for the sixth method. In contrast, there is no such restriction for us to establish the asymptotic properties of the maximum marginal likelihood estimator.

When we compare the estimated change-point sequence with the true sequence, a common Euclidean metric cannot be used since there is no one-to-one correspondence between each estimated change-point and the true one. This fact makes the evaluation of change-point estimators challenging. To the best of our knowledge, no single distance metric can provide a satisfactory result. Without a proper metric, the variability of estimated change-point sequences is not defined either. Thus, we will use the following three criteria in which the discrepancies between the estimated and the true sequences of change-points are examined from different angles.

- Criterion I: The difference between the estimated number of change-points and the true number of change-points.
- Criterion II: The frequency of correctly identifying certain segment of interest, or the overall proportion of segments correctly identified by the change-point estimator. For a segment to be considered correctly identified, the two change-points that define a given segment need to be exactly estimated with no other change-point estimated in between.
- Criterion III: The distance between the estimated change-points and the true change-points:
  (A) The distance from a true change-point to the estimated set of change-points.
  (B) The distance from an estimated change-point to the true set of change-points.

Criterion I is straightforward: The closer the number of estimated change-points to the truth, the better. The distance in Criterion III (A) can be thought of as a measure of the false negative rate, or underfitting of the model. A short distance in III (A) would suggest that the true change-point is roughly contained in the estimated change-point set, while a large distance is a sign that the true change-point is not detected by the estimator. Similarly, the distance in Criterion III (B) is a measure of the false positive rate, or overfitting of the model. A short distance in III (B) would suggest that the estimated change-point is close to one of the true change-points, while a large distance suggests that the estimated change-point is simply an overfit.

5.1 Scenario I: Stepwise Signal With Fixed Change-Points

The first scenario we explore is borrowed from Lai et al. (2005) in which 13 different algorithms used in analyzing array CGH data were evaluated. Each simulated dataset contains 500 indexed observations, divided between alternating "normal" and "abnormal" regions. The signal in the "abnormal" regions is higher than that in the "normal" regions. The five abnormal segments are at indexes 49-50, 147-151, 245-254, 340-359, and 430-469. The lengths of the abnormal segments are 2, 5, 10, 20, and 40, respectively, so we could study the performance of different estimators on detecting segments with different lengths. Figure 1 shows one such dataset along with the step function. For the data distribution, we consider three different settings:

1. Normal distribution with equal variance (EV).
   This is the original assumption used in Lai et al. (2005). Observations follow $N(0, 0.25^2)$ in normal regions and $N(1, 0.25^2)$ in abnormal regions.
2. Normal distribution with unequal variance (UEV).
   Observations follow $N(0, 0.25^2)$ in normal regions and $N(1.5, 0.5^2)$ in the abnormal regions so that a high-level signal is associated with large noise.
3. Poisson distribution.
   Observations follow Pois(25) in normal regions and Pois(50) in abnormal regions.

A total of 1000 independent datasets are generated under each distributional assumption. Change-points are estimated using our maximum marginal likelihood estimator (employing both the Norm-A prior and the Norm-B prior for normal data, and the Pois-P Prior for Poisson data) and the six methods previously described. The results are tabulated according to the three criteria. Table 1 shows the mean and the standard deviation (in parentheses) of the difference between the estimated and the true numbers of change-points. The frequencies of correctly identifying each abnormal segment are listed in Table 2. The average distances based on Criterion III (A) (B) are summarized in Tables 3 and 4, respectively.

As indicated by Tables 1 and 4, the fused lasso method tends to overestimate the number of change-points under all the three data distributions. In addition, the performance of the fused lasso method significantly deteriorates when the equal-variance assumption does not hold. The Potts functional method performs well under the equal-variance normal case and the Poisson case, but tends to overestimate the number of change-points under the unequal-variance normal case. The quasi-likelihood and mBIC methods both work well under the equal-variance normal and Poisson cases, but with smaller chances to detect the shortest abnormal segment with length 2 than the Potts functional method (Table 2), and both tend to underestimate the number of change-points (Tables 1 and 3). Under the unequal-variance case, the performance of the quasi-likelihood method is quite good and better than the mBIC method. For the SMUCE estimator, its power of detecting the shortest abnormal segment is even weaker than the quasi-likelihood and mBIC estimators, but still stronger than the CBS estimator which misses the shortest abnormal segment most of the time (Table 2). Otherwise, the

Table 1. The mean and standard deviation of the difference between the estimated number of change-points and the true number of change-points under Scenario I

	Norm-A /Pois-P	Norm-B	Lasso	Potts-func	Quasi-lik	mBIC	SMUCE	CBS
Normal	6.01	−0.04	22.20	−0.32	−0.58	−0.70	−1.17	−1.77
(EV)	(3.48)	(1.21)	(5.10)	(0.96)	(0.98)	(1.00)	(0.73)	(0.67)
Normal	2.73	0.04	22.13	48.38	1.51	3.19	0.24	−1.89
(UEV)	(2.02)	(1.1)	(5.77)	(79.76)	(1.94)	(2.56)	(1.02)	(0.74)
Poisson	1.35		93.11	2.91	−0.35	0.16	−0.86	−1.82
	(1.56)		(6.94)	(22.21)	(0.83)	(1.15)	(0.70)	(0.68)

performance of both the SMUCE and the CBS estimators is quite good.

Under Poisson data, the performance of our method is comparable to the quasi-likelihood and mBIC methods and better than the Potts functional, SMUCE and CBS methods, especially with regard to identifying the shortest segment with length 2. Under Poisson data, our method slightly inclines to overfit the data, while the quasi-likelihood, SMUCE and CBS methods tend to underfit, as suggested by Tables 1, 3, and 4. Under the normal cases, the Norm-B prior is more conservative than the Norm-A prior and can be expected to be more effective for signals with long segments and few change-points. This is supported by the results summarized in Tables 1, 3, and 4, where the Norm-B prior often gives the best results, especially with unequal-variance data. However, the Norm-B prior is not as good as the Norm-A prior in detecting the shortest abnormal segment with length 2 (Table 2) and tends to underfit the data (Table 3). The power of the Norm-A prior lies in detecting short segments, which makes it an option worthy of consideration when the false negative may bring undesirable consequences. The performance of the Norm-A prior is also comparable or better than the other methods under unequal-variance normal data.

It can then be concluded that, overall speaking, our method based on the Norm-B prior is the most effective method under this scenario—step signal with fixed change-points. The Norm-A and Pois-P priors are most effective in identifying short segments. Finally, our method (with both Norm-A and Norm-B

Table 2. The frequency of correctly identifying each abnormal segment, indexed by the segment length (L) under Scenario I

	L	Norm-A /Pois-P	Norm-B	Lasso	Potts-func	Quasi-lik	mBIC	SMUCE	CBS
	2	86.1%	62.5%	7.5%	69.6%	62.1%	57.8%	16.8%	0.6%
Normal	5	89.0%	87.7%	88.2%	90.3%	90.2%	90.3%	86.8%	90.4%
	10	88.3%	88.0%	66.0%	90.4%	90.5%	90.5%	90.5%	90.7%
(EV)	20	87.7%	90.5%	26.4%	90.9%	91.3%	91.2%	91.6%	90.7%
	40	83.2%	87.6%	7.1%	83.2%	88.7%	89.0%	89.4%	89.0%
	2	86.9%	74.3%	28.1%	70.0%	78.7%	79.8%	66.5%	2.1%
Normal	5	80.2%	88.9%	78.6%	56.6%	79.5%	72.9%	83.9%	81.1%
	10	74.2%	88.2%	37.9%	47.0%	70.8%	59.9%	82.7%	83.9%
(UEV)	20	66.1%	83.9%	10.4%	35.5%	59.7%	45.5%	74.3%	82.6%
	40	58.4%	80.1%	1.5%	23.7%	46.2%	30.7%	56.9%	79.0%
	2	87.8%		84.9%	81.3%	72.7%	79.3%	30.3%	2.1%
	5	90.6%		20.1%	87.6%	92.9%	88.6%	92.6%	90.3%
Poisson	10	86.4%		2.3%	84.4%	90.3%	85.2%	90.4%	90.6%
	20	87.6%		0.0%	81.7%	92.0%	83.0%	92.2%	89.8%
	40	86.4%		0.0%	76.1%	90.5%	80.3%	91.0%	88.1%

Table 3. The mean and standard deviation of the distance from the true change-point to the estimated set of change-points under Scenario I

	Norm-A /Pois-P	Norm-B	Lasso	Potts-func	Quasi-lik	mBIC	SMUCE	CBS
Normal	1.06	5.89	1.55	4.83	6.37	7.27	6.67	17.14
(EV)	(4.09)	(20.98)	(5.68)	(18.87)	(22.10)	(23.74)	(22.45)	(36.02)
Normal	1.22	4.35	1.66	1.74	2.86	2.25	2.34	19.12
(UEV)	(5.70)	(17.7)	(7.23)	(9.49)	(13.51)	(11.26)	(11.44)	(39.49)
Poisson	1.31		0.63	2.79	4.35	3.32	4.29	17.75
	(6.19)		(0.48)	(13.27)	(17.77)	(15.00)	(17.34)	(37.12)

Table 4. The mean and standard deviation of the distance from the estimated change-point to the set of true change-points under Scenario I

	Norm-A /Pois-P	Norm-B	Lasso	Potts-func	Quasi-lik	mBIC	SMUCE	CBS
Normal	5.81	1.24	9.07	1.07	0.98	0.95	0.94	1.06
(EV)	(10.70)	(3.45)	(11.70)	(2.11)	(1.40)	(1.16)	(1.40)	(1.95)
Normal	1.89	1.04	7.22	14.29	1.83	2.44	1.39	1.10
(UEV)	(4.64)	(1.90)	(9.97)	(13.38)	(3.12)	(3.90)	(2.55)	(2.12)
Poisson	2.56		16.13	4.53	0.98	1.15	0.93	1.11
	(6.68)		(13.59)	(9.10)	(1.50)	(1.95)	(1.40)	(2.44)

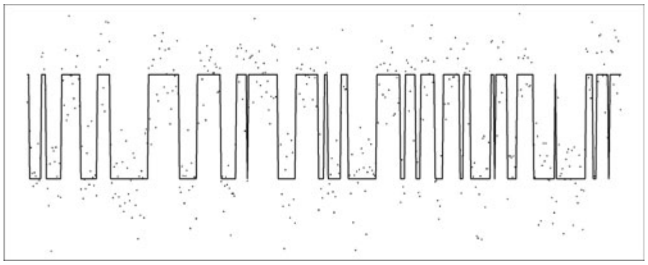

Figure 4. One realization of the step function and the observations, Scenario II.

priors) is able to handle the unequal-variance case better than the other methods.

5.2 Scenario II: Stepwise Signal With Randomized Markov Change-Points

The scenario explored in this section is inspired by the enzymatic cycle of a single enzyme molecule, in which the enzyme switches between different conformations. The fluorescence marker on the enzyme molecule releases a high-intensity photon stream when the enzyme is in one conformation but releases fewer photons when the enzyme is in another conformation. This system is often modeled as a two-state Markov chain (Lu, Xun, and Xie 1998), but more complex patterns have been discovered and studied as well (English et al. 2006; Kou 2008; Du and Kou 2012).

To emulate such systems, we employ a two-state discrete-time Markov chain to simulate the change-point sequence. The probabilities of staying in state 1 and 2 are 0.9 and 0.95, respectively, so the mean sojourn times (i.e., the average length) for states 1 and 2 are $10 = 1/(1 - 0.9)$ and $20 = 1/(1 - 0.95)$, respectively. The starting state is drawn from the stationary distribution.

In each simulation run, we first simulate the change-points according to the two-state Markov process and then generate 500 observations on top according to each of the three sets of distributional assumptions described in Scenario I. The average number of change-points is around 30. As a result, there are more short segments, and the inference is thus harder than that of Scenario I. Figure 4 shows a realization of such data with the mean function plotted as a solid line.

The estimation results are summarized in Tables 5–8. Note in this scenario all the change-points are random so we list the overall proportion of correctly identified segments in Table 6.

Table 5 shows that the Potts functional method performs quite well under equal-variance normal case, but the overestimation biases are huge in other cases. The fused lasso method also overestimates the number of change-points under the equal-variance normal case and Poisson case and is not accurate in identifying the segments (Table 6). The quasi-likelihood, mBIC, SMUCE, and CBS methods work better for the unequal-variance data than the lasso method, but all these four methods tend to ignore short segments and thus underestimate the number of change-points (Tables 5 and 7).

Owing to the existence of many short segments in the stepwise signal, our method with the conservative Norm-B prior also tends to underestimate the number of change-points (Table 5). Yet the results obtained through the Norm-B prior are still reasonably good compared to the other approaches: Under the unequal-variance case, our estimator using the Norm-B prior is only clearly outperformed by the mBIC estimator and is comparable or better than other estimators; under the equal-variance case, our estimator using the Norm-B prior is not as good as the Potts functional, quasi-likelihood, and mBIC estimators but comparable to the SMUCE estimator and better than the fused lasso and CBS estimators. Our estimator based on the Norm-A prior dominates all other methods under the unequal-variance case and is comparable to the Potts functional method under the equal-variance case. The performance of our method with the Pois-P prior is also significantly better than the other approaches for Poisson data. Thus, simulations in this scenario again point out the effectiveness of our method (with the Norm-A and Pois-P priors) in analyzing a stepwise signal with short segments and unequal variance.

Based on the discussion in Scenarios I and II, it can be seen that while the Norm-A prior is more sensitive to short segments, it is also sensitive to the extreme values found in the long segments. The Norm-B prior, on the other hand, is more robust and yields more conservative outcomes. We suggest both priors be used in practice and that comparison with domain scientific knowledge can then be made to make a final choice.

5.3 Scenario III: Stepwise Signal With Many Levels

In this scenario, we explore a setting where the variance of the segment means dominates the within-segment variance: $\text{var}(\mu_j) \gg E(\sigma_j^2)$. This scenario is partly inspired by the linear stepwise movement of a molecular motor along a microtubule (Yildiz et al. 2004; Nan, Sims, and Xie 2008). A molecular motor is a biomolecule that carries cargo loading back and forth in (and out of) a cell. A fluorescence marker attached to the motor molecule can be used to track its trajectory along the microtubule. The plot of the molecular motor's movement

Table 5. The mean and standard deviation of the difference between the estimated number of change-points and the true number of change-points under Scenario II

	Norm-A /Pois-P	Norm-B	Lasso	Potts-func	Quasi-lik	mBIC	SMUCE	CBS
Normal	0.39	−8.40	9.51	−0.30	−6.81	−6.10	−8.35	−12.30
(EV)	(3.11)	(4.46)	(11.50)	(7.17)	(4.29)	(3.40)	(4.08)	(5.82)
Normal	5.36	−8.91	2.01	161.63	−9.11	−6.42	−8.28	−12.53
(UEV)	(5.28)	(4.68)	(12.94)	(26.58)	(5.14)	(4.08)	(4.23)	(5.69)
Poisson	−0.49		114.11	143.97	−6.35	−6.58	−6.26	−11.82
	(2.59)		(8.61)	(55.60)	(4.21)	(4.15)	(3.22)	(5.40)

Table 6. The overall proportion of correctly identified segments under Scenario II

	Norm-A /Pois-P	Norm-B	Lasso	Potts-func	Quasi-lik	mBIC	SMUCE	CBS
Normal (EV)	77.6%	58.9%	27.1%	75.5%	64.4%	66.8%	51.8%	45.6%
Normal (UEV)	67.3%	51.7%	23.8%	29.4%	52.5%	59.0%	45.5%	42.4%
Poisson	79.7%		23.3%	28.8%	67.4%	64.6%	61.4%	47.2%

against time follows a stairwise pattern, where the length of a segment represents the waiting time in a particular location.

To emulate such a system, we establish a fixed step function with six different levels and the total number of change-points is 16. Figure 5 shows a realization of such data together with the step function. We employ three distributions for testing:

1. Normal distribution with equal variance (EV).
   Observations follow normal distributions with means 1, 2, 3, 4, 5, and 6 for the six different levels and a common variance 0.25^2.
2. Normal distribution with unequal variance (UEV).
   Observations follow normal distributions with means 1.5, 3, 4.5, 6, 7.5, and 9 for the six different levels. The variances are 0.25^2 and 0.5^2, alternating.
3. Poisson distribution.
   Observations follow Poisson distributions with means 25, 50, 75, 100 , 125, and 150 for the six levels.

As discussed in Section 4, given that var$(\mu_j) \gg E(\sigma_j^2)$, we will use the Norm-C prior for the normal data. For the Poisson data, the Pois-P prior is still applicable. The simulation and estimation results are summarized in Tables 9–12.

Table 7. The mean and standard deviation of the distance from the true change-point to the estimated set of change-points under Scenario II

	Norm-A /Pois-P	Norm-B	Lasso	Potts-func	Quasi-lik	mBIC	SMUCE	CBS
Normal	0.70	3.16	2.10	0.84	2.29	2.03	2.00	5.67
(EV)	(3.54)	(7.76)	(4.84)	(4.06)	(6.62)	(6.25)	(5.22)	(11.96)
Normal	0.75	3.72	3.58	0.09	3.45	2.39	2.09	5.82
(UEV)	(3.42)	(8.46)	(6.99)	(1.43)	(8.27)	(6.60)	(5.01)	(11.60)
Poisson	0.67		0.03	0.18	2.17	2.28	1.36	5.11
	(3.49)		(0.20)	(2.01)	(6.53)	(6.60)	(4.19)	(10.86)

Table 8. The mean and standard deviation of the distance from the estimated change-point to the set of true change-points under Scenario II

	Norm-A /Pois-P	Norm-B	Lasso	Potts-func	Quasi-lik	mBIC	SMUCE	CBS
Normal	0.40	0.08	2.54	0.56	0.08	0.08	0.22	0.13
(EV)	(2.36)	(0.48)	(6.12)	(2.47)	(0.62)	(0.49)	(1.14)	(0.82)
Normal	1.59	0.21	2.59	7.89	0.16	0.26	0.57	0.18
(UEV)	(5.35)	(1.57)	(6.10)	(9.30)	(0.84)	(1.44)	(2.74)	(0.93)
Poisson	0.29		6.68	7.18	0.07	0.11	0.19	0.12
	(1.59)		(8.76)	(8.71)	(0.53)	(0.98)	(1.19)	(0.80)

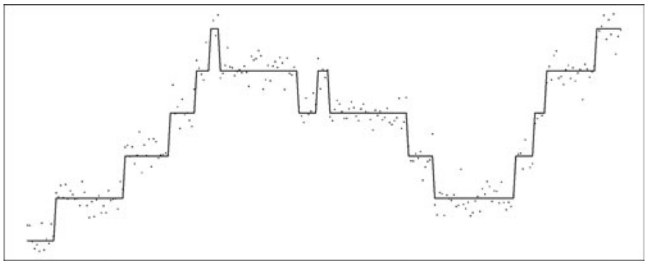

Figure 5. One simulation realization and the stepwise mean function, Scenario III.

Based on the numerical results, both the fused lasso and Potts functional methods significantly overestimate the number of change-points (Tables 9, 10, and 12). On the other hand, the estimators based on quasi-likelihood, mBIC, SMUCE, CBS and our method yield much better results under all criteria for all three data distributions. These five different estimators exhibit roughly similar performances. Among these five, the quasi-likelihood and mBIC estimators hold a slight edge over the other three estimators for equal-variance normal data and Poisson data; under unequal-variance normal data, the quasi-likelihood estimator and our estimators show better results than the others.

In summary, it appears that our marginal likelihood method is the most versatile among all the methods tested here. Its performance is at least comparable to the best of the other methods under the scenarios and data distributions tested. In addition, our method has a considerable advantage when the variances vary and can be good at detecting short segments with an appropriate prior setting. Finally, our method is adaptable in the sense that it is essentially an empirical Bayes method that self-adjusts to the data and that the users can choose the appropriate prior based on the domain knowledge and the context. In the next section, we will apply our method to two real datasets.

6. ANALYZING REAL DATA

6.1 Array CGH Data

Locating the aberration regions in a DNA sequence is important for understanding the pathogenesis of cancer and many other diseases. Array CGH is a technique developed for such a purpose. A typical array CGH sequence consists of the log-ratios of normalized intensities from disease versus control samples, indexed by the genome numbers. The regions of concentrated high or low log-ratios departing from 0 indicate amplification or loss of chromosomal segments. Thus, a key question in analyzing array CGH data is to detect those abnormal regions.

Here we will use our marginal likelihood method to study two samples of array CGH data analyzed in Lai et al. (2005) (*http://compbio.med.harvard.edu/Supplements/Bioinformatics 05b.html*). The data are normalized from the raw data from Bredel et al. (2005), which concerns primary glioblastoma multiforme (GBM), a malignant type of brain tumor. In particular, the two samples represent chromosome 7 in GBM29 from 40 to 65 Mb and chromosome 13 in GBM31. We apply both Norm-A and Norm-B priors to analyze these two samples. The estimated step functions along with the CGH data are shown in Figure 6 for sample GBM29 and in Figure 7 for sample GBM31.

In sample GBM29, three regions of high-amplitude amplifications exist and have been well studied. Based on Figure 6, both estimators successfully identify all three high amplifications even though the first two regions are separated only by four probes. In sample GBM31, a large region of low-magnitude loss exists, as indicated by comparing the estimated signal with the dashed reference line in Figure 7. Both estimators pick up spikes with unusual log-ratios. In either case, our result provides solid evidence that the magnitudes of the signal are lower than the

Table 9. The mean and standard deviation of the difference between the estimated number of change-points and the true number of change-points under Scenario III

	Norm-C /Pois-P	Lasso	Potts-func	Quasi-lik	mBIC	SMUCE	CBS
Normal	−0.23	16.67	68.3	0.03	0.06	−0.52	0.13
(EV)	(0.59)	(3.77)	(32.50)	(0.29)	(0.30)	(0.70)	(0.60)
Normal	0.63	18.26	83.76	0.22	1.00	0.42	−0.09
(UEV)	(0.97)	(4.06)	(4.42)	(0.77)	(1.36)	(1.19)	(0.79)
Poisson	−2.14	107.8	79.7	−1.45	−0.99	−3.31	−2.16
	(1.34)	(6.88)	(18.38)	(1.50)	(1.58)	(0.78)	(1.47)

Table 10. The overall proportion of correctly identified segments under Scenario III

	Norm-C /Pois-P	Lasso	Potts-func	Quasi-lik	mBIC	SMUCE	CBS
Normal (EV)	87.2%	20.7%	21.8%	89.5%	89.4%	81.2%	84.1%
Normal (UEV)	84.8%	17.1%	12.0%	83.5%	81.4%	76.9%	80.9%
Poisson	50.2%	1.0%	8.2%	52.7%	53.5%	38.6%	45.9%

Table 11. The mean and standard deviation of the distance from the true change-point to the estimated set of change-points under Scenario III

	Norm-C /Pois-P	Lasso	Potts-func	Quasi-lik	mBIC	SMUCE	CBS
Normal	0.16	1.20	0.03	0.07	0.06	0.24	0.12
(EV)	(0.91)	(2.67)	(0.17)	(0.37)	(0.32)	(0.80)	(0.54)
Normal	0.08	1.13	0.03	0.13	0.11	0.20	0.21
(UEV)	(0.36)	(2.54)	(0.16)	(0.65)	(0.48)	(0.70)	(0.92)
Poisson	1.42	0.05	0.16	1.12	0.95	1.68	1.50
	(3.02)	(0.23)	(0.46)	(2.61)	(2.36)	(2.82)	(3.06)

Table 12. The mean and standard deviation of the distance from the estimated change-point to the set of true change-points under Scenario III

	Norm-C /Pois-P	Lasso	Potts-func	Quasi-lik	mBIC	SMUCE	CBS
Normal	0.06	2.30	5.09	0.07	0.08	0.12	0.19
(EV)	(0.27)	(3.20)	(4.65)	(0.45)	(0.45)	(0.52)	(0.98)
Normal	0.27	2.34	5.48	0.23	0.48	0.56	0.20
(UEV)	(1.35)	(3.19)	(4.81)	(1.22)	(1.92)	(2.23)	(0.97)
Poisson	0.37	5.32	5.14	0.43	0.45	0.55	0.49
	1.06	(4.58)	(4.58)	(1.30)	(1.29)	(1.45)	(1.33)

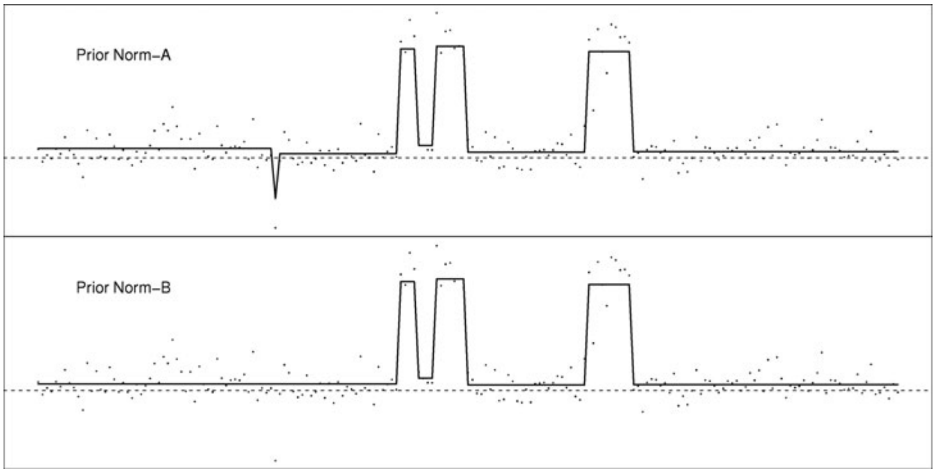

Figure 6. Array CGH data of GBM29, with the estimated step functions based on the Norm-A and Norm-B priors. A horizontal dashed line with intercept 0 is also plotted for reference.

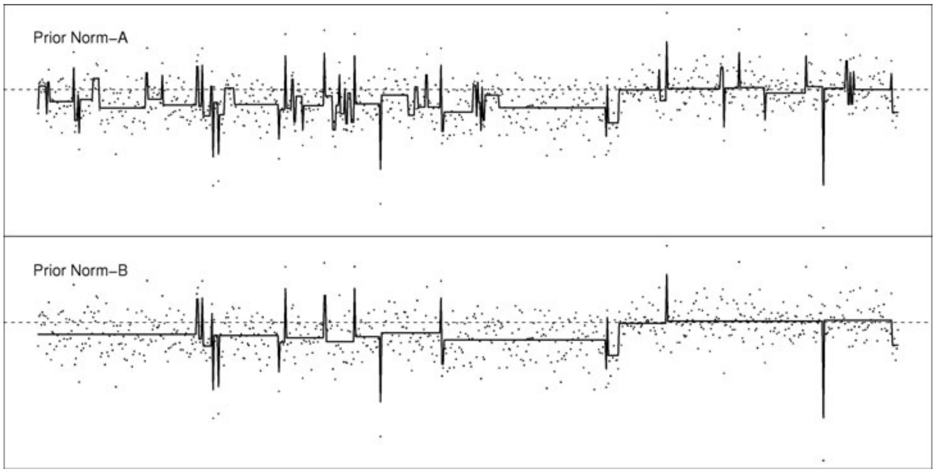

Figure 7. Array CGH data of GBM31, with the estimated step functions based on the Norm-A and Norm-B priors. A horizontal dashed line with intercept 0 is also plotted for reference.

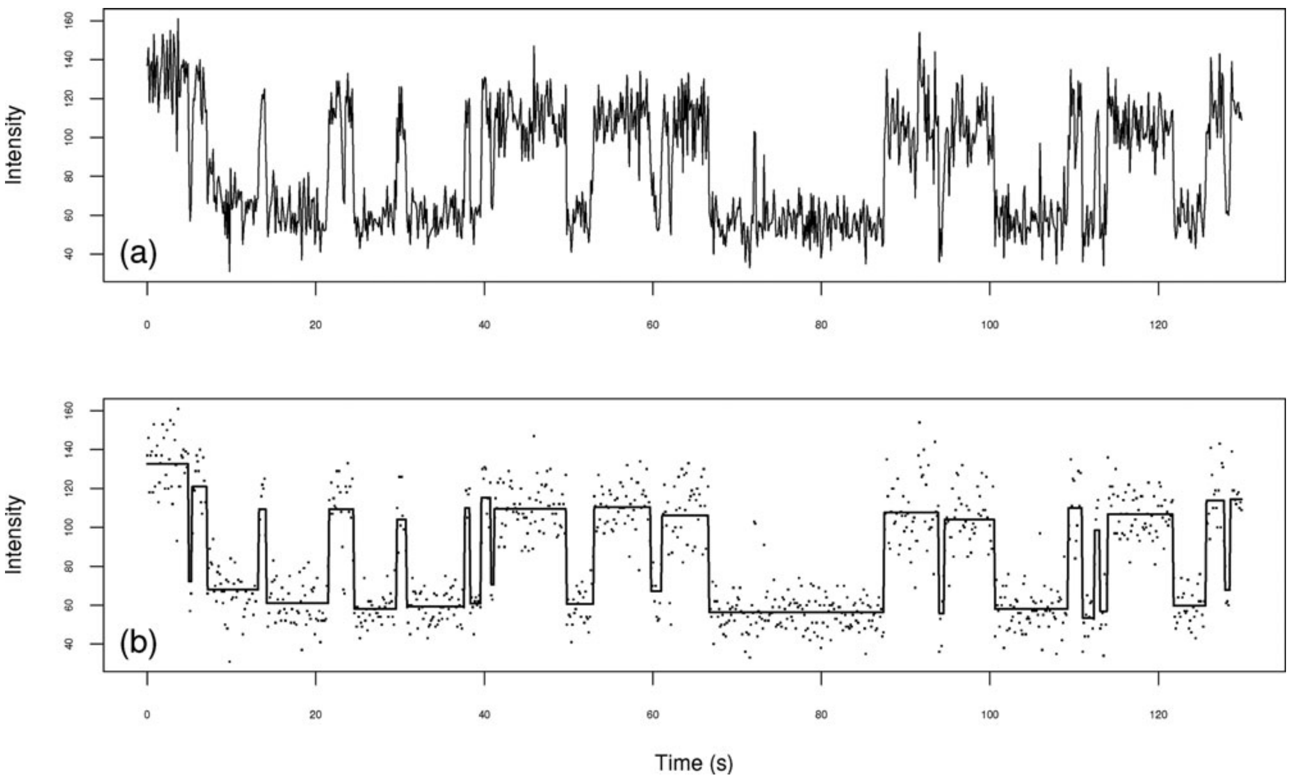

Figure 8. (a) The experimental trajectory of the fluorescence intensity of a single cholesterol oxidase. (b) The estimated step function based on the Norm-B prior.

reference line on the left two-thirds of the data sequence (except for occasional spikes). Furthermore, our estimators, especially the estimator based on the Norm-B prior, suggest that the magnitudes of loss are not constant within this region. For example, the magnitude of loss of the leftmost segment is clearly less than the magnitude of loss of the segment near the center.

Although our estimators based on different priors produce largely similar results, the discrepancy is apparent. However, it is often insufficient to determine which estimator is superior purely based on statistical grounds. For instance, in sample GBM 29, the difference of the estimators based on the two priors originates from a single-probe outlier. This outlier can either be a real aberration or the result of an experimental error, and specific scientific knowledge would be necessary to determine its nature.

It is also worth noting that, in Lai et al. (2005), 13 different CGH data analysis algorithms were also used to analyze those two samples, and our method is comparable to the best of those algorithms. Our method thus can prove to be a useful tool to detect aberrations in array CGH data.

6.2 Enzymatic Cycle of a Single Cholesterol Oxidase Molecule

A cholesterol oxidase is an enzyme that catalyzes the oxidation of cholesterol. The active site of the enzyme (E) binds a flavin adenine dinucleotide (FAD), which is naturally fluorescent, but the fluorescence is lost when FAD is reduced by cholesterol to FADH_2. The resulting complex E–FADH_2 will then be oxidized by O_2 and return to the fluorescent state E–FAD, starting the next cycle. This enzymatic cycle can be represented by the following diagram:

$$\text{E–FAD} \underset{k_{-1}}{\overset{k_1}{\rightleftarrows}} \text{E–FADH}_2, \tag{6.1}$$

where k_1 and k_{-1} represent the corresponding kinetic reaction rates, measured in sec^{-1}. This cycle is often modeled as a two-state continuous-time Markov chain in which k_1 and k_{-1} serve as the transition intensities (so that the dwell times in the states E–FAD and E–FADH_2 have exponential distributions with rates k_1 and k_{-1}, respectively).

New advances in nanosciences in the last two decades have opened the door for scientists to study such processes on a microscopic molecule-by-molecule basis (Nie and Zare 1997; Xie and Trautman 1998; Xie and Lu 1999; Tamarat et al. 2000; Weiss 2000; Moerner 2002; Kou 2009; Qian and Kou 2014). In such experiments, a single enzyme molecule is immobilized and its fluorescence intensity over time is recorded (Lu, Xun, and Xie 1998). Figure 8(a) shows the experimental fluorescence intensity trajectory of a single cholesterol oxidase.

In agreement with Equation (6.1), the observed trajectory clearly suggests the existence of two different states with high and low fluorescence intensities (corresponding to the two states E–FAD and E–FADH_2 respectively). The exact segmentation of this stepwise signal is unknown due to the noise but can be estimated with our method.

It is common to model the fluorescence intensity by a Poisson distribution. However, we found that the variance of the intensity in each state is much larger than the mean. As a result, the unequal-variance Gaussian assumption appears to be more appropriate, and we apply the conservative Norm-B prior due to the large noise. The estimated step function is shown in Figure 8(b).

Our estimate suggests that this signal can be divided into 33 segments: 17 segments with high intensities are associated with the state E–FAD, while the other 16 segments are associated with the state E–FADH$_2$. Based on this segmentation, k_1 and k_{-1} can be estimated as 0.279 ± 0.133 sec^{-1} and 0.231 ± 0.113 sec^{-1}, respectively.

If one assumes a two-state Markov process (Wang and Wolynes 1995), then the autocorrelation function of the fluorescence intensity trajectory for (6.1) is $\exp(-(k_1 + k_{-1})t)$. Thus, the sum $k_1 + k_{-1}$ can also be inferred from the empirical autocorrelation function (without segmenting the signal) *under* the Markov model. The best exponential fitting of the autocorrelation estimates $k_1 + k_{-1}$ to be 0.431, in good agreement with our estimate. The autocorrelation method, however, can only estimate their sum, not k_1 or k_{-1} individually. Furthermore, the estimation based on the autocorrelation strongly depends on the two-state Markov model. In contrast, our method does not require any assumption on the underlying mechanism and provides a direct segmentation of the stepwise signal. The information gained through our method can be used to test models and estimate the model parameters, offering a valuable insight to validate, modify, and improve existing models.

7. CONCLUSION

In this article, we formulated a maximum marginal likelihood estimator for stepwise signal estimation. We investigated the impact and the choice of prior, as well as the asymptotic properties of our estimator. We also carried out an extensive simulation study and applied this method to two real data problems.

Our analytical results show that, under mild conditions, our maximum marginal likelihood estimator of change-points is asymptotically consistent. In the finite-sample scenario, our investigation in Section 4 illustrated the importance of choosing an appropriate prior. In the simulation, our maximum marginal likelihood estimator coupled with an empirical Bayes choice of the hyperparameters was demonstrated to be competitive compared to the other methods, specially in the following cases:

(i) when the equal-variance assumption does not hold; and
(ii) when many short segments are present.

In addition, our method works well for two real data examples discussed in Section 6.

A stepwise signal appears in many applications in both natural and social sciences. In particular, in biology, chemistry, and biophysics, the fluorescence stepwise signal is often the main source researchers rely on to infer the time evolution of the underlying systems. Locating the change-points is often the first and key step in such quests. Much research has been devoted to create algorithms for this purpose. However, important questions are not thoroughly discussed, including how to configure and auto-adjust the algorithms so that they can work for a broad range of real-data problems, and how to judge the estimation outcome. We hope our discussion on these questions, along with our proposed method, will generate further interest in research along this direction. For example, an interesting question is to quantify the variability of change-point estimates.

The R and Matlab packages of our marginal likelihood method can be downloaded at *http://www.people.fas.harvard.edu/∼skou/publication.htm*.

8. SUPPLEMENTARY MATERIALS

The detailed proofs of Lemmas 3.1, 3.2, Theorem 3.1 and Corollary 3.1 are given in the online supplementary material.

[Received November 2013. Revised December 2014.]

REFERENCES

Auger, I. E., and Lawrence, C. E. (1989), “Algorithms for the Optimal Identification of Segment Neighborhoods,” *Bulletin of Mathematical Biology*, 51, 39–54. [315,316]

Bai, J., and Perron, P. (1998), “Estimating and Testing Linear Models With Multiple Structural Changes,” *Econometrica*, 66, 47–78. [314]

—— (2003), “Computation and Analysis of Multiple Structural Change Models,” *Journal of Applied Econometrics*, 18, 1–22. [314]

Barry, D., and Hartigan, J. (1993), “A Bayesian Analysis for Change-Point Problems,” *Journal of the American Statistical Association*, 88, 309–319. [315,316]

Bekaert, G., Harvey, C. R., and Lumsdaine, R. L. (2002), “Dating the Integration of World Equity Markets,” *Journal of Financial Economics*, 65, 203–247. [314]

Bellman, R., and Roth, R. (1969), “Curve Fitting by Segmented Straight Lines,” *Journal of the American Statistical Association*, 64, 1079–1084. [315,316]

Bement, T. R., and Waterman, M. S. (1977), “Locating Maximum Variance Segments in Sequential Data,” *Mathematical Geosciences*, 9, 55–61. [315,316]

Bhattacharya, P. K. (1994), “Some Aspects of Change-Point Analysis,” in *Change-Point Problems, IMS Monograph 23* , eds. E. Carlstein, H. Muller, and D. Siegmund, Hayward, CA: Institute of Mathematical Statistics, pp. 28–56. [314]

Boysen, L., Kempe, A., Liebscher, V., Munk, A., and Wittich, O. (2009), “Consistencies and Rates of Convergence of Jump-Penalized Least Squares Estimators,” *The Annals of Statistics*, 37, 157–183. [314,315,321]

Braun, J. V., Braun, R. K., and Müller, H.-G. (2000), “Multiple Changepoint Fitting via Quasilikelihood, With Application to DNA Sequence Segmentation,” *Biometrika*, 87, 301–314. [314,321]

Braun, J. V., and Müller, H.-G. (1998), “Statistical Methods for DNA Sequence Segmentation,” *Statistical Science*, 13, 142–162. [314]

Bredel, M., Bredel, C., Juric, D., Harsh, G. R., Vogel, H., Recht, L. D., and Sikic, B. I. (2005), “High-Resolution Genome-Wide Mapping of Genetic Alterations in Human Glial Brain Tumors,” *Cancer Research*, 65, 4088–4096. [326]

Carlin, B., Gelfand, A., and Smith, A. (1992), “Hierarchical Bayesian Analysis of Changepoint Problems,” *Applied Statistics*, 41, 2, 389–405. [314,316]

Chernoff, H., and Zacks, S. (1964), “Estimating the Current Mean of a Normal Distribution Which is Subjected to Changes in Time,” *The Annals of Mathematical Statistics*, 35, 999–1018. [314]

Chib, S. (1998), “Estimation and Comparison of Multiple Change-Point Models,” *Journal of Econometrics*, 86, 221–241. [315,316,319]

Davis, R. A., Lee, T. C. M., and Rodriguez-Yam, G. A. (2006), “Structural Break Estimation for Nonstationary Time Series Models,” *Journal of the American Statistical Association*, 101, 223–239. [315]

Du, C., and Kou, S. C. (2012), “Correlation Analysis of Enzymatic Reaction of a Single Protein Molecule,” *The Annals of Applied Statistics*, 6, 950–976. [324]

English, B. P., Min, W., van Oijen, A. M., Lee, K. T., Luo, G., Sun, H., Cherayil, B. J., Kou, S. C., and Xie, X. S. (2006), “Ever-Fluctuating Single Enzyme Molecules: Michaelis-Menten Equation Revisited,” *Nature Chemical Biology*, 2, 87–94. [314,324]

Fearnhead, P. (2005), “Exact Bayesian Curve Fitting and Signal Segmentation,” *IEEE Transactions on Signal Processing*, 53, 2160–2166. [315,316,319]

—— (2006), “Exact and Efficient Bayesian Inference for Multiple Changepoint Problems,” *Statistics and Computing*, 16, 203–213. [315,316,319]

Fearnhead, P., and Liu, Z. (2007), “On-line Inference for Multiple Changepoint Problems,” *Journal of the Royal Statistical Society,* Series B, 69, 589–605. [315]

Frick, K., Munk, A., and Sieling, H. (2014), “Multiscale Change-Point Inference,” *Journal of the Royal Statistical Society,* Series B, 76, 495–580. [314,318,321]

Fuh, C.-D. (2003), "SPRT and CUSUM in Hidden Markov Models," *Annals of Statistics*, 31, 942–977. [314]

—— (2004), "Asymptotic Operating Characteristics of an Optimal Change Point Detection in Hidden Markov Models," *Annals of Statistics*, 32, 2305–2339. [314]

Giordani, P., and Kohn, R. (2008), "Efficient Bayesian Inference for Multiple Change-Point and Mixture Innovation Models," *Journal of Business & Economic Statistics*, 26, 66–77. [315]

Green, P. (1995), "Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination," *Biometrika*, 82, 711–732. [315]

Hinkley, D. V. (1970), "Inference About the Change-Point in a Sequence of Random Variables," *Biometrika*, 57, 1–17. [314]

Jackson, B., Scargle, J. D., Barnes, D., Arabhi, S., Alt, A., Gioumousis, P., Gwin, E., Sangtrakulcharoen, P., Tan, L., and Tsai, T. T. (2005), "An Algorithm for Optimal Partitioning of Data on an Interval," *IEEE Signal Processing Letters*, 12, 105–108. [315,317]

Jarrett, R. G. (1979), "A Note on the Intervals Between Coal-Mining Disasters," *Biometrika*, 66, 191–193. [314]

Johnson, T. D., Elashoff, R. M., and Harkema, S. J. (2003), "A Bayesian Change-Point Analysis of Electromyographic Data: Detecting Muscle Activation Patterns and Associated Applications," *Biostatistics*, 4, 143–164. [314]

Killick, R., Fearnhead, P., and Eckley, I. A. (2012), "Optimal Detection of Changepoints With a Linear Computational Cost," *Journal of the American Statistical Association*, 107, 1590–1598. [315,317]

Koop, G., and Potter, S. M. (2007), "Estimation and Forecasting in Models With Multiple Breaks," *The Review of Economic Studies*, 74, 763–789. [315,316]

—— (2009), "Prior Elicitation in Multiple Change-Point Models," *International Economic Review*, 50, 751–772. [316]

Kou, S. C. (2008), "Stochastic Networks in Nanoscale Biophysics: Modeling Enzymatic Reaction of a Single Protein," *Journal of the American Statistical Association*, 103, 961–975. [314,324]

—— (2009), "A Selective View of Stochastic Inference and Modeling Problems in Nanoscale Biophysics," *Science in China,* A, 52, 1181–1211. [328]

Lai, T., and Xing, H. (2011), "A Simple Bayesian Approach to Multiple Change-Points," *Statistica Sinica*, 21, 539–569. [315,316]

Lai, W. R., Johnson, M. D., Kucherlapati, R., and Park, P. J. (2005), "Comparative Analysis of Algorithms for Identifying Amplifications and Deletions in Array CGH Data," *Bioinformatics*, 21, 3763–3770. [314,322,326,328]

Lu, H. P., Xun, L., and Xie, X. S. (1998), "Single-Molecule Enzymatic Dynamics," *Science*, 282, 1877–1882. [314,324,328]

Moerner, W. E. (2002), "A Dozen Years of Single-Molecule Spectroscopy in Physics, Chemistry, and Biophysics," *The Journal of Physical Chemistry*, B, 106, 910–927. [328]

Moreno, E., Javier Girón, F., and García-Ferrer, A. (2013), "A Consistent On-line Bayesian Procedure for Detecting Change Points," *Environmetrics*, 24, 342–356. [316]

Nan, X., Sims, P. A., and Xie, X. S. (2008), "Organelle Tracking in a Living Cell With Microsecond Time Resolution and Nanometer Spatial Precision," *ChemPhysChem*, 9, 707–712. [324]

Nie, S., and Zare, R. (1997), "Optical Detection of Single Molecules," *Annual Review of Biophysics and Biomolecular Structure*, 26, 567–596. [328]

Olshen, A. B., Venkatraman, E., Lucito, R., and Wigler, M. (2004), "Circular Binary Segmentation for the Analysis of Array-Based DNA Copy Number Data," *Biostatistics*, 5, 557–572. [315,322]

O Ruanaidh, J. J. K., and Fitzgerald, W. J. (1996), *Numerical Bayesian Methods Applied to Signal Processing*, New York: Springer-Verlag. [314]

Pesaran, M. H., Pettenuzzo, D., and Timmermann, A. (2006), "Forecasting Time Series Subject to Multiple Structural Breaks," *The Review of Economic Studies*, 73, 1057–1084. [316]

Qian, H., and Kou, S. C. (2014), "Statistics and Related Topics in Single-Molecule Biophysics," *Annual Review of Statistics and Its Application*, 1, 465–492. [328]

Scott, A. J., and Knott, M. (1974), "A Cluster Analysis Method for Grouping Means in the Analysis of Variance," *Biometrics*, 30, 507–512. [315]

Smith, A. F. M. (1975), "A Bayesian Approach to Inference About a Change-Point in a Sequence of Random Variables," *Biometrika*, 62, 407–416. [314]

Tamarat, Ph, Maali, A., Lounis, B., and Orrit, M. (2000), "Ten Years of Single-Molecule Spectroscopy," *The Journal of Physical Chemistry,* Series A, 104, 1–16. [328]

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005), "Sparsity and Smoothness via the Fused Lasso," *Journal of the Royal Statistical Society,* Series B, 67, 91–108. [314,321]

Tibshirani, R., and Wang, P. (2008), "Spatial Smoothing and Hot Spot Detection for CGH Data Using the Fused Lasso," *Biostatistics*, 9, 18–29. [314,321]

Venkatraman, E., and Olshen, A. B. (2007), "A Faster Circular Binary Segmentation Algorithm for the Analysis of Array CGH Data," *Bioinformatics*, 23, 657–663. [315,322]

Wang, J., and Wolynes, P. (1995), "Intermittency of Single Molecule Reaction Dynamics in Fluctuating Environments," *Physical Review Letters*, 74, 4317–4320. [329]

Weiss, S. (2000), "Measuring Conformational Dynamics of Biomolecules by Single Molecule Fluorescence Spectroscopy," *Nature Structural Biology*, 7, 724–729. [328]

Wyse, J., and Friel, N. (2010), "Simulation-Based Bayesian Analysis for Multiple Changepoints," arXiv preprint (arXiv:1011.2932) [315]

Xie, X. S., and Lu, H. P. (1999), "Single-Molecule Enzymology," *Journal of Biological Chemistry*, 274, 15967–15970. [328]

Xie, X. S., and Trautman, J. K. (1998), "Optical Studies of Single Molecules at Room Temperature," *Annual Review of Physical Chemistry*, 49, 441–480. [328]

Yang, T. Y., and Kuo, L. (2001), "Bayesian Binary Segmentation Procedure for a Poisson Process With Multiple Changepoints," *Journal of Computational and Graphical Statistics*, 10, 772–785. [315]

Yao, Y.-C. (1988), "Estimating the Number of Change-Points via Schwarz' Criterion," *Statistics & Probability Letters*, 6, 181–189. [314,321]

Yao, Y.-C., and Au, S. T. (1989), "Least-Squares Estimation of a Step Function," *Sankhyā: The Indian Journal of Statistics,* Series A, 51, 370–381. [314]

Yildiz, A., Tomishige, M., Vale, R. D., and Selvin, P. R. (2004), "Kinesin Walks Hand-Over-Hand," *Science*, 303, 676–678. [314,324]

Zhang, N. R., and Siegmund, D. O. (2007), "A Modified Bayes Information Criterion With Applications to the Analysis of Comparative Genomic Hybridization Data," *Biometrics*, 63, 22–32. [314,317,321]