The Annals of Applied Statistics
2017, Vol. 11, No. 4, 2200–2221
https://doi.org/10.1214/17-AOAS1075

EMPIRICAL BAYESIAN ANALYSIS OF SIMULTANEOUS CHANGEPOINTS IN MULTIPLE DATA SEQUENCES

BY ZHOU FAN[1] AND LESTER MACKEY[2]

Stanford University and Microsoft Research

Copy number variations in cancer cells and volatility fluctuations in stock prices are commonly manifested as changepoints occurring at the same positions across related data sequences. We introduce a Bayesian modeling framework, BASIC, that employs a changepoint prior to capture the co-occurrence tendency in data of this type. We design efficient algorithms to sample from and maximize over the BASIC changepoint posterior and develop a Monte Carlo expectation-maximization procedure to select prior hyperparameters in an empirical Bayes fashion. We use the resulting BASIC framework to analyze DNA copy number variations in the NCI-60 cancer cell lines and to identify important events that affected the price volatility of S&P 500 stocks from 2000 to 2009.

1. Introduction. Figure 1 displays three examples of aligned sequence data. Panel (a) presents DNA copy number measurements at sorted genome locations in four human cancer cell lines [Varma et al. (2014)]. Panel (b) shows the daily stock returns of four U.S. stocks over a period of ten years. Panel (c) traces the interatomic distances between four pairs of atoms in a protein molecule over the course of a computer simulation [Lindorff-Larsen et al. (2011)]. Each sequence in each panel is reasonably modeled as having a number of discrete "changepoints," such that the characteristics of the data change abruptly at each changepoint but remain homogeneous between changepoints. In panel (a), these changepoints demarcate the boundaries of DNA stretches with abnormal copy number. In panel (b), changepoints indicate historical events that abruptly impacted the volatility of stock returns. In panel (c), changepoints indicate structural changes in the 3-D conformation of the protein molecule. For each of these examples, it is important to understand when and in which sequences changepoints occur. However, the number and locations of these changepoints are typically not known a priori and must be estimated from the data. The problem of detecting changepoints in sequential data has a rich history in the statistics literature, and we refer the reader to Basseville and Nikiforov (1993), Chen and Gupta (2012) for a more detailed review and further applications.

Received July 2016; revised April 2017.

[1]Supported by a Hertz Foundation Fellowship and an NDSEG Fellowship (DoD, Air Force Office of Scientific Research, 32 CFR 168a).

[2]Supported by a Terman Fellowship.

Key words and phrases. Changepoint detection, empirical Bayes, Markov chain Monte Carlo, copy number variation, stock price volatility.

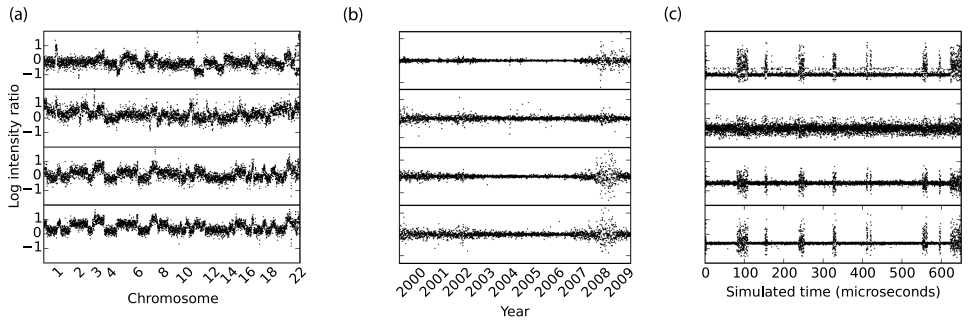


FIG. 1. (a) *DNA copy numbers in four cancer cell lines, indicated by fluorescence intensity log-ratios from array-CGH experiments*. (b) *Daily returns of four U.S. stocks*. (c) *Distances between four pairs of atoms in a computer simulation of a protein molecule*.

In many modern applications, we have available not just a single data sequence but rather many related sequences measured at the same locations or time points. These sequences often exhibit changepoints occurring at the same sequential locations. For instance, copy number variations frequently occur at common genomic locations in cancer samples [Pollack and Brown (1999)] and in biologically related individuals [Zhang et al. (2010)], economic and political events can impact the volatility of many stock returns in tandem, and a conformational change in a region of a protein molecule can affect distances between multiple atomic pairs [Fan et al. (2015)]. As recognized in many recent papers, discussed below, an analysis of multiple sequences jointly may yield greater statistical power in detecting their changepoints than analyses of the sequences individually. In addition, a joint analysis may more precisely identify the times or locations at which changepoints occur and better highlight the locations where changepoints most frequently recur across sequences.

Motivated by these considerations, we introduce a Bayesian modeling framework, *BASIC*, for carrying out a *B*ayesian *A*nalysis of *SI*multaneous *C*hangepoints. In single-sequence applications, Bayesian changepoint detectors have been shown to exhibit favorable performance in comparison with other available methods and have enjoyed widespread use [Barry and Hartigan (1993), Chernoff and Zacks (1964), Chib (1998), Fearnhead (2006), Stephens (1994), Yao (1984), Adams and MacKay (2007)]. In Section 2, we propose an extension of Bayesian changepoint detection to the multi-sequence setting by defining a hierarchical prior over latent changepoints, which first specifies the sequential locations at which changepoints may occur and then specifies the sequences that contain a changepoint at each such location.

Inference in the BASIC model is carried out through efficient, tailored Markov chain Monte Carlo (MCMC) procedures (Section 3.1) and optimization procedures (Section 3.2) designed to estimate the posterior probabilities of changepoint events and the maximum-a-posteriori (MAP) changepoint locations, respectively. These

procedures employ dynamic programming sub-routines to avoid becoming trapped in local maxima of the posterior distribution. To free the user from prespecifying prior hyperparameters, we adopt an empirical Bayes approach [Robbins (1956)] to automatic hyperparameter selection using Monte Carlo expectation maximization (MCEM) [Wei and Tanner (1990)] (Section 3.4).

To demonstrate the applicability of our model across different application domains, we use our methods to analyze two different data sets. The first is a set of array comparative genomic hybridization (aCGH) copy number measurements of the NCI-60 cancer cell lines [Varma et al. (2014)], four of which have been displayed in Figure 1(a). In Section 5, we use our method to highlight focal copy number variations that are present in multiple cell lines; many of the most prominent variations that we detect are consistent with known or suspected oncogenes and tumor suppressor genes. The second data set consists of the daily returns of 401 U.S. stocks in the S&P 500 index from the year 2000 to 2009, four of which have been displayed in Figure 1(b). In Section 6, we use our method to identify important events in the history of the U.S. stock market over this time period, pertaining to the entire market as well as to individual groups of stocks.

Comparison with existing methods: Early work on changepoint detection for multivariate data [Healy (1987), Srivastava and Worsley (1986)] studied the detection of a change in the joint distribution of all observed variables. Our viewpoint is instead largely shaped by Zhang et al. (2010), who formulated the problem as detecting changes in the marginal distributions of subsets of these variables. A variety of methods have been proposed to address variants of this problem, many with a particular focus on analysis of DNA copy number variation. These methods include segmentation procedures using scan statistics [Jeng, Cai and Li (2013), Siegmund, Yakir and Zhang (2011), Zhang et al. (2010)], model-selection penalties [Fan et al. (2015), Zhang and Siegmund (2012)], total-variation denoising [Nowak et al. (2011), Zhou et al. (2013)] and other Bayesian models [Bardwell and Fearnhead (2017), Dobigeon, Tourneret and Davy (2007), Harlé et al. (2016), Shah et al. (2007)]. Here, we briefly highlight several advantages of our present approach.

Comparing modeling assumptions, several methods [Bardwell and Fearnhead (2017), Jeng, Cai and Li (2013)] focus on the setting in which each sequence exhibits a baseline behavior, and changepoints demarcate the boundaries of nonoverlapping "aberrant regions" that deviate from this baseline. Shah et al. (2007) further assumes a hidden Markov model with a small finite set of possible signal values for each sequence. However, data in many applications are not well described by these simpler models. For instance, in cancer samples, short focal copy number aberrations may fall inside longer aberrations of entire chromosome arms and overlap in sequential position, and true copy numbers might not belong to a small set of possible values if there are fractional gains and losses due to sample heterogeneity. Conversely, the Bayesian models of Dobigeon, Tourneret and Davy (2007), Harlé et al. (2016) are very general, but their priors and inference

procedures involve 2^J parameters (where J is the number of sequences), rendering inference intractable for applications with many sequences. By introducing a prior that is exchangeable across sequences, we strike a different balance between model generality and tractability of inference.

Comparing algorithmic approaches, we observe in simulation (Section 4) that total-variation denoising can severely overestimate the number of changepoints, rendering them ill-suited for applications in which changepoint-detection accuracy (rather than signal reconstruction error) is of interest. In contrast to recursive segmentation procedures, our algorithms employ sequencewise local moves, which we believe are better suited to multi-sequence problems with complex changepoint patterns. These local moves are akin to the penalized likelihood procedure of Fan et al. (2015), but, in contrast to Fan et al. (2015), where the likelihood penalty shape and magnitude are ad hoc and user-specified, our empirical Bayes approach selects prior hyperparameters automatically using MCEM. Finally, the BASIC approach provides a unified framework that accommodates a broad range of data types and likelihood models, can detect changes of various types (e.g., in variance as well as in mean), and returns posterior probabilities for changepoint events in addition to point estimates.

2. The BASIC model. Suppose $X \in \mathbb{R}^{J\times T}$ is a collection of J aligned data sequences, each consisting of T observations. The BASIC model for X is a generative process defined by three inputs: an observation likelihood $p(\cdot|\theta)$ parameterized by $\theta \in \Theta \subseteq \mathbb{R}^d$, a prior distribution π_Θ on the parameter space Θ, and a changepoint frequency prior π_Q on $[0, 1]$. For each sequence position t, a latent variable $q_t \in [0, 1]$ is drawn from π_Q and represents the probability of any sequence having a changepoint between its $(t-1)$th and tth data points. Then, for each sequence position t and sequence j, a latent variable $Z_{j,t} \in \{0, 1\}$ is drawn with $\Pr[Z_{j,t} = 1] = q_t$ and indicates whether there is a changepoint in sequence j between its $(t-1)$th and tth data points. Finally, for each t and j, a latent likelihood parameter $\theta_{j,t} \in \Theta$ and an observed data point $X_{j,t}$ are drawn, such that $\theta_{j,t}$ remains constant (as a function of t) in each data sequence between each pair of consecutive changepoints of that sequence and is generated anew from the prior π_Θ at each changepoint, and $X_{j,t}$ is a conditionally independent draw from $p(\cdot|\theta_{j,t})$. This process is summarized as follows:

$$
\begin{aligned}
q_2, \ldots, q_T &\overset{\text{i.i.d.}}{\sim} \pi_Q, \\
Z_{j,t}|q_t &\overset{\text{ind}}{\sim} \text{Bernoulli}(q_t) \qquad \forall j = 1, \ldots, J \text{ and } t = 2, \ldots, T, \\
\theta_{1,1}, \ldots, \theta_{J,1} &\overset{\text{i.i.d.}}{\sim} \pi_\Theta, \\
\theta_{j,t}|Z_{j,t}, \theta_{j,t-1} &\begin{cases} \overset{\text{ind}}{\sim} \pi_\Theta & \text{if } Z_{j,t} = 1 \\ = \theta_{j,t-1} & \text{if } Z_{j,t} = 0 \end{cases} \qquad \forall j = 1, \ldots, J \text{ and } t = 2, \ldots, T,
\end{aligned}
$$

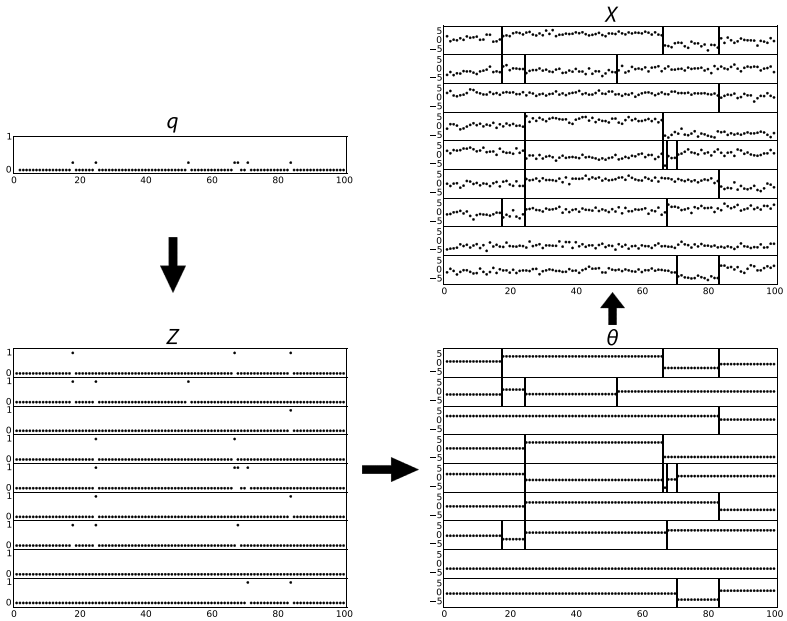


FIG. 2. *An illustration of the BASIC model. In this illustration, distinct values of* θ *are drawn from* $\pi_\Theta = \text{Normal}(0, 5)$, *and values of* X *are drawn from* $p(\cdot|\theta) = \text{Normal}(\theta, 1)$.

$$X_{j,t}|\theta_{j,t} \overset{\text{ind}}{\sim} p(\cdot|\theta_{j,t}) \qquad \forall j = 1, \ldots, J \text{ and } t = 1, \ldots, T.$$

For notational convenience, we arrange $Z_{j,t}$ into a matrix $Z \in \{0, 1\}^{J \times T}$, fixing $Z_{j,1} = 0$ for all $j = 1, \ldots, J$. Figure 2 illustrates this generative model in the case where the piecewise-constant parameter $\theta_{j,t}$ represents the mean of the distribution of $X_{j,t}$, and $X_{j,t}$ is normally distributed around this mean with fixed unit variance. Our primary goal in this model will be to infer the latent changepoint variables Z upon observing the data X.

A key input to the model is the prior distribution π_Q over $[0, 1]$, which controls how frequently changepoints occur and to what extent they co-occur across sequences. Rather than requiring the user to prespecify this prior, Section 3.4 develops an empirical Bayes MCEM procedure to select π_Q automatically. Specifically, we parametrize π_Q as a mixture distribution

$$\pi_Q = \sum_{k \in S} w_k \nu_k, \tag{1}$$

where $\{\nu_k\}_{k \in S}$ is a fixed finite dictionary of probability distributions over $[0, 1]$ and $\{w_k\}_{k \in S}$ are nonnegative mixture weights summing to 1, and the MCEM maximum marginal likelihood procedure selects the weights $\{w_k\}_{k \in S}$. In our applications, we will simply take the dictionary $\{\nu_k\}_{k \in S}$ to be discrete point masses over a fine grid of points in $[0, 1]$.

The choices of the likelihood model $p(\cdot|\theta)$ and the prior distribution π_Θ are application dependent. For our analysis of DNA copy number variations in Section 5, we use a normal model for $p(\cdot|\theta)$ where θ parametrizes the normal mean, and π_Θ is the normal conjugate prior. For our analysis of stock return volatility in Section 6, we use a Laplace model for $p(\cdot|\theta)$ with mean 0 and scale parameter θ, and π_Θ is the inverse-Gamma conjugate prior. We provide details on these and several other common models in Appendix S1. Our inference procedures are tractable whenever the marginal

$$P_j(t,s) := \int \prod_{r=t}^{s-1} p(X_{j,r}|\theta)\pi_\Theta(d\theta) \tag{2}$$

may be computed quickly from $P_j(t,s-1)$ and $P_j(t-1,s)$. This holds, in particular, whenever $p(\cdot|\theta)$ is an exponential family model with π_Θ the conjugate prior, as $P_j(t,s)$ may be computed by updating a fixed number of sufficient statistics. Any unspecified hyperparameters of π_Θ can also be selected automatically using the MCEM procedure of Section 3.4.

We have assumed for notational convenience that each data sequence is generated from the same parametric family $p(\cdot|\theta)$ with the same prior π_Θ. In applications where sequences represent different types of quantities, the choices of $p(\cdot|\theta)$ and π_Θ should vary across sequences, and our posterior inference algorithms are easily extended to accommodate this setting.

3. Inference procedures. In this section, we give a high-level overview of our algorithms for posterior inference in the BASIC model, deferring details to Appendices S2–S4. Our primary task is to perform posterior inference of the unobserved latent changepoint variables Z, given the observed data X. Assuming π_Q and π_Θ are fixed and known, Section 3.1 presents an MCMC procedure for sampling from the posterior distribution $\Pr(Z|X)$, and Section 3.2 presents an optimization algorithm to locally maximize this posterior distribution over Z to yield a MAP estimate. Section 3.4 presents an MCEM method to select π_Q and π_Θ, following the empirical Bayesian principle of maximum marginal likelihood. An efficient implementation of all inference algorithms is available on the authors' websites.

We emphasize that even though the BASIC model is specified hierarchically, our inference algorithms directly sample from and maximize over the posterior distribution of only Z, analytically marginalizing over the other latent variables q and θ. Furthermore, these procedures use dynamic programming subroutines that exactly sample from and maximize over the joint conditional distribution of many or all variables in a single row or column of Z, that is, changepoints in a single sequence or at a single location across all sequences. We verify in Appendix S5 that this greatly improves mixing of the sampler over a naïve Gibbs sampling scheme that individually samples each $Z_{j,t}$ from its univariate conditional distribution.

3.1. *Sampling from the posterior distribution*. To sample from $\Pr(Z|X)$, we propose the following high-level MCMC procedure:

1. For $j = 1, \ldots, J$, resample $Z_{j,\cdot}$ from $\Pr(Z_{j,\cdot}|X, Z_{(-j),\cdot})$.
2. For $t = 2, \ldots, T$, resample $Z_{\cdot,t}$ from $\Pr(Z_{\cdot,t}|X, Z_{\cdot,(-t)})$.
3. For $b = 1, \ldots, B$, randomly select t such that $Z_{j,t} = 1$ for at least one j, choose $s = t - 1$ or $s = t + 1$, and perform a Metropolis–Hastings step to swap $Z_{\cdot,t}$ and $Z_{\cdot,s}$.

We treat the combination of steps 1–3 above as one complete iteration of our MCMC sampler. Here, $Z_{j,\cdot}$, $Z_{(-j),\cdot}$, $Z_{\cdot,t}$ and $Z_{\cdot,(-t)}$ respectively denote the jth row, all but the jth row, the tth column and all but the tth column of Z. In step 3, B is the number of swap attempts, which we set in practice as $B = 10T$.

To sample $Z_{j,\cdot}|Z_{(-j),\cdot}$ in step 1, we adapt the dynamic programming recursions developed in Fearnhead (2006) to our setting, which require $O(T^2)$ time for each j. To sample $Z_{\cdot,t}|Z_{\cdot,(-t)}$ in step 2, we develop a novel dynamic programming recursion which performs this sampling in $O(J^2)$ time for each t. Step 3 is included to improve the positional accuracy of detected changepoints, and the swapping of columns of Z typically amounts to shifting all changepoints at position t to a new position $t + 1$ or $t - 1$ that previously had no changepoints. This step may be performed in $O(JT)$ time [when $B = O(T)$], and so one complete iteration of steps 1–3 may be performed in time $O(JT^2 + J^2T)$. Details of all three algorithmic procedures are provided in Appendix S2.

3.2. *Maximizing the posterior distribution*. To maximize $\Pr(Z|X)$ over Z, we similarly propose iterating the following three high-level steps:

1. For $j = 1, \ldots, J$, maximize $\Pr(Z|X)$ over $Z_{j,\cdot}$.
2. For $t = 2, \ldots, T$, maximize $\Pr(Z|X)$ over $Z_{\cdot,t}$.
3. For each t such that $Z_{j,t} = 1$ for at least one j, swap $Z_{\cdot,t}$ with $Z_{\cdot,t-1}$ or $Z_{\cdot,t+1}$ if this increases $\Pr(Z|X)$, and repeat.

We terminate the procedure when one iteration of all three steps leaves Z unchanged. In applications, we first perform MCMC sampling to select π_Q and π_Θ using the MCEM procedure to be described in Section 3.4, and then initialize Z in the above algorithm to a rounded average of the sampled values. Under this initialization, we find empirically that the above algorithm converges in very few iterations.

To maximize $\Pr(Z|X)$ over $Z_{j,\cdot}$ in step 1, we adapt the dynamic programming recursions developed in Jackson et al. (2005) to our setting, which require $O(T^2)$ time for each j. Maximization over $Z_{\cdot,t}$ in step 2 is easy to perform in $O(J \log J)$ time for each t. Step 3 is again included to improve the positional accuracy of detected changepoints, and after an $O(JT)$ initialization, each swap of step 3 may be performed in $O(J)$ time. Hence one complete iteration of steps 1–3 may be performed in time $O(JT \log J + JT^2)$. Details of all three algorithmic procedures are provided in Appendix S3.

3.3. *Reduction to linear cost in T.* In practice, T may be large, and it is desirable to improve upon the quadratic computational cost in T. For sampling, one may use the particle filter approach of Fearnhead and Liu (2007) in place of the exact sampling procedure in step 1, adding a Metropolis–Hastings rejection step in the particle-MCMC framework of Andrieu, Doucet and Holenstein (2010) to correct for the approximation error. For maximization, one may use the PELT idea of Killick, Fearnhead and Eckley (2012) to prune the computation in step 1, with modifications for a position-dependent cost as described in Fan et al. (2015).

In our applications we adopt a simpler approach of dividing each row $Z_{j,\cdot}$ into contiguous blocks and sampling or maximizing over the blocks sequentially; details of this algorithmic modification are provided in Appendices S2–S3. This reduces the computational cost of one iteration of MCMC sampling to $O(J^2T)$ and of one iteration of posterior maximization to $O(JT\log J)$, provided the block sizes are $O(1)$. In all of our simulated and real data examples, we use a block size of 50 data points per sequence. We examine the effect of block size choice in Appendix S5.

3.4. *Empirical Bayes selection of priors π_Q and π_Θ.* To select π_Q and π_Θ automatically using the empirical Bayes principle of maximum marginal likelihood, we assume π_Q is a mixture as in equation (1) over a fixed dictionary $\{\nu_k\}$, and we estimate the weights $\{w_k\}$. We also assume that π_Θ is parametrized by a low-dimensional parameter η, and we estimate η. We denote $P_j(t,s)$ in equation (2) by $P_j(t,s|\eta)$.

Let $\mathcal{S}(Z_{j,\cdot})$ denote the data segments $\{(1,t_1),(t_1,t_2),\dots,(t_k,T+1)\}$ induced by changepoints $Z_{j,\cdot}$, that is, $Z_{j,t_1}=\cdots=Z_{j,t_k}=1$ and $Z_{j,t}=0$ for all other t. Let $N_l=\#\{t\ge 2:\sum_{j=1}^J Z_{j,t}=l\}$ be the total number of positions where exactly l sequences have a changepoint. Our MCEM approach to maximizing the marginal likelihood over candidate priors operates on the "complete" marginal log-likelihood,

$$\begin{aligned}
&\log\Pr\bigl(X,Z|\{w_k\},\eta\bigr)\\
&\quad=\log\Pr(X|Z,\eta)+\log\Pr\bigl(Z|\{w_k\}\bigr)\\
&\quad=\Biggl(\sum_{j=1}^J\sum_{(t,s)\in\mathcal{S}(Z_{j,\cdot})}\log P_j(t,s|\eta)\Biggr)\\
&\qquad+\sum_{l=0}^J N_l\log\Biggl(\sum_{k\in S}w_k\int q^l(1-q)^{J-l}\nu_k(dq)\Biggr).
\end{aligned}$$

Starting with the initializations $\{w_k^{(0)}\}$ and $\eta^{(0)}$, EM iteratively computes the expected complete marginal log-likelihood (E-step)

$$l^{(i)}\bigl(\{w_k\},\eta\bigr)=\mathbb{E}_{Z|X,\{w_k^{(i-1)}\},\eta^{(i-1)}}\bigl[\log\Pr\bigl(X,Z|\{w_k\},\eta\bigr)\bigr]$$

and maximizes this quantity to select new prior estimates (M-step)

$$\{w_k^{(i)}\}, \eta^{(i)} = \operatorname*{argmax}_{\{w_k\},\eta} l^{(i)}(\{w_k\}, \eta).$$

MCEM approximates the E-step by a Monte Carlo sample average,

$$\mathbb{E}_{Z|X,\{w_k^{(i-1)}\},\eta^{(i-1)}}\big[\log \Pr(X, Z|\{w_k\}, \eta)\big] \approx \frac{1}{M}\sum_{m=1}^{M} \log \Pr\big(X, Z^{(m)}|\{w_k\}, \eta\big),$$

where $Z^{(1)}, \ldots, Z^{(M)}$ are MCMC samples under the prior estimates $\{w_k^{(i-1)}\}$ and $\eta^{(i-1)}$. Maximization over $\{w_k\}$ and η are decoupled in the M-step:

$$\{w_k^{(i)}\} = \operatorname*{argmax}_{\{w_k\}} \sum_{m=1}^{M}\sum_{l=0}^{J} N_l^{(m)} \log\Big(\sum_{k\in S} w_k \Big(\int q^l(1-q)^{J-l}\nu_k(dq)\Big)\Big),$$

$$\eta^{(i)} = \operatorname*{argmax}_{\eta} \sum_{m=1}^{M}\sum_{j=1}^{J} \sum_{(t,s)\in\mathcal{S}(Z_{j,\cdot}^{(m)})} \log P_j(t, s|\eta),$$

where $N_l^{(m)} = \#\{t \geq 2 : \sum_{j=1}^{J} Z_{j,t}^{(m)} = l\}$. Maximization over $\{w_k\}$ is convex, and we use a tailored KL-divergence-minimization algorithm for this purpose. We use a generic optimization routine to maximize over the low-dimensional parameter η. In our applications, we take $\{\nu_k\}_{k\in S}$ to be point masses at a grid of points with spacing $1/J$ and spanning the range $[0, J/2)$, and we initialize $\{w_k^{(0)}\}$ to assign large weight at 0 and spread the remaining weight uniformly over the other grid points. We initialize $\eta^{(0)}$ by dividing the data sequences into blocks and matching moments. Details of the optimization and initialization procedures are given in Appendix S4.

4. Simulation studies.

4.1. *Assessing inference on a small example.* We first illustrate our inference procedures on the small data example shown in Figure 2, with $J = 9$ sequences and $T = 100$ data points per sequence. This data was generated according to the BASIC model [with $\theta := (\mu, \sigma^2)$, $p(\cdot|\theta) = \text{Normal}(\mu, \sigma^2)$, π_Θ given by $\mu \sim \text{Normal}(0, 5)$ and $\sigma^2 = 1$, and $\pi_Q = 0.9\delta_0 + 0.1\delta_{2/9}$].

Figure 3 shows the effectiveness of the empirical Bayesian MCEM approach to inference in this setting. Panel (a) shows the marginal posterior changepoint probabilities $\Pr(Z_{j,t} = 1|X)$ computed with 50 MCMC samples after a 50-sample burn-in in an idealized setting where the sampling is performed under the true priors π_Q and π_Θ that generated the data. The results of panel (a) represent an idealized gold standard, as "true priors" are typically unknown in practice. Panel (c) demonstrates, however, that performance comparable to the gold standard can

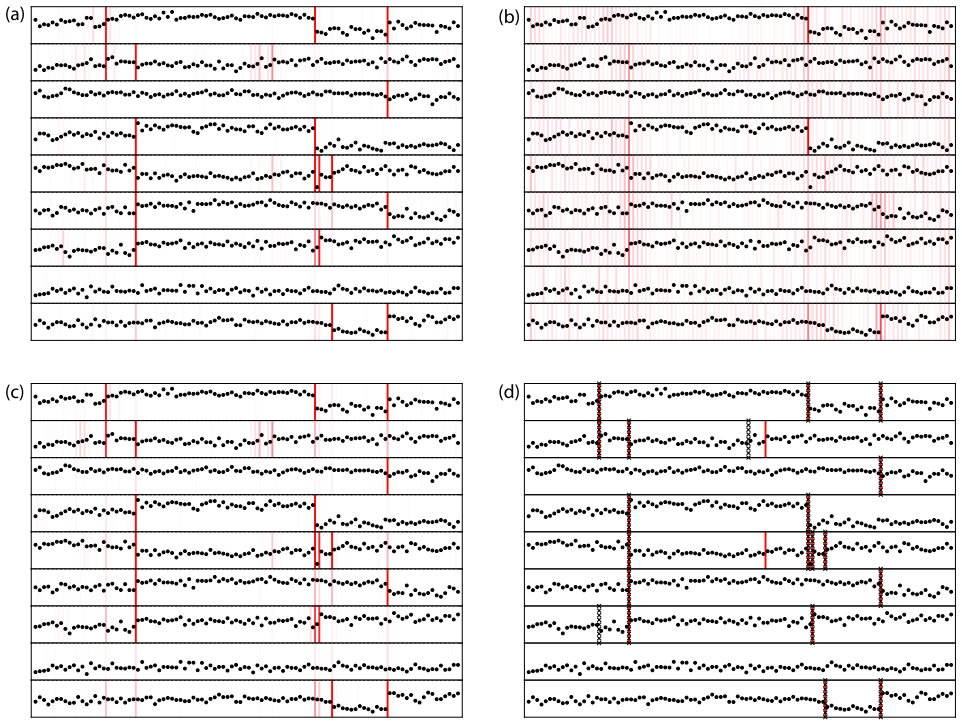


FIG. 3. *BASIC posterior inference on data generated from the BASIC model* (*see Section* 4.1). *Heatmaps* (a)–(c) *display the marginal posterior probabilities of change* $\Pr(Z_{j,t} = 1|X)$ *estimated by MCMC using* (a) *the true data-generating priors* π_Q *and* π_Θ (*which in practice are unknown*), (b) *grossly incorrect priors*, *and* (c) *MCEM-selected priors*. *The MCEM procedure in* (c) *is initialized with the incorrect priors of* (b) *but recovers accuracy comparable to the idealized setting in* (a). *Under the MCEM priors of* (c), *panel* (d) *displays the MAP changepoint estimate in red and the true changepoints as black crosses*.

be obtained using MCEM-selected priors, even when the MCEM algorithm is initialized with a grossly incorrect prior guess. In particular, panel (b) displays $\Pr(Z_{j,t} = 1|X)$ under the grossly incorrect prior choices $\mu \sim \mathcal{N}(0, 10)$, $\sigma^2 = 10$, and $\pi_Q = 0.2\delta_0 + 0.2\delta_{1/9} + 0.2\delta_{2/9} + 0.2\delta_{3/9} + 0.2\delta_{4/9}$, while panel (c) displays $\Pr(Z_{j,t} = 1|X)$ when prior parameters are initialized to the same grossly incorrect choices and updated with an MCEM update after iterations 5, 10, 20, 30 and 50 of the burn-in. Notably, the posterior inferences using MCEM priors [panel (c)] are comparable to those of the idealized setting [panel (a)], despite this incorrect initialization. Finally, panel (d) shows the MAP estimate of Z using the priors estimated in panel (c). In this example, the MAP estimate misses two true changepoints and makes two spurious detections.

We repeated this simulation with 100 different data sets generated from the BASIC model. Table 1 summarizes results using three error measures (all averaged across the 100 experiments): the squared error of the posterior mean changepoint

TABLE 1

Errors averaged over 100 *instances of the Section* 4.1 *simulation. Posterior inference using MCEM-selected priors recovers accuracy comparable to the idealized setting of using the true data-generating priors (“True priors”), even when initialized with grossly incorrect priors (“Wrong priors”)*

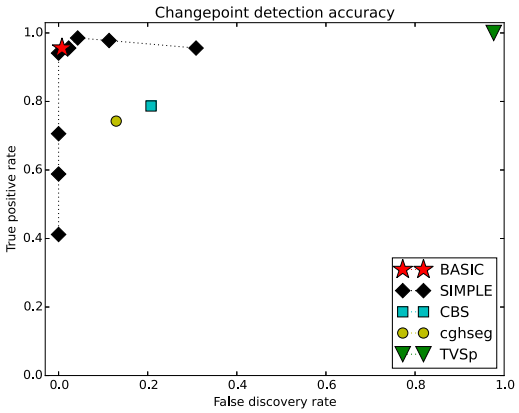
	True priors	**Wrong priors**	**MCEM priors**
Squared error of $\mathbb{E}\{Z\|X\}$	8.1	17.9	8.3
Squared error of $\mathbb{E}\{\theta\|X\}$	50.3	151	51.1
0–1 changepoint error of Z^{MAP}	10.3	14.9	10.1

indicators $\sum_{j,t}(\mathbb{E}\{Z_{j,t}|X\} - Z^{\text{true}}_{j,t})^2$, the squared error of the posterior mean signal reconstruction $\sum_{j,t}(\mathbb{E}\{\theta_{j,t}|X\} - \theta^{\text{true}}_{j,t})^2$, and the 0–1 error of detected changepoints in the MAP estimate. All evaluation metrics indicate that posterior inference using the MCEM-selected prior consistently leads to accuracy comparable to the idealized gold standard of using the true data-generating prior. As a reference point for the difficulty of this simulated data, the average 0–1 changepoint error of applying a univariate changepoint method (PELT with default MBIC penalty in the “changepoint” R package [Killick, Fearnhead and Eckley (2012)]) to each data sequence individually is 12.6, which is 25% higher than that of our MAP estimate under the MCEM-selected prior.

4.2. *Comparing detection accuracy on artificial CNV data.* The identification of copy number variations (CNVs) in aCGH data for cancer cells represents one primary motivation for our work. As there is typically no known “gold standard” for the locations of all CNVs in real aCGH data, we will assess changepoint detection accuracy in a simulation study, applying our inference procedures to 50 simulated aCGH data sequences using the simulator from Louhimo et al. (2012).[3] This simulator generates six CNVs that are either focal high-level (2-copy loss or 6-to-8-copy gain), focal medium-level (1-copy loss or 4-copy gain), or broad low-level (1-copy gain). The prevalence of each CNV across samples ranges between 5% and 50%. The simulator accounts for sample heterogeneity, with each sample corresponding to a random mixture of normal and abnormal cells.

To apply BASIC, we performed 100 iterations of MCMC sampling after 100 iterations of burn-in, using a normal likelihood model with changing mean and fixed (unknown) variance, and with MCEM updates of prior parameters after iterations 10, 20, 40, 60 and 100 of the burn-in. We then performed MAP estimation using the resulting empirical Bayes priors, with Z initialized to the MCMC sample average. On this data, the BASIC MAP estimate achieved 100% accuracy; we report results in Appendix S6.

[3]This simulator also generates corresponding gene expression data; we ignored this additional data, as integration of these two data types is not the focus of our paper.



Signal reconstruction error

Method	$\sum_{j,t}(\mu_{j,t}^{\text{est}} - \mu_{j,t}^{\text{true}})^2$
BASIC	10.40
SIMPLE	10.42
CBS	21.82
cghseg	29.23
TVSp	54.22

FIG. 4. *Changepoint detection accuracy and signal reconstruction squared error for various methods on simulated aCGH data from Louhimo et al.* (2012) (*see Section* 4.2). *Left*: *Fraction of true changepoints detected across all sequences, versus fraction of all changepoint detections that are false discoveries. Right*: *Total signal reconstruction squared error, where* $\mu_{j,t}^{\text{est}}$ *is the estimated* $\log_2$ *ratio at probe* t *in sequence* j, *and* $\mu_{j,t}^{\text{true}}$ *is its true value. For SIMPLE, we report the highest accuracy obtained across all values of its tuning parameter.*

One way in which this synthetic data is easier than the real aCGH data we analyze in Section 5 is that focal and broad CNVs span at least 50 and 500 probes, respectively, whereas they are shorter in our data of Section 5 and also in certain previous single-sample comparison studies [Lai et al. (2005)]. To increase the difficulty in this regard, we subsampled every tenth point of each synthetic data sequence and analyzed the resulting sequences in which focal CNVs span 5 probes and broad CNVs span 50. Results on this more challenging data set are reported here.

The accuracy of the BASIC MAP estimate is shown as the red star in Figure 4, where we plot the fraction of true changepoints discovered against the false-discovery proportion. Shown also in Figure 4 are the results of several alternative methods: SIMPLE [Fan et al. (2015)] to represent the penalized likelihood approach, TVSp [Zhou et al. (2013)] to represent total-variation regularization, circular binary segmentation (CBS) [Olshen et al. (2004)] applied separately to each sequence to represent a popular method of unpooled analysis, and cghseg [Picard et al. (2011)] to represent a popular method of pooled analysis. We set the convergence tolerance of TVSp to 10^{-14} and ignored changes with mean shift less than 10^{-3} to avoid identifying breakpoints because of numerical inaccuracy. We applied SIMPLE with a normal likelihood model; as the method does not prescribe a default value for the main tuning parameter, we plot its performance as the tuning parameter varies. All remaining parameters of the methods were set to their default values or selected using the provided cross-validation routines.

Detection accuracy of the BASIC MAP estimate is near perfect and competitive with the other tested methods—examination of its output reveals that it misses a focal (5-probe) medium-level loss in two sequences and a broad low-level gain in one sequence, and it makes one spurious segmentation in one sequence. Detection by cghseg is conservative, missing 10 focal gains and losses across all sequences. In addition, as cghseg does not attempt to identify changepoints at common sequential positions, it inaccurately identifies the location of 15 additional changepoints, which contributes both to an increased false discovery proportion and a reduced true discovery proportion. (This positional inaccuracy ranges between 1 and 5 probes.) Single-sequence CBS suffers from the same changepoint location inaccuracy. It is less conservative than cghseg, truly missing only 3 aberrations across all sequences, but also identifying 2 nonexistent aberrations. TVSp partitions the data sequence into too many segments, yielding false-discovery proportion close to 1 for changepoint discovery. We do note that TVSp and its tuning-parameter selection procedure are designed to minimize the signal-reconstruction squared error rather than the changepoint identification error. However, we report the signal-reconstruction errors alongside Figure 4 and observe that TVSp is also less accurate by this metric.

SIMPLE yields performance close to that of BASIC under optimal tuning, but the authors of Fan et al. (2015) provide little guidance on how to choose the tuning parameter. In the BASIC framework, the analogous hyperparameters of π_Q are selected automatically by MCEM.

5. Copy number aberrations in the NCI-60 cancer cell lines. We applied our BASIC model to analyze CNVs in aCGH data for the NCI-60 cell lines, a set of 60 cancer cell lines derived from human tumors in a variety of tissues and organs, as reported in Varma et al. (2014). We discarded measurements on the sex chromosomes, removed outlier measurements, and centered each sequence to have median 0; we discuss these preprocessing steps in Appendix S7. We fit the BASIC model using a normal likelihood with changing mean and fixed variance, applying the same procedure as in Section 4.2. The runtime of our analysis on the pooled data ($J = 125$, $T = 40{,}217$) was 2 hours.

In this data, measurements for 59 of the 60 cell lines were made with at least two technical replicates. We used this to test the changepoint detection consistency of various methods by constructing two data sets of 59 sequences corresponding to the two replicates and applying each method to the data sets independently. A detected changepoint is "coincident" across replicates if it is also detected in the same cell line at the same probe location in the other replicate. Figure 5 plots the total number of coincident detections versus the fraction of all changepoint detections that are coincident for the methods tested in Section 4.2. (We omit the comparison with TVSp due to its high false-discovery rate for changepoint identification.) BASIC has better performance than single-sample CBS, yielding more

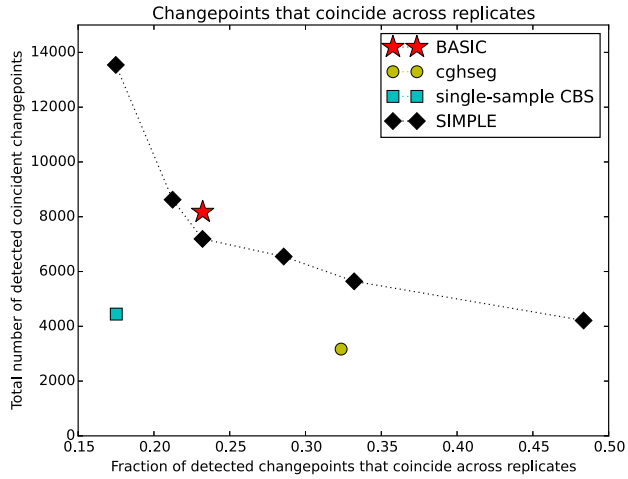


FIG. 5. *Comparison of methods by the total number and rate of detected changepoints that are coincident across two technical replicates of real aCGH data for* 59 *cancer cell lines* (*see Section* 5). *The performance of SIMPLE varies with its unspecified tuning parameter.*

coincident detections also at a higher coincidence rate. BASIC is less conservative than cghseg, detecting more coincident changes but at a lower coincidence rate. Recall that the performance of SIMPLE varies with its unspecified tuning parameter. For comparable tunings of SIMPLE, BASIC yields slightly better performance: for the same level of changepoint coincidence across replicates, BASIC detects more changepoints, and for comparable numbers of detected changepoints, BASIC achieves a higher level of changepoint coincidence.

We emphasize that a noncoincident detection is not necessarily wrong—for a changepoint demarcating a low-level aberration against which a method does not have full detection power, a method may detect this change in one replicate but not the other. Conversely, a coincident detection need not correspond to a true CNV if technical artifacts are present in both replicates. The coincidence rate is not high for any tested method. Reasons for this include the following: (1) changepoints due to technical drift, a common occurrence [Olshen et al. (2004)] which is particularly severe in some of the sequences of this data set; (2) probe artifacts that differ across replicates; and (3) low-level nonshared aberrations with boundary points that are difficult to precisely identify. The coincidence rate may be increased for all methods by applying post-processing procedures to remove changepoints due to technical drift and probe artifacts, although these procedures are usually ad hoc.

Our BASIC framework provides not only a point estimate of changepoints, but also posterior probability estimates that may be valuable in interpreting results and also performing this type of post-processing. Figure 6 displays the $\log_2$-ratio measurements and the BASIC MAP estimate of changepoints in chromosome 1 for four distinct melanoma cell lines alongside the estimated marginal posterior

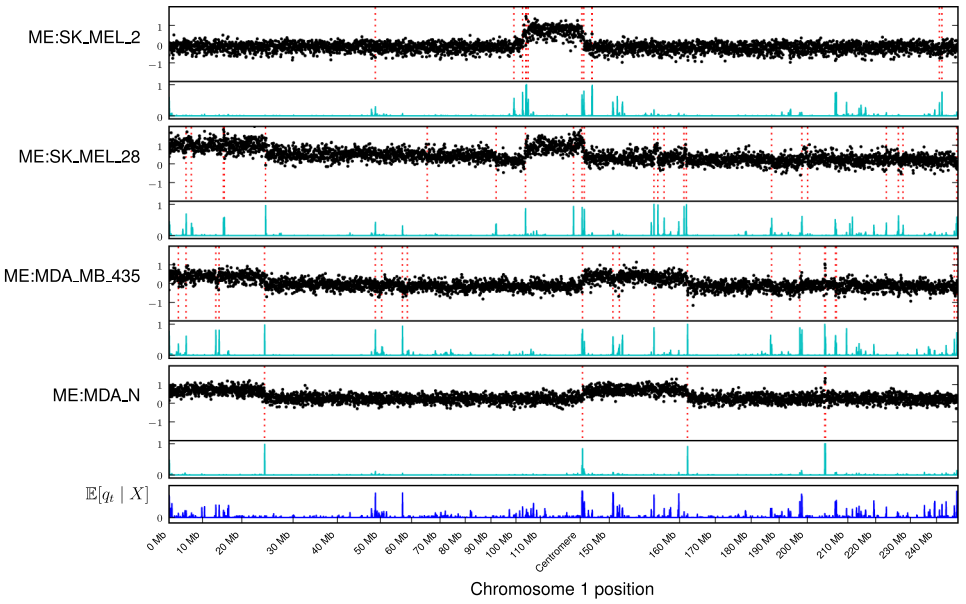


FIG. 6. *Chromosome* 1 *aCGH measurements for four NCI-*60 *melanoma cell lines* (*black points*) *and associated BASIC estimates of marginal posterior changepoint probabilities using* 100 *MCMC samples* (*teal curves*). *Red dashed lines indicate BASIC MAP changepoint estimates. The estimated posterior mean of* q_t *is displayed below in blue, providing a cross-sample summary of changepoint prevalence across all* 125 *analyzed sequences.*

changepoint probabilities. Figure 6 also displays the posterior mean estimate of q_t (computed from the sampled Z matrices), which provides a cross-sample summary of the prevalence of shared changepoints across all analyzed sequences at each probe location.

To illustrate one use of this posterior information, we performed a pooled analysis of all sequences (including all replicates to increase detection power and accuracy) in order to highlight genomic locations that contain focal and shared CNVs. First, we identified all pairs of genomic locations s and t on the same chromosome at distance less than 3×10^6 base pairs apart[4] such that at least two distinct cell lines had posterior probability greater than 90% of containing changepoints at both s and t. The interval between s and t is the identified CNV, and the sequences having posterior probability greater than 90% of change at s and t are the identified carriers of that CNV. To reduce false discoveries due to technical noise of the aCGH experiments, we restricted attention to those pairs for which this interval contained at least three microarray probes. Then, for each such pair, we computed the mean value of the data in the interval between s and t for the carrier sequences

[4]We use 3 million base pairs as the cutoff to distinguish focal from nonfocal CNVs.

and compared this to the mean value in small intervals before s and after t. Figure 7 shows the 20 identified CNVs that exhibit the greatest absolute difference between these mean values, displaying up to five distinct carriers of each CNV. CNVs that overlap in genomic position are grouped together in the figure.

Many of the CNVs highlighted in Figure 7 contain genes that have been previously studied in relation to cancer; we have annotated the figure with some of these gene names. CDKN2A and CDKN2B are well-known tumor suppressor genes whose deletion and mutation have been observed across many cancer types [Kamb et al. (1994), Nobori (1994)]. FBXW7 is a known tumor suppressor gene that plays a role in cellular division [Akhoondi et al. (2007)]. MYC is a well-known oncogene that is commonly amplified in many cancers [Dang (2012)]. URI1 is a known oncogene in ovarian cancer [Theurillat et al. (2011)]. FAF1 is believed to be a tumor suppressor gene involved in the regulation of apoptosis [Menges, Altomare and Testa (2009)]. Deletion of A2BP1 has been previously observed in colon cancer tumors and gastric cancer cell lines [Trautmann et al. (2006), Tada et al. (2010)]. Deletion of APOBEC3 has been observed in breast cancer [Long et al. (2013), Xuan et al. (2013)], although we detect its deletion in cell lines of cancers of the central nervous system and the lung. Deletion of CFHR3 and CFHR1 is not specifically linked to cancer, but it is a common haplotype that has been observed in many healthy individuals [Hughes et al. (2006)]. Many of the remaining CNVs in Figure 7 appear to represent true copy number variations present in the data (rather than spurious detections by our algorithm), but we could not validate the genes present in the corresponding genomic regions against the cancer genomics literature.

6. Price volatility in S&P 500 stocks. As a second example, we applied the BASIC model to analyze the volatility in returns of U.S. stocks from the year 2000 to 2009. We collected from Yahoo Finance the daily adjusted closing prices of stocks that were in the S&P 500 index fund over the entire duration of this 10-year period, and we computed the daily return of each stock on each trading day t as $(p_t - p_{t-1})/p_{t-1}$, where p_t is its closing price on day t and p_{t-1} is its closing price on the previous day. Our data consists of the returns for $J = 401$ stocks over $T = 2514$ trading days, and the total runtime of our pooled analysis was 1 hour.

Previous authors have applied univariate changepoint detection methods to analyze daily returns of the Dow Jones Industrial Index from 1970 to 1972, modeling the data as normally distributed with zero mean and piecewise constant variance [Adams and MacKay (2007), Hsu (1977)]. We observed empirically for our data that the tails of the distribution of daily returns are heavier than normal, and we instead applied BASIC using a Laplace likelihood with fixed zero mean and piecewise constant scale. We used the same MCMC/MCEM/MAP inference procedure as in Section 4.2.

Shown in Figure 8 are the daily returns for American International Group Inc. (AIG), Aon Corp. (AON), Bank of America Corp. (BAC) and The Bank of New

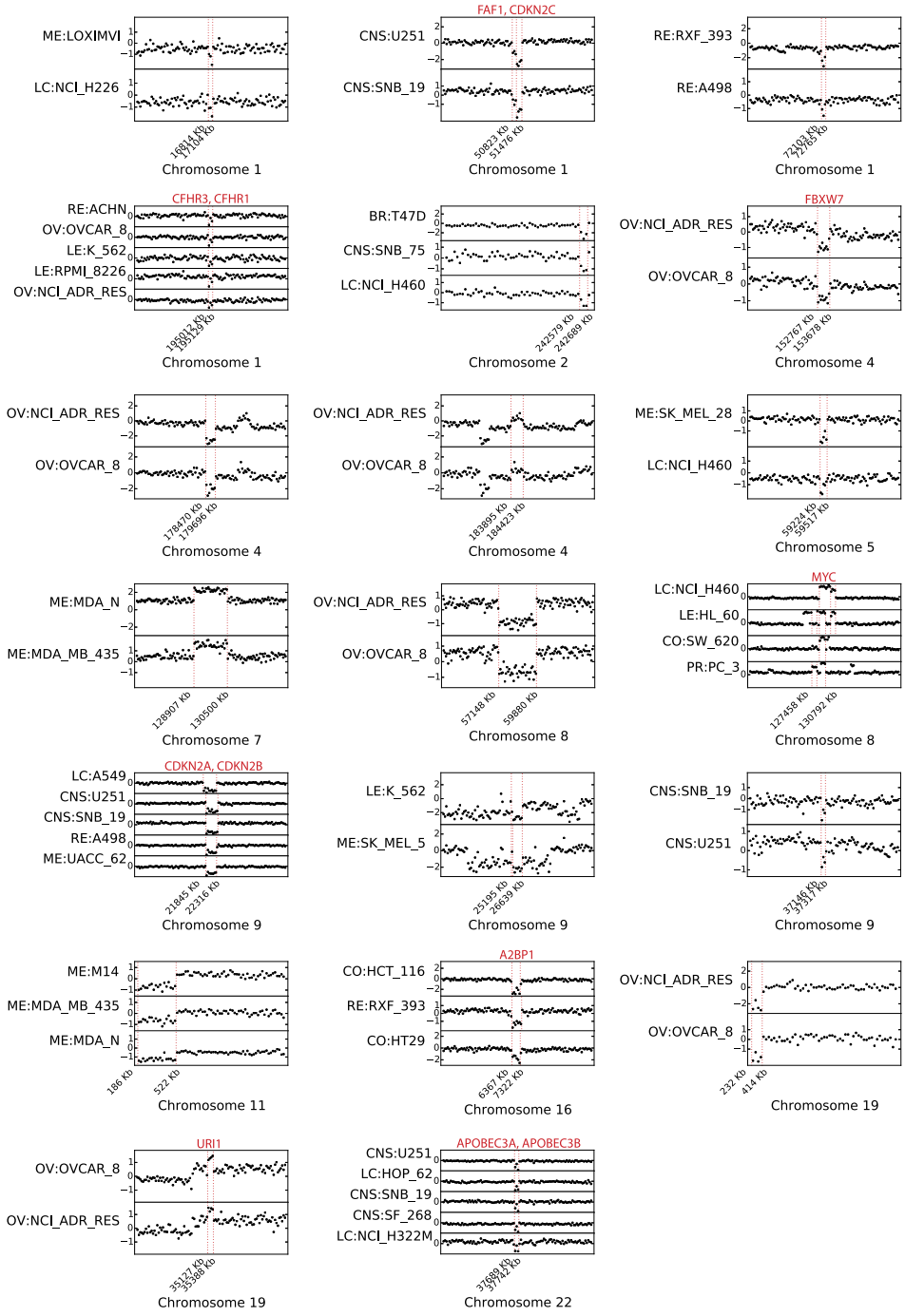


FIG. 7. *The* 20 *most prominent focal CNVs present in at least two of the NCI-*60 *cancer cell lines. Genes of interest in the aberrant regions are highlighted in red.*

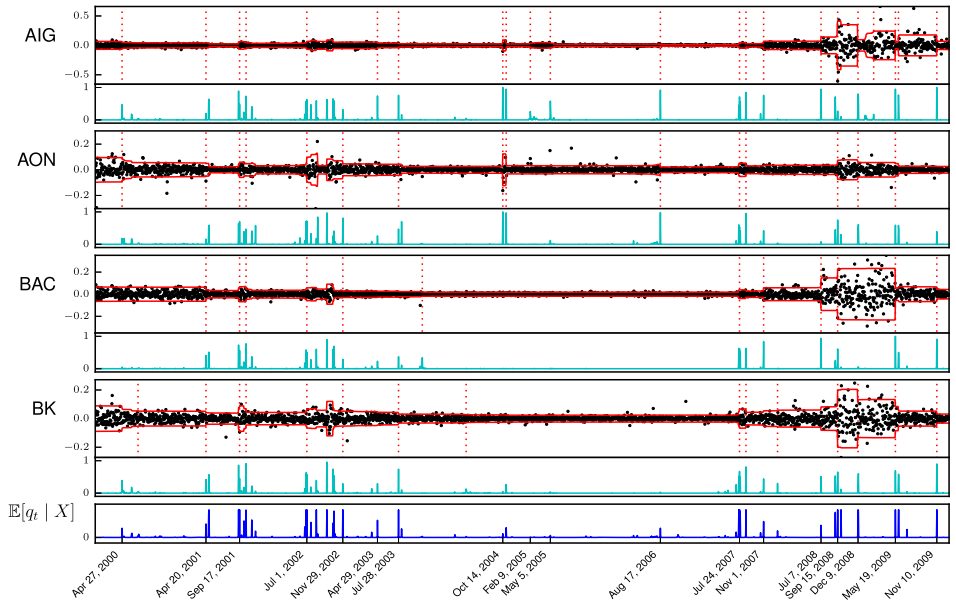


FIG. 8. *Daily returns of four U.S. stocks from* 2000 *to* 2009, *with MAP changepoint estimates* (*from a joint analysis of* 401 *stocks*) *shown in dashed red and model-based volatility estimates shown in solid red. The estimated posterior mean of* q_t *is displayed below in blue.*

York Mellon Corp. (BK), together with MAP changepoint estimates and estimated marginal posterior change probabilities. Shown also is the cross-sample changepoint summary provided by the posterior mean of q_t. Within this 10-year period, the 15 trading days with the highest posterior mean for q_t are, in chronological order: Sep 6 2001, Sep 17 2001, Jun 27 2002, Jul 1 2002, Aug 9 2002, Sept 24 2002, Nov 29 2002, Jul 24 2007, Aug 20 2007, Sep 15 2008, Sep 29 2008, Dec 9 2008, Jun 2 2009, Jun 3 2009, and Nov 10 2009. The changepoints from 2001 to 2002 are attributable to the collapse of the dot-com bubble of the late 1990s and early 2000s, and those from 2007 to 2009 are attributable to the U.S. financial crisis. Several of these dates correspond to important events in U.S. stock market history, including Sep 17 2001 when the markets first reopened after the World Trade Center terrorist attacks, Jul 1 2002 when WorldCom stock fell in value by 93%, Sept 15 2008 when Lehman Brothers filed for Chapter 11 bankruptcy, and Sept 29 2008 when the U.S. House of Representatives rejected a proposed bailout plan for the financial crisis.

Many other detected changepoints were local to small numbers of individual stocks. For instance, the changepoint detected on Oct 14 2004 and visible in the first two sequences of Figure 8 was shared across the seven stocks AIG, AON, Coventry Health Care, Hartford Financial Services, Marsh & McLennan, Merk & Co. and Unum Group. Six of these seven stocks belong to the insurance industry,

and the changepoint represents a brief spike in price volatility due to an insurance scandal that was revealed on Oct 14 2004 when AIG publicly disclosed its involvement, along with Marsh & McLennan and others, in an illegal market division scheme, and civil and criminal charges were announced against Marsh & McLennan and employees at AIG pertaining to various allegations of corporate misbehavior.[5] Other examples of detected "locally-shared" changepoints include Oct 10 2000, marking the beginning of a period of increased price volatility in the tech companies Amazon.com, Cisco Systems, EMC Corporation, JSD Uniphase, Oracle Corporation and Yahoo! Inc.; and Feb 16 2005, coinciding with the date on which the international Kyoto Protocol treaty on carbon emissions took effect and marking the start of a period of increased price volatility in the energy companies Dominion Resources, Devon Energy, Public Service Enterprise Group and Exxon Mobil.

We may also use our methods to produce a smooth estimate of the historical volatility of stock prices by computing the posterior mean of the Laplace scale parameter $\theta_{j,t}$ for each sequence j and each day t using the sampled Z matrices. The Laplace scale parameter $\theta_{j,t}$ implies a standard deviation of $\sqrt{2}\theta_{j,t}$; red lines in Figure 8 are plotted at ± 2 standard deviations to pictorially illustrate this volatility estimate. This estimate is smooth and resilient to outliers, while still exhibiting rapid adjustments to real structural changes in the data.

Acknowledgments. We would like to thank Ron Dror, David Siegmund, Janet Song and Weijie Su for helpful discussions and comments on an early draft of this paper. We would also like to thank the referees and Associate Editor for suggestions that led to many improvements in our data analyses.

SUPPLEMENTARY MATERIAL

Supplementary Appendices (DOI: 10.1214/17-AOAS1075SUPP; .pdf). The Supplementary Appendices [Fan and Mackey (2017)] contain the following additional materials, as referenced in the main text: Description of common likelihood models and associated priors, details of inference procedures, comparison of MCMC sampler with naïve Gibbs sampling, and additional details of copy number analysis for the NCI-60 cell lines.

REFERENCES

ADAMS, R. P. and MACKAY, D. J. (2007). Bayesian online changepoint detection. Technical report. Available at arXiv:0710.3742 [stat.ML].

AKHOONDI, S. et al. (2007). FBXW7/hCDC4 is a general tumor suppressor in human cancer. *Cancer Res.* **67** 9006–9012.

[5]Source: "Just how rotten?" *The Economist*, Special Report, 21 October 2004.

ANDRIEU, C., DOUCET, A. and HOLENSTEIN, R. (2010). Particle Markov chain Monte Carlo methods. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **72** 269–342. MR2758115

BARDWELL, L. and FEARNHEAD, P. (2017). Bayesian detection of abnormal segments in multiple time series. *Bayesian Anal.* **12** 193–218. MR3597572

BARRY, D. and HARTIGAN, J. A. (1993). A Bayesian analysis for change point problems. *J. Amer. Statist. Assoc.* **88** 309–319. MR1212493

BASSEVILLE, M. and NIKIFOROV, I. V. (1993). *Detection of Abrupt Changes*: *Theory and Application*. Prentice Hall, Englewood Cliffs, NJ. MR1210954

CHEN, J. and GUPTA, A. K. (2012). *Parametric Statistical Change Point Analysis*: *With Applications to Genetics*, *Medicine*, *and Finance*, 2nd ed. Birkhäuser/Springer, New York. MR3025631

CHERNOFF, H. and ZACKS, S. (1964). Estimating the current mean of a normal distribution which is subjected to changes in time. *Ann. Math. Stat.* **35** 999–1018. MR0179874

CHIB, S. (1998). Estimation and comparison of multiple change-point models. *J. Econometrics* **86** 221–241. MR1649222

DANG, C. V. (2012). MYC on the path to cancer. *Cell* **149** 22–35.

DOBIGEON, N., TOURNERET, J.-Y. and DAVY, M. (2007). Joint segmentation of piecewise constant autoregressive processes by using a hierarchical model and a Bayesian sampling approach. *IEEE Trans. Signal Process.* **55** 1251–1263. MR2464988

FAN, Z. and MACKEY, L. (2017). Supplement to "Empirical Bayesian analysis of simultaneous changepoints in multiple data sequences." DOI:10.1214/17-AOAS1075SUPP.

FAN, Z., DROR, R. O., MILDORF, T. J., PIANA, S. and SHAW, D. E. (2015). Identifying localized changes in large systems: Change-point detection for biomolecular simulations. *Proc. Natl. Acad. Sci. USA* **112** 7454–7459.

FEARNHEAD, P. (2006). Exact and efficient Bayesian inference for multiple changepoint problems. *Stat. Comput.* **16** 203–213. MR2227396

FEARNHEAD, P. and LIU, Z. (2007). On-line inference for multiple changepoint problems. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **69** 589–605. MR2370070

HARLÉ, F., CHATELAIN, F., GOUY-PAILLER, C. and ACHARD, S. (2016). Bayesian model for multiple change-points detection in multivariate time series. *IEEE Trans. Signal Process.* **64** 4351–4362. MR3528848

HEALY, J. D. (1987). A note on multivariate CUSUM procedures. *Technometrics* **29** 409–412.

HSU, D.-A. (1977). Tests for variance shift at an unknown time point. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **26** 279–284.

HUGHES, A. E. et al. (2006). A common CFH haplotype, with deletion of CFHR1 and CFHR3, is associated with lower risk of age-related macular degeneration. *Nat. Genet.* **38** 1173–1177.

JACKSON, B. et al. (2005). An algorithm for optimal partitioning of data on an interval. *IEEE Signal Process. Lett.* **12** 105–108.

JENG, X. J., CAI, T. T. and LI, H. (2013). Simultaneous discovery of rare and common segment variants. *Biometrika* **100** 157–172. MR3034330

KAMB, A. et al. (1994). A cell cycle regulator potentially involved in genesis of many tumor types. *Science* **264** 436–439.

KILLICK, R., FEARNHEAD, P. and ECKLEY, I. A. (2012). Optimal detection of changepoints with a linear computational cost. *J. Amer. Statist. Assoc.* **107** 1590–1598. MR3036418

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.

LINDORFF-LARSEN, K., PIANA, S., DROR, R. O. and SHAW, D. E. (2011). How fast-folding proteins fold. *Science* **334** 517–520.

LONG, J. et al. (2013). A common deletion in the APOBEC3 genes and breast cancer risk. *J. Natl. Cancer Inst.* **105** 573–579.

LOUHIMO, R., LEPIKHOVA, T., MONNI, O. and HAUTANIEMI, S. (2012). Comparative analysis of algorithms for integration of copy number and expression data. *Nat. Methods* **9** 351–355.

MENGES, C. W., ALTOMARE, D. A. and TESTA, J. R. (2009). FAS-associated factor 1 (FAF1): Diverse functions and implications for oncogenesis. *Cell Cycle* **8** 2528–2534.

NOBORI, T. (1994). Deletions of the cyclin-dependent kinase-4 inhibitor gene in multiple human cancers. *Trends in Genetics* **10** 228.

NOWAK, G., HASTIE, T., POLLACK, J. R. and TIBSHIRANI, R. (2011). A fused lasso latent feature model for analyzing multi-sample aCGH data. *Biostatistics* **12** 776–791.

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.

PICARD, F., LEBARBIER, E., HOEBEKE, M., RIGAILL, G., THIAM, B. and ROBIN, S. (2011). Joint segmentation, calling, and normalization of multiple CGH profiles. *Biostatistics* **12** 413–428.

POLLACK, J. R. and BROWN, P. O. (1999). Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nat. Genet.* **23** 41–46.

ROBBINS, H. (1956). An empirical Bayes approach to statistics. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, 1954–1955, *Vol. I* 157–163. Univ. California Press, Berkeley. MR0084919

SHAH, S. P., LAM, W. L., NG, R. T. and MURPHY, K. P. (2007). Modeling recurrent DNA copy number alterations in array CGH data. *Bioinformatics* **23** i450–i458.

SIEGMUND, D., YAKIR, B. and ZHANG, N. R. (2011). Detecting simultaneous variant intervals in aligned sequences. *Ann. Appl. Stat.* **5** 645–668. MR2840169

SRIVASTAVA, M. S. and WORSLEY, K. J. (1986). Likelihood ratio tests for a change in the multivariate normal mean. *J. Amer. Statist. Assoc.* **81** 199–204. MR0830581

STEPHENS, D. A. (1994). Bayesian retrospective multiple-changepoint identification. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **43** 159–178.

TADA, M. et al. (2010). Prognostic significance of genetic alterations detected by high-density single nucleotide polymorphism array in gastric cancer. *Cancer Science* **101** 1261–1269.

THEURILLAT, J.-P. et al. (2011). URI is an oncogene amplified in ovarian cancer cells and is required for their survival. *Cancer Cell* **19** 317–332.

TRAUTMANN, K. et al. (2006). Chromosomal instability in microsatellite-unstable and stable colon cancer. *Clin. Cancer Res.* **12** 6379–6385.

VARMA, S., POMMIER, Y., SUNSHINE, M., WEINSTEIN, J. N. and REINHOLD, W. C. (2014). High resolution copy number variation data in the NCI-60 cancer cell lines from whole genome microarrays accessible through CellMiner. *PLoS ONE* **9** e92047.

WEI, G. C. and TANNER, M. A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *J. Amer. Statist. Assoc.* **85** 699–704.

XUAN, D. et al. (2013). APOBEC3 deletion polymorphism is associated with breast cancer risk among women of European ancestry. *Carcinogenesis* **34** 2240–2243.

YAO, Y.-C. (1984). Estimation of a noisy discrete-time step function: Bayes and empirical Bayes approaches. *Ann. Statist.* **12** 1434–1447. MR0760698

ZHANG, N. R. and SIEGMUND, D. O. (2012). Model selection for high-dimensional, multi-sequence change-point problems. *Statist. Sinica* **22** 1507–1538. MR3027097

ZHANG, N. R., SIEGMUND, D. O., JI, H. and LI, J. Z. (2010). Detecting simultaneous changepoints in multiple sequences. *Biometrika* **97** 631–645. MR2672488

ZHOU, X., YANG, C., WAN, X., ZHAO, H. and YU, W. (2013). Multisample aCGH data analysis via total variation and spectral regularization. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **10** 230–235.

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
390 SERRA MALL
STANFORD, CALIFORNIA 94305
USA
E-MAIL: zhoufan@stanford.edu

MICROSOFT RESEARCH NEW ENGLAND
1 MEMORIAL DRIVE
CAMBRIDGE, MASSACHUSETTS 02142
USA
E-MAIL: lmackey@microsoft.com