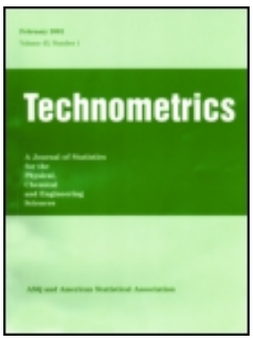

Technometrics

ISSN: 0040-1706 (Print) 1537-2723 (Online) Journal homepage: http://www.tandfonline.com/loi/utch20

Bayesian Detection of Changepoints in Finite-State Markov Chains for Multiple Sequences

Petter Arnesen, Tracy Holsclaw & Padhraic Smyth

To cite this article: Petter Arnesen, Tracy Holsclaw & Padhraic Smyth (2016) Bayesian Detection of Changepoints in Finite-State Markov Chains for Multiple Sequences, Technometrics, 58:2, 205-213, DOI: 10.1080/00401706.2015.1044118

To link to this article: http://dx.doi.org/10.1080/00401706.2015.1044118

Bayesian Detection of Changepoints in Finite-State Markov Chains for Multiple Sequences

Petter ARNESEN

Department of Mathematical Sciences
Norwegian University of Science and Technology
Trondheim 7491, Norway
(*petterar@math.ntnu.no*)

Tracy HOLSCLAW

Department of Computer Science and the Department of Statistics
University of California
Irvine, CA 92697
(*tholscla@ams.ucsc.edu*)

Padhraic SMYTH

Department of Computer Science
University of California
Irvine, CA 92697
(*smyth@ics.uci.edu*)

We consider the analysis of sets of categorical sequences consisting of piecewise homogenous Markov segments. The sequences are assumed to be governed by a common underlying process with segments occurring in the same order for each sequence. Segments are defined by a set of unobserved changepoints where the positions and number of changepoints can vary from sequence to sequence. We propose a Bayesian framework for analyzing such data, placing priors on the locations of the changepoints and on the transition matrices and using Markov chain Monte Carlo (MCMC) techniques to obtain posterior samples given the data. Experimental results using simulated data illustrate how the methodology can be used for inference of posterior distributions for parameters and changepoints, as well as the ability to handle considerable variability in the locations of the changepoints across different sequences. We also investigate the application of the approach to sequential data from an application involving monsoonal rainfall patterns. Supplementary materials for this article are available online.

KEY WORDS: Changepoint model; Cross-validation; Hidden Markov model; Multiple sequences.

1. INTRODUCTION

Finite-state Markov chains are widely used to model sequential data in applications such as weather models (Gabriel and Neumann 1962), speech recognition (Rabiner 1989), bioinformatics (Durbin et al. 1998), and more (Guttorp 1995). A common assumption is that the chain is homogenous, often motivated by a desire to keep the number of model parameters tractable. In practice, however, inhomogeneity in various forms is often present.

In particular in this article, we investigate the problem of modeling sets of categorical-valued sequences where each sequence is assumed to be generated by an ordered set of segments, with unobserved segment boundaries that can vary from sequence to sequence. Each segment has its own Markov dynamics, representing common "phases" for some underlying process. As a specific example, consider the modeling of rainfall at a particular location. Markov chains have a long history of use as stochastic models of rainfall (Newnham 1916; Gold 1929; Cochran 1938; Wilks and Wilby 1999; Chen, Brissette, and Leconte 2010). Figure 1 is an illustration of annual daily rainfall sequences for a weather station in northern India.

The appearance of the Indian monsoon in early summer causes a visible phase-change in the sequence of daily rainfall for each year, motivating the type of segment-based model that is the focus of this article. Being able to infer monsoon onset and withdrawal dates, and to effectively handle interannual variability, is of considerable interest in agriculture modeling and climate science (e.g., Joseph, Eischeid, and Pyle 1994). Other examples of datasets characterized by multiple sequences with common segmental structure include tree branching patterns (Guédon et al. 2001), sequences of bird songs (Craig 1943; Raftery and Tavaré 1994), wind-speed data (Berchtold 1999), DNA sequences (Berchtold 2002; Fearnhead and Liu 2007), and dialog transcripts (Levin and Pieraccini 2000).

In this article, we model a particular type of shared structure where a categorical sequence is assumed to consist of $K+1$ *segments* separated by K unobserved *changepoints*. Within each segment the observed data are assumed to be generated by a homogenous finite-state Markov chain. We will assume that each of the $K+1$ segments has its own Markov parameters, resulting in $K+1$ transition matrices. The location of each changepoint is modeled by a discrete distribution conditioned on the location of the previous changepoint. We will consider

Technometrics, MAY 2016, VOL. 58, NO. 2
DOI: 10.1080/00401706.2015.1044118

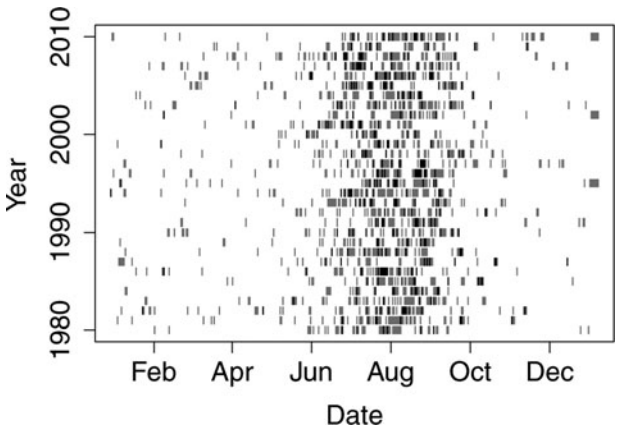

Figure 1. Thirty-one years of rainfall categorized into no/light rainfall (white), medium rainfall (gray), and heavy rainfall (black).

the case of a set of L observed sequences (of potentially varying length) where the locations of the K changepoints can vary across the sequences. The L sequences share information in that the parameters of the Markov transition matrices are global, and the order in which the segments appear in each sequence are assumed to be the same, corresponding to common underlying phases. Our framework allows any segment to be skipped in a particular sequence, and thus, K is in effect the maximal number of changepoints that can occur in any particular sequence.

There is a large amount of prior literature on changepoint detection for sequential data, covering a variety of aspects of this problem, including online/offline estimation, single versus multiple sequences, Bayesian versus non-Bayesian inference, and methods for inferring the number of changepoints. Much of this prior work has focused on the case of independent observations and/or noncategorical data. An exception is the double-Markov chain approach proposed by Berchtold (1999) using the expectation-maximization (EM) algorithm for inference, and which was further developed within a Bayesian framework by Fitzpatrick and Marchev (2013). Because of the Markov assumption at the segment-transition level, the model imposes an implicit geometric distribution on segment lengths, which is not always ideal in practice (we will follow-up on this point later in the article). The geometric assumption can be relaxed by using a hidden semi-Markov approach, for example (Guédon 2003), but this imposes computational restrictions, namely, that the time complexity of inference (e.g., via the forward-backward algorithm, Rabiner 1989) can scale as $O(T^2)$, where T is the length of a sequence.

Polansky (2007) proposed a likelihood-based framework for segmentation assuming a process that switches between different Markov chains with unknown parameters, but for the case of a single sequence rather than multiple sequences. A restricted variant of the multiple sequence changepoint problem is the case where sequences are required to be of the same length and the changepoints are assumed to occur at the same location in each sequence, which can also be viewed as a single multivariate sequence (Lai and Xing 2011; Xing, Sun, and Chen 2012; Fitzpatrick and Marchev 2013). Zhang and Siegmund (2012) relaxed this approach to allow subsets of sequences (of equal length and for the case of independent real-valued observations) to share changepoints. In the approach proposed here, we allow each sequence to have its own changepoint locations and sequences to have different lengths.

While most prior work has focused on likelihood-based inference and point estimates of changepoints, here we use full Bayesian inference for both changepoint locations and model parameters. Our MCMC inference algorithm has a time complexity per iteration that is linear in the length T of a sequence, avoiding the $O(T^2)$ time complexity incurred by Rigaill, Lebarbier, and Robin (2012). Fearnhead and Liu (2011) developed a Bayesian changepoint detection that is linear in T in terms of time complexity, but for the case of a single real-valued sequence.

Our focus on multiple sequences also provides a natural context for using cross-validation (across sequences) for model selection, providing a practical alternative to approaches such as the Bayesian information criterion (BIC; Guédon 2003; Polansky 2007; Fitzpatrick and Marchev 2013; Luong, Rozenholc, and Nuel 2013), and related penalized likelihood approaches (Zhang and Siegmund 2007; Rigaill, Lebarbier, and Robin 2012; Cleynen and Lebarbier 2014), which are not always appropriate or effective for changepoint problems (e.g., see the discussion in Cleynen et al. 2014).

In Section 2, we introduce our proposed changepoint Markov model, with results on model selection in Section 3. In Section 4.1, we present an example of simulated data to illustrate different aspects of the proposed approach. Section 4.2 describes the application of our approach to a dataset consisting of multiple years of daily rainfall data in India. We conclude the article in Section 5 with a brief discussion of the main results as well as suggestions for future directions. Discussions on MCMC sampling, how to handle missing data, additional simulated examples, and an application to a second real dataset are provided in the supplementary materials.

2. A CHANGEPOINT MARKOV MODEL FOR CATEGORICAL SEQUENCES

In this section, we define our Bayesian model for detection of a fixed number of changepoints in multiple sequences. The MCMC sampling algorithm we construct for sampling from the resulting posterior distribution is given in Section S.1 in the supplementary materials.

2.1 Homogenous Markov Chains

Given a sequence of discrete observations $\mathbf{y} = (y_1, \ldots, y_T)$, such that $y_j \in \{1, \ldots, n\}$ for all $j = 1, \ldots, T$, $\mathbf{y}$ is a realization from a finite-state Markov chain if the joint probability of $\mathbf{y}$ can be written as

$$p(\mathbf{y}) = p_1(y_1)\prod_{j=2}^{T} p_j(y_j|y_{j-1}, \ldots, y_1) = \prod_{j=1}^{T} p_j(y_j|y_{j-1}), \tag{1}$$

where the conditional distribution of y_j given $y_{j-1}, \ldots, y_1$ only depends on the previous state y_{j-1}, and we assume some given initial state y_0 such that $p_1(y_1) = p_1(y_1|y_0)$. If the transition distribution $p_j(y_j|y_{j-1}) = p(y_j|y_{j-1})$ for all $j = 1, \ldots, T$, we say that the Markov chain is homogenous. The transition

probabilities can be organized into an $n \times n$ transition matrix $\mathbf{Q}$, where the rows indicate the value y_{j-1} and the columns represent the values of y_j. We relax these assumptions below to allow the Markov chain to be *piecewise homogenous*, allowing inhomogeneity via local segmentation of the sequence $\boldsymbol{y}$.

2.2 Modeling Local Segments

To define a piecewise homogenous Markov chain, we divide the T observations into $K+1$ segments by introducing K changepoints $\tau_0 = 0 \le \tau_1 \le \cdots \le \tau_K \le T$, and write $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_K)$ as the vector of changepoints. If $\tau_{i-1} < \tau_i$ then we denote $s_i = \{\tau_{i-1}+1, \ldots, \tau_i\}$ to be the ith segment and if $\tau_{i-1} = \tau_i$ (a segment of length zero) then $s_i = \emptyset$. Introducing the notation $\boldsymbol{y}_s = (y_j | j \in s)$ for $s \subseteq \{1, \ldots, T\}$, the ith segment will include the data points $\boldsymbol{y}_{s_i}$. We assume the locations of the K changepoints to be generated by a distribution of the form

$$p(\boldsymbol{\tau}|T, \bar{\boldsymbol{\theta}}) = p(\tau_1, \ldots, \tau_K|T, \bar{\boldsymbol{\theta}}) = \prod_{i=1}^{K} p(\tau_i|\tau_{i-1}, T, \boldsymbol{\theta}_i), \quad (2)$$

where $p(\tau_i|\tau_{i-1}, T, \boldsymbol{\theta}_i)$ is a discrete parametric distribution for changepoint τ_i, with parameter vector $\boldsymbol{\theta}_i$, and where we let $\bar{\boldsymbol{\theta}}$ denote the collection of the parameter vectors $\boldsymbol{\theta}_i$ for $i = 1, \ldots, K$. In the remainder of the article, we assume the distribution for τ_i to be the negative binomial distribution truncated to the interval (τ_{i-1}, T) (additional details in Section 2.4). The two-parameter negative binomial distribution provides additional flexibility in modeling segment lengths compared to a single parameter distribution such as the geometric distribution. This can lead to more accurate detection of changepoints, as we will see later in the article.

2.3 The Piecewise Homogenous Markov Chain

We assume that within each segment the sequential observations in $\boldsymbol{y}$ evolve according to a fixed transition matrix, that is, $\boldsymbol{y}$ is generated by a piecewise homogenous Markov chain within each segment. With K changepoints, we have a total of $K+1$ transition matrices $\mathbf{Q}^{(1)}, \ldots, \mathbf{Q}^{(K+1)}$, each having size $n \times n$. In what follows we present the case where the $K+1$ transition matrices are modeled separately and independently, but it is straightforward to constrain some of these matrices to be the same (as in our rainfall example later in the article) and to reduce the parameter count accordingly. Let $\bar{\mathbf{Q}}$ denote the collection of these $K+1$ transition matrices, and let $\mathbf{Q}^{(i)}_{k,l}$ denote the element at the kth row and the lth column of matrix i. Given an initial state y_0, the data likelihood is

$$p(\mathbf{y}|\boldsymbol{\tau}, \bar{\mathbf{Q}}, y_0) = \prod_{j=1}^{T} p(y_j|y_{j-1}, \bar{\mathbf{Q}}, \boldsymbol{\tau}) = \prod_{i=1}^{K+1} \prod_{j \in s_i} \mathbf{Q}^{(i)}_{y_{j-1}, y_j}. \quad (3)$$

Assume for the moment that no segments are of length zero, that is, no segments are skipped. If y_j is the first data point in segment s_i, $i > 1$, then we assume it to be distributed according to the conditional distribution of y_j given y_{j-1} (the last point in segment s_{i-1}) using the transition matrix for segment s_i, $\mathbf{Q}^{(i)}$. The extension to segments of length zero is straightforward.

We adopt a fully Bayesian approach conditioned on fixed K for our model. For priors for the transition matrices, we assume each of the rows to be independently distributed according to the Dirichlet distribution. In particular, $\mathbf{Q}^{(i)}_{k,\cdot} \sim \text{Dir}(\boldsymbol{\alpha}^{(i)}_k)$, where $\mathbf{Q}^{(i)}_{k,\cdot}$ is the kth row of the ith transition matrix and where $\text{Dir}(\boldsymbol{\alpha}^{(i)}_k)$ denotes the Dirichlet distribution with parameter vector $\boldsymbol{\alpha}^{(i)}_k$ of appropriate length. We set all elements in $\boldsymbol{\alpha}^{(i)}_k$ to be equal to 1, for all i and k in our examples, rendering the prior equivalent to having seen one transition from each category to every other category including itself.

2.4 Modeling Multiple Sequences of Variable Length

To handle multiple sequences, consider L conditionally independent sequences of observations, $\boldsymbol{y}^{(l)}$, $l = 1, \ldots, L$, with lengths $T_1, \ldots, T_L$, where each sequence consists of $K+1$ segments occurring in the same order as described above. Also let $y^{(l)}_0$ denote the initial value for each of the l sequences. Assuming the sequences to be conditionally independent, the likelihood for multiple sequences is the product of the likelihoods for each individual sequence (as defined earlier). Segments are allowed to be of zero length (i.e., skipped), effectively allowing the number of changepoints per sequence to differ (see the simulation studies in Section S.3.1 and S.3.2 in the supplementary material for examples of this property). Let $\boldsymbol{\tau}^{(l)} = (\tau^{(l)}_1, \ldots, \tau^{(l)}_K)$ denote the changepoints in sequence l. There are a number of options for modeling how the locations of the changepoints $\boldsymbol{\tau}^{(l)}$ are related to the lengths of the sequences T_l. One could allow the changepoints to have distributions defined in absolute units (e.g., of time) and treat the total length of the sequence as a random quantity. The approach we take here is to assume that the distribution on changepoints (or equivalently, on segment length) is specified in terms of position *relative* to the total length of the sequence, where we treat the observed total sequence lengths T_l as fixed quantities and condition on them. In practice the choice of parameterization will depend on the specific nature of the application, and for the special case where the sequences are all of the same length, the absolute and relative approaches will be equivalent. In particular, we assume that the position of the changepoints $\tau^{(l)}_i$ have a negative binomial distribution, with parameters $\boldsymbol{\theta}_i = (r_i, b_i) \in (0, 1) \times (0, 1)$, truncated to the range $(\tau^{(l)}_{i-1}, T_l)$. We write

$$p(\tau^{(l)}_i|\tau^{(l)}_{i-1}, T_l, \boldsymbol{\theta_i}) \propto \frac{\Gamma(\tau^{(l)}_i + \gamma(\boldsymbol{\theta}_i, T_l))}{\tau^{(l)}_i!\,\Gamma(\gamma(\boldsymbol{\theta}_i, T_l))} b_i^{\gamma(\boldsymbol{\theta}_i, T_l)} (1-b_i)^{\tau^{(l)}_i},$$
$$\tau^{(l)}_i = \tau^{(l)}_{i-1}, \ldots, T_l, \quad (4)$$

where $\gamma(\boldsymbol{\theta}_i, T_l)$ is defined via the expression $r_i T_l = \gamma(\boldsymbol{\theta}_i, T_l)(1-b_i)/b_i$, which corresponds to the expected value of the negative binomial distribution without truncation. The parameter r_i will therefore be related to the expected position of changepoint $\tau^{(l)}_i$ scaled by the length T_l of sequence l, while b_i will be related to variance of the distribution. This truncated distribution can be efficiently computed as it does not include a computationally demanding normalizing constant. We also assume a priori that the $\boldsymbol{\theta}_i$, $i = 1, \ldots, K$ are independent, and that r_i and b_i are independent and uniformly distributed in the unit interval, $U(0, 1)$.

3. MODEL SELECTION AND COMPARISON

Previous work on estimating the number of changepoints has often used formulations such as BIC (Guédon 2003; Polansky 2007; Fitzpatrick and Marchev 2013; Luong, Rozenholc, and Nuel 2013). The BIC criterion, however, is not directly applicable to changepoint problems (as for instance discussed by Zhang and Siegmund 2007), leading to alternative penalized likelihood formulations (Zhang and Siegmund 2007; Cleynen and Lebarbier 2014; Cleynen et al. 2014). More fully Bayesian approaches have also been pursued, for example, by Fearnhead and Liu (2011) and Rigaill, Lebarbier, and Robin (2012) but for single-sequences with real or count-valued IID observations.

For multiple sequences, Xing, Sun, and Chen (2012) investigated a Bayesian approach for inferring boundaries for piecewise homogenous Markov chains with unobserved changepoints, similar to the problem we address in this article. However, their approach assumes that all sequences are the same length and have changepoints in the same position, essentially restricting the approach to the case of a single sequence with a multivariate distribution. For IID observations, Zhang and Siegmund (2012) also considered the scenario of multiple equi-length sequences with aligned changepoints across sequences, using a modified BIC criterion for model selection.

An alternative approach that could be pursued is that of trans-dimensional MCMC algorithms based on a joint posterior distribution for the model and parameter space (Green 1995). Reversible jump MCMC (RJMCMC) algorithms (Green 1995) could be used in this context, but we would anticipate slow convergence and potential sensitivity issues when choosing proposal distributions (particularly for the jump proposals). Another option is the Dirichlet process framework, although it would not be straightforward to apply this approach given the computational challenges that would result from the lack of conjugacy (Neal 2000).

Given these various issues with Bayesian and information-based criteria (at least in the context of our proposed model), we instead use cross-validation with log-probability scores (e.g., see Vehtari and Lampinen 2002; Gneiting and Raftery 2007; Gelman, Hwang, and Vehtari 2014; Li et al. 2014) for both choosing the number of changepoints and for model selection among alternative frameworks. We take advantage of the fact that we are working with multiple sequences and use cross-validation at the sequence level, using the log-probability of held-out sequences as our scoring function. Our simulation results (next section) suggest that this approach is feasible even with a relatively small number of sequences (e.g., 10). When training our model, we hold out t of our L sequences, referred to as the test set, and train our model using only $L-t$ of the available sequences, referred to as the training set. Denote the $L-t$ sequences in a training set by $\boldsymbol{Y}_D$ and the t sequences in the test set by $\boldsymbol{Y}_{-D}$. To evaluate the quality of the model, we estimate the logarithm of the probability of observing the test set $\boldsymbol{Y}_{-D}$ given the training set $\boldsymbol{Y}_D$ and the model. We average this log-probability for a number of different train/test set pairs using cross-validation. The Monte Carlo estimate of the log-probability of one of the sequences $\boldsymbol{y}^{(l)}$ in the test set is

$$\ln p(\boldsymbol{y}^{(l)}|\boldsymbol{Y}_D) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{M}\sum_{j=1}^{M}\ln p(\boldsymbol{y}^{(l)}|\boldsymbol{\tau}^{(l)[j]}, \bar{\mathbf{Q}}^{[i]}, y_0^{(l)}), \quad (5)$$

where $\boldsymbol{\tau}^{(l)[j]} \sim p(\boldsymbol{\tau}^{(l)}|T_l, \bar{\boldsymbol{\theta}}^{[i]})$ for $j = 1, \ldots, M$, and $(\bar{\boldsymbol{\theta}}^{[i]}, \bar{\mathbf{Q}}^{[i]}) \sim p(\bar{\boldsymbol{\theta}}, \bar{\mathbf{Q}}|\boldsymbol{Y}_D)$ for $i = 1, \ldots, N$, and where we use superscript $[\cdot]$ to denote simulated samples. After convergence of the algorithm and for N independent samples of $\bar{\boldsymbol{\theta}}$ and $\bar{\mathbf{Q}}$, we simulate M samples from $p(\boldsymbol{\tau}^{(l)}|T_l, \bar{\boldsymbol{\theta}})$. For each of these M samples, we calculate the value of the logarithm of the data likelihood in (3), and then compute the mean. For the results in this article we used $M = 1000$ samples, although sensitivity analysis (not shown) indicates that $M = 100$ is sufficient to obtain consistent estimates of the log-probability score. For N, which is the number of independent samples from the MCMC algorithm, we used $N = 1000$ following the discussion in Section S.1 in the supplementary materials on burn-in period and thinning. For all of our results we use 10-fold cross-validation, that is, we partition our data into 10 test and training sets (unless otherwise stated).

The total log-probability score is defined to be the sum of the individual scores divided by the sum of the length of the sequences

$$S(\boldsymbol{Y}_{-D}|\boldsymbol{Y}_D) = \frac{\sum_{l=1}^{t}\ln p(\boldsymbol{y}^{(l)}|\boldsymbol{Y}_D)}{\sum_{l=1}^{t}T_l}, \quad (6)$$

where $\boldsymbol{y}^{(1)}, \ldots, \boldsymbol{y}^{(t)}$ are the sequences in the test set. Using the equation above we can compare different models, either our proposed model with different numbers of changepoints, or our proposed model versus an alternative model. When calculating the cross-validated log-probability scores for two models $\mathcal{M}_1$ and $\mathcal{M}_2$, we use the same train/test sets and report the difference between the scores for each train/test set, that is, we calculate $S(\boldsymbol{Y}_{-D}|\boldsymbol{Y}_D, \mathcal{M}_1) - S(\boldsymbol{Y}_{-D}|\boldsymbol{Y}_D, \mathcal{M}_2)$ for each train/test set.

4. ANALYSIS AND RESULTS

In this section, we analyze both a simulated and real dataset. For the simulated data we generate the sequences and changepoints, allowing us to test our model and compare to alternative methods. Additional simulated datasets are provided in the supplementary material to further explore different aspects of our model. We also analyze a real rainfall dataset in this section, where we use our model to detect the changepoints of monsoon onset and withdrawal. An additional real-world example, involving segmentation of branching patterns for trees, is described in Section S.3.4 of the supplement.

In all of our analyses, we compare our method to three baseline models. The first baseline SI (Segmental-Independence) is the same as our proposed model but assumes the observations $\boldsymbol{y}$ are independent within each segment. The second baseline is a standard hidden Markov model (HMM) with a left-to-right transition matrix. The third baseline is the double hidden Markov chain model (dHMM) from Berchtold (1999) with a left-to-right transition matrix. These three baselines differ from our proposed model in that the SI and HMM models assume independence of the observations $\boldsymbol{y}$ (rather than Markov dependence), and

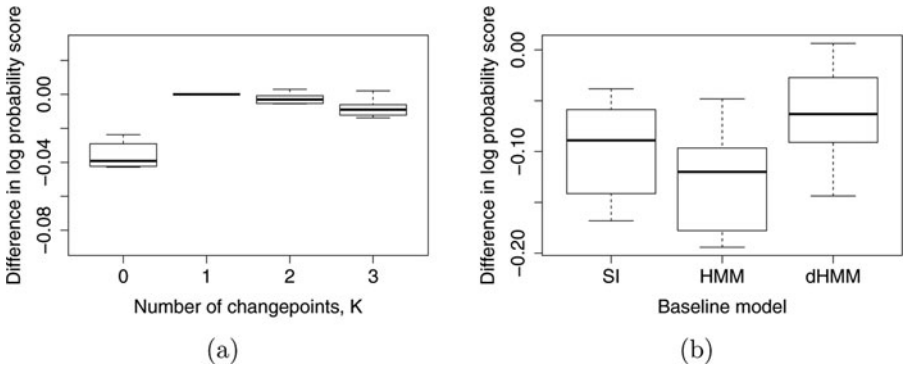

Figure 2. Each boxplot shows the log-probability scores, across the validation sets, for a different model. The y-axis is defined as the log-probability score for (a) other numbers of changepoints and (b) other baseline models, minus the log-probability score for the model with one changepoint.

the HMM and dHMM models assume geometric distributions on the changepoint locations (rather than a negative-binomial distribution). For each of the three baselines we use a Bayesian inference procedure similar to that described in Section S.1 in the supplementary materials, and we report the results for the models with the number of changepoints corresponding to the highest log-probability score (which for all the baselines turned out to be the same number of changepoints as in our preferred model). In addition to comparing our model to these baseline models, we also compare our results to those obtained with a model where each sequence is analyzed independently. In particular, we use the single-sequence model of Fearnhead (2006) for comparison, originally proposed for real-valued data with independence. See the supplementary materials for these results.

4.1 Synthetic Data Example: Scenario 1

We simulated $L = 10$ binary sequences, all of length 200, from our proposed model. For each sequence $l = 1, \ldots, 10$, we simulated one changepoint $\tau_1^{(l)}$, and the position of the changepoint was generated from the truncated negative binomial distribution as explained in Section 2.4, using the parameters $r_1 = 0.5$ and $b_1 = 0.8$, such that each segment will have length about 100. The transition matrices used to generate the binary data in each segment had diagonal entries $q_{1,i}$ and $q_{2,i}$, with i denoting the segment. We used $q_{1,1} = q_{2,1} = 0.8$ in the first segment, and $q_{1,2} = 0.5$ and $q_{2,2} = 0.4$ in the second segment.

We used our cross-validation model selection procedure described in Section 3 to determine the number of changepoints. We defined 10 train/test sets by leaving out one of the 10 sequences in each fold, resulting in a "hold-one-sequence-out" cross-validation test. The result of this model comparison is shown in Figure 2(a).

It is clear from the figure that the model with a single changepoint ($K = 1$) is the preferred one. As the number of changepoints increases beyond a single changepoint the cross-validated log-probability scores become slightly worse.

The cross-validated log-probability scores of the three baseline models relative to our model are shown in Figure 2(b), and as we can see our model performs the best. All models were fit using the optimal changepoint number ($K = 1$).

We then trained our model with the correct number of changepoints ($K = 1$), using all 10 sequences. The estimated parameter values with 95% credible intervals (CI) (corresponding to the 2.5% maximum and minimum percentiles of the posterior samples after convergence) are shown in Tables S1 and S2 in the supplementary materials. All of the parameters are well estimated, although there is considerable uncertainty concerning the estimated value of $b = b_1$. For a more detailed discussion of the posterior analysis for b and r, see the supplementary materials' Section S.2.1.

Figure 3(a) shows, for each sequence, the estimated marginal probability that each observation in that sequence belongs to segment 2. The true changepoint is marked (with an $\times$) for each sequence and, as we can see, the changepoints are well

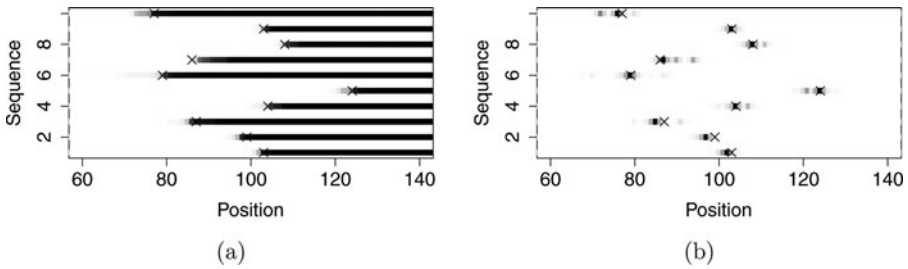

Figure 3. The estimated marginal probabilities of (a) the classification of each observation to segment 2 (probability 1 is black) and (b) a particular position being a changepoint, where the gray scale has been adjusted for each sequence so that the position with the maximum probability is black and the position with the minimum probability is white. The true changepoint locations are marked with ($\times$) in both plots. Note that only positions 60 to 140 are shown in each sequence.

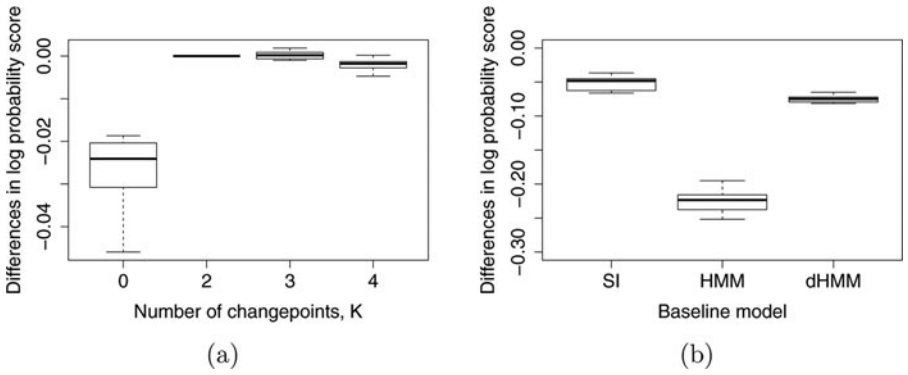

Figure 4. Each boxplot shows the log-probability scores, across the validation sets, for a different model. The y-axis is defined as the log-probability score for (a) other numbers of changepoints and (b) other baseline models, minus the log-probability score for the model with two changepoints.

recovered. In Figure 3(b), we see the marginal probability plot for the positions of the changepoints. It is worth commenting on the fact that the marginal probabilities in Figure 3(b) do not necessarily vary smoothly as a function of location. That is, there are observations that are considered to be unlikely candidates for a changepoint even though the previous and next observations have a high probability of being a changepoint.

4.2 Real Data Analysis: Monsoon Rainfall

To illustrate the applicability of the model to a real-world problem, we analyze the Indian monsoon data described earlier in Section 1. The onset and withdrawal of the annual summer monsoon is of critical importance in India since it directly impacts agricultural production, water resource management, and hydroelectricity production (Lima and Lall 2009). There is no precise definition of the monsoon season, but there is a general understanding that the onset is the time of consistent and substantial increase in rainfall over a regional area and the withdrawal is the time that marks the return to a dry period (Fasullo and Webster 2003; Joseph, Sooraj, and Rajan 2006). In terms of understanding the climatological variability of the monsoon over time, a first step is to label the changepoints of the onset and withdrawal in the historic record (Joseph, Sooraj, and Rajan 2006), enabling (e.g.) prediction of onset and withdrawal as a function of exogenous variables such as large-scale atmospheric quantities (Pai and Nair 2009). Our proposed model provides a framework to not only detect but also to quantify our uncertainty about the estimation of the onset and withdrawal dates and our results can be viewed as an alternative to other approaches that use non-Markov models for this purpose (e.g., Stern 1982; Lima and Lall 2009).

Finite-state Markov models have long been used to characterize daily rainfall occurrence in climatological applications (Gabriel and Neumann 1962; Katz 1974). Rainfall occurrence is typically defined as any amount more than 0.2 mm or 0.01 in. A second cutoff can be used to distinguish light and heavy rainfall events (anything over 20 mm is considered a heavy rainfall event). In our analysis below, the daily rainfall amount is assigned to one of three discrete categories: no/light rainfall with 0–0.2 mm ($y_j = 1$), medium rainfall with 0.2–20 mm ($y_j = 2$), or heavy rainfall with >20 mm ($y_j = 3$) (Katz 1977).

Our dataset consists of daily rainfall measurements over a 31 year period (1979–2010) from a weather station located at latitude 24.65 and longitude 77.32 in the monsoon region of India. (The data were obtained from the U.S. National Centers for Environmental Prediction (NCEP) Climate Prediction Center (CPC) Global Summary of the Day (GSOD) Observations, *http://rda.ucar.edu/datasets/ds512.0*). The data are plotted in Figure 1 with each year plotted as a sequence. Rainfall amounts were categorized into three states as described above and the 153 missing observations (1.35% of the data) were imputed during inference as described in Section S.1 in the supplementary materials. We denote each day of the year as $j = 1, \ldots, 365$ (leap days are removed). We assume that the first and last segments in each year have the same Markov transition matrix, reflecting the fact that the end of one sequence on December 31 is contiguous in time with the start of the next sequence on January 1. This constraint rules out the possibility of a single changepoint in the model, and thus, the number of possible changepoints we can consider is $K = 0, 2, 3, 4, \ldots$.

To determine the optimal number of changepoints using cross-validation, we randomly partitioned our sequences (or years) into 10 training/test sets where nine of the test sets contain three sequences and one has four sequences. Figure 4(a) compares the cross-validated log probability scores of the model with two changepoints, to models with zero, three, and four changepoints. The models with two and three changepoints are very close in performance and outperform models with zero or four changepoints. We select the model with two changepoints ($K = 2$) as our preferred model given that it is simpler and it corresponds to our physical intuition about the monsoon phenomenon (i.e., onset and withdrawal of the monsoon season).

Figure 4(b) compares the model with two changepoints with the baseline models SI, HMM, and dHMM. As we can see, our proposed model significantly outperforms the three baselines.

A comparison of the joint model and the approach based on individual sequences is provided in Section S.2.2 in the supplementary materials.

Figure 5(a) shows the estimates of the marginal probability of each day belonging to the monsoon season, providing a visual interpretation of the estimated interannual variability in the dates of the Indian monsoon onset and withdrawal across 31 years. The marginal distribution across all 31 years is shown at the top

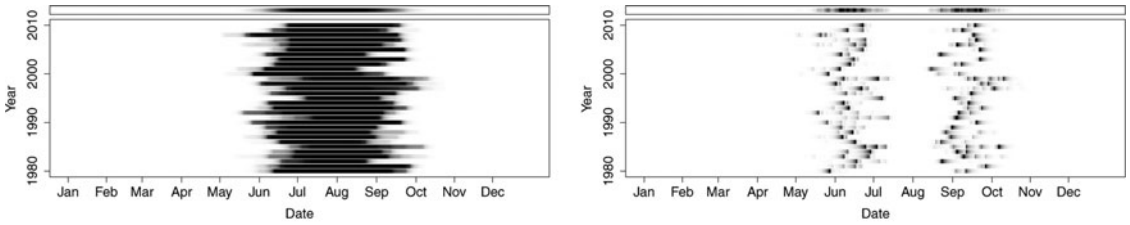

Figure 5. The estimated marginal probability distribution for (a) the classification of each day to the monsoon season (probability 1 is black) and (b) a particular day being a changepoint, where the gray scale has been adjusted for each year so that the day with the maximum probability is black and the day with the minimum probability is white.

of the figure with each of the years plotted per row below. Most days in July and August, across all years, are highly probable to be classified as monsoon days (black) and the days at the beginning and end of the monsoon season have less certainty of being in the monsoon (gray).

The probability of each individual day being a changepoint (onset or withdrawal of the monsoon) can also be obtained from the model. Figure 5(b) shows the probability of a particular day being a changepoint in each of the 31 years. Days that are darker gray are more likely to be the monsoon onset or withdrawal date. The top line shows the marginal over all years. We see that the highest probability of onset is in June and the highest probability of withdrawal is in September.

The estimated transition matrices for our model with two changepoints, the first for the dry (nonmonsoon) season and the second for the monsoon season are as follows:

$$\hat{\boldsymbol{Q}}^{\text{dry}} = \begin{bmatrix} 0.916 & 0.026 & 0.002 \\ 0.742 & 0.232 & 0.027 \\ 0.423 & 0.367 & 0.210 \end{bmatrix},$$

$$\hat{\boldsymbol{Q}}^{\text{monsoon}} = \begin{bmatrix} 0.720 & 0.216 & 0.069 \\ 0.439 & 0.436 & 0.125 \\ 0.258 & 0.489 & 0.253 \end{bmatrix},$$

with rows/columns 1, 2, 3, corresponding to no rainfall, light rainfall, and heavy rainfall categories, respectively. The monsoon and nonmonsoon seasons are distinctly different in terms of their Markov transition probabilities, with transitions into light and heavy rainfall categories being significantly higher in the monsoon season compared to the dry season. The estimates for the changepoint parameters r_1, b_1, r_2, and b_2 are 0.446, 0.810, 0.703, and 0.885, respectively, with 95% credible intervals [0.431, 0.461], [0.550, 0.985], [0.687, 0.719], and [0.693, 0.993], respectively.

Upon examining the model with three changepoints ($K = 3$), we found it added an additional changepoint in the middle of the monsoon season. However, the estimated transition matrices for these two segments did not differ significantly from one another.

Table 1. Computational time per iteration in seconds for two of the datasets discussed in the article, with code in Matlab on a computer with 2.80 GHz CPU

Dataset	$K = 0$	$K = 1$	$K = 2$	$K = 3$
Simulated data 1	0.002	0.015	0.065	0.138
Monsoon data	0.054	0.163	0.432	0.795

Our results are broadly consistent with other findings. For example, Fasullo and Webster (2003) analyzed data from a station further south in India, estimating that the monsoon onset dates at this location range from May 19 to June 20 with the median being June 6, and the withdrawal dates range from August 12 to September 27 with the median being September 4. These dates are consistent with our estimates at a more northerly location, given that the monsoon approaches from the south in India, and thus, we would expect our calculated onset and withdrawal dates to be somewhat later than those of Fasullo and Webster (2003). We obtain a mean estimate of June 13 for the monsoon onset with 95% probability interval (PI) of [May 23, July 4] and the mean date for the monsoon withdrawal is September 13 with an estimated 95% PI of [August 24, October 3]. These estimates are based on the data shown at the top of Figure 5(b) where the 95% PI is computed by summing up the estimated marginal probabilities symmetrically around the mean until a 95% interval is reached.

5. DISCUSSION AND CONCLUSIONS

We introduced a piecewise homogenous Markov chain model where changepoint positions are modeled by a discrete distribution, in particular, a truncated version of the negative binomial distribution. The model is constructed to handle multiple sequences of variable length where each sequence moves between the underlying Markov chains in the same order. We show that the changepoints are well recovered on synthetic data, resulting in accurate estimates for the parameters used to define our model. To illustrate the utility of the model on real-world datasets, we applied the model to daily rainfall data collected for 31 years in northern India. For this dataset, the model is able to detect the onset and withdrawal of the monsoon season, and produced inferences of parameters that were scientifically interpretable. In addition, the model outperformed alternative approaches such as hidden Markov and double Markov models.

Our Bayesian framework allows different sequences to draw strength from each other both when finding the changepoints and for parameter estimation. Estimates of uncertainty about both parameters and latent variables, which arise naturally from our MCMC inference algorithm, provide an appealing interpretation of the uncertainty regarding position of the changepoints. This is of particular interest, for example, in analyzing the onset and withdrawal of the Indian monsoon season. Uncertainty quantification is an increasingly important component of climate data analysis, and the type of Bayesian approach used in this article can provide a useful data-driven alternative to

more traditional methods such as using threshold values (Lima and Lall 2009) or relying entirely on definitions based on prior knowledge (Fasullo and Webster 2003).

Comparisons of the computation times for two of the data examples in the article are shown in Table 1, comparing the times for the smallest dataset (the first simulated dataset) and the largest dataset (the monsoon data). The numbers shown are for one MCMC iteration when fitting the model to the full dataset (we used parallelization for our cross-validation runs). We see that as the number of changepoints increase the method is more computationally expensive, but overall the method is relatively fast even on the larger dataset. Additional speedups could be obtained (e.g.) by parallelizing the analysis across sequences, coding the algorithm in a more efficient language than Matlab, and so on.

There are a number of additional extensions to the model that may be potentially useful to explore. For example, a useful direction would be the development of conditional models $p(\boldsymbol{y}|\boldsymbol{x})$, where each sequence $\boldsymbol{y} = (y_1, \ldots, y_T)$ is accompanied by a sequence $\boldsymbol{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T)$ of exogenous variables and where each $\boldsymbol{x}_t$ could be multivariate, $t = 1 \ldots, T$. The exogenous variables $\boldsymbol{x}$ could influence the $\boldsymbol{y}$'s directly and/or the locations of the changepoints. Another direction would be models for handling high-dimensional and/or real-valued observation sequences, where one could assume the existence of a categorical latent sequence $\boldsymbol{z} = (z_1, \ldots, z_T)$ as a low-dimensional representation of the observed sequence data $\boldsymbol{y}$, extending the general approach presented here to piecewise homogenous hidden Markov chains.

SUPPLEMENTARY MATERIALS

Additional supporting information: A document containing additional results, simulations, sensitivity analysis, etc. (pdf file).

Software implementation of the changepoint Markov model: The Matlab code for implementing the changepoint Markov model and the simulated datasets used in the article (GNU zipped tar file).

ACKNOWLEDGMENTS

This work was supported in part by U.S. Department of Energy award DOE-SC0006619 (TH and PS), U.S. Office of Naval Research under MURI grant N00014-08-1-1015 (TH and PS), and U.S. National Science Foundation award IIS-1320527 and a Google Faculty Award (PS).

[Received August 2014. Revised April 2015.]

REFERENCES

Berchtold, A. (1999), "The Double Markov Chain," *Communications in Statistics—Theory and Methods*, 28, 2569–2589. [205,206,208]

——— (2002), "High-Order Extensions of the Double Chain Markov Model," *Stochastic Models*, 18, 193–227. [205]

Chen, J., Brissette, F., and Leconte, R. (2010), "A Daily Stochastic Weather Generator for Preserving Low-Frequency of Climate Variability," *Journal of Hydrology*, 388, 480–490. [205]

Cleynen, A., Koskas, M., Lebarbier, E., Rigaill, G., and Robin, S. (2014), "Segmentor3IsBack: An R Package for the Fast and Exact Segmentation of Seq-Data," *Algorithms for Molecular Biology*, 9, 1–11. [206,208]

Cleynen, A., and Lebarbier, E. (2014), "Segmentation of the Poisson and Negative Binomial Rate Models: A Penalized Estimator," *ESAIM: Probability and Statistics*, 18, 750–769. [206,208]

Cochran, W. (1938), "An Extension of Gold's Method of Examining the Apparent Persistence of One Type of Weather," *Quarterly Journal of the Royal Meteorological Society*, 64, 631–634. [205]

Craig, W. (1943), "The Song of the Wood Pewee," Ph.D. dissertation, University of the State of New York, Albany. [205]

Durbin, R., Eddy, S., Krogh, A., and Mitchison, G. (1998), *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*, Cambridge, UK: Cambridge University Press. [205]

Fasullo, J., and Webster, P. J. (2003), "A Hydrological Definition of Indian Monsoon Onset and Withdrawal," *Journal of Climate*, 16, 3200–3211. [210,211]

Fearnhead, P. (2006), "Exact and Efficient Inference for Multiple Changepoint Problems," *Statistics and Computing*, 16, 203–213. [209]

Fearnhead, P., and Liu, Z. (2007), "Online Inference for Multiple Changepoint Problems," *Journal of the Royal Statistical Society*, Series B, 69, 589–605. [205]

——— (2011), "Efficient Bayesian Analysis of Multiple Changepoint Models With Dependence Across Segments," *Statistics and Computing*, 21, 217–229. [206,208]

Fitzpatrick, M., and Marchev, D. (2013), "Efficient Bayesian Estimation of the Multivariate Double Chain Markov Model," *Statistics and Computing*, 23, 467–480. [206,208]

Gabriel, K., and Neumann, J. (1962), "A Markov Chain Model for Daily Rainfall Occurrence at Tel Aviv," *Quarterly Journal of the Royal Meteorological Society*, 88, 90–95. [205,210]

Gelman, A., Hwang, J., and Vehtari, A. (2014), "Understanding Predictive Information Criteria for Bayesian Models," *Statistics and Computing*, 24, 997–1016. [208]

Gneiting, T., and Raftery, A. E. (2007), "Strictly Proper Scoring Rules, Prediction, and Estimation," *Journal of the American Statistical Association*, 102, 359–378. [208]

Gold, E. (1929), "Note on the Frequency of Occurrence of Sequences in a Series of Events of Two Types," *Quarterly Journal of the Royal Meteorological Society*, 55, 307–309. [205]

Green, P. J. (1995), "Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination," *Biometrika*, 82, 711–732. [208]

Guédon, Y. (2003), "Estimating Hidden Semi-Markov Chains From Discrete Sequences," *Journal of Computational and Graphical Statistics*, 12, 604–639. [206,208]

Guédon, Y., Barthelemy, D., Caraglio, Y., and Costes, E. (2001), "Pattern Analysis in Branching and Axillary Flowering Sequences," *Journal of Theoretical Biology*, 212, 481–520. [205]

Guttorp, P. (1995), *Stochastic Modeling of Scientific Data*, Boca Raton, FL: CRC Press. [205]

Joseph, P. V., Eischeid, J. K., and Pyle, R. J. (1994), "Interannual Variability of the Onset of the Indian Summer Monsoon and its Association With Atmospheric Features, El Niño, and Sea Surface Temperature Anomalies," *Journal of Climate*, 7, 81–105. [205]

Joseph, P. V., Sooraj, K. P., and Rajan, C. K. (2006), "The Summer Monsoon Onset Process Over South Asia and an Objective Method for the Date of Monsoon Onset Over Kerala," *International Journal of Climatology*, 26, 1871–1893. [210]

Katz, R. W. (1974), "Computing Probabilities Associated With the Markov Chain Model for Precipitation," *Journal of Applied Meteorology*, 13, 953–954. [210]

——— (1977), "Precipitation as a Chain-Dependent Process," *Journal of Applied Meteorology*, 16, 671–676. [210]

Lai, T. L., and Xing, H. (2011), "A Simple Bayesian Approach to Multiple Change-Points," *Statistica Sinica*, 21, 539–569. [206]

Levin, E., and Pieraccini, R. (2000), "A Stochastic Model of Human-Machine Interaction for Learning Dialog Strategies," in *IEEE Transactions on Speech and Audio Processing*, pp. 11–23. [205]

Li, L., Qiu, S., Zhang, B., and Feng, C. X. (2015), "Approximating Cross-Validatory Predictive Evaluation in Bayesian Latent Variables Models With Integrated IS and WAIC," 1–17. [208]

Lima, C. H. R., and Lall, U. (2009), "Hierarchical Bayesian Modeling of Multisite Daily Rainfall Occurrence: Rainy Season Onset, Peak, and End," *Water Resources Research*, 45, 1–14. [210,212]

Luong, T.-M., Rozenholc, Y., and Nuel, G. (2013), "Fast Estimation of Posterior Probabilities in Change-Point Models Through a Constrained Hidden Markov Model," *Computational Statistics and Data Analysis*, 68, 129–140. [206,208]

Neal, R. M. (2000), "Markov Chain Sampling Methods for Dirichlet Process Mixture Models," *Journal of Computational and Graphical Statistics*, 9, 249–265. [208]

Newnham, E. (1916), "The Persistence of Wet and Dry Weather," *Quarterly Journal of the Royal Meteorological Society*, 42, 153–162. [205]

Pai, D. S., and Nair, R. M. (2009), "Summer Monsoon Onset Over Kerala: New Definition and Prediction," *Journal of Earth System Science*, 118, 123–135. [210]

Polansky, A. M. (2007), "Detecting Change-Points in Markov Chains," *Computational Statistics and Data Analysis*, 51, 6013–6026. [206,208]

Rabiner, L. (1989), "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proceedings of the IEEE*, 77, 257–285. [205,206]

Raftery, A., and Tavaré, S. (1994), "Estimation and Modelling Repeated Patterns in High Order Markov Chains With the Mixture Transition Distribution Model," *Applied Statistics*, 43, 179–199. [205]

Rigaill, G., Lebarbier, E., and Robin, S. (2012), "Exact Posterior Distributions and Model Selection Criteria for Multiple Change-Point Detection Problems," *Statistics and Computing*, 22, 917–929. [206,208]

Stern, R. D. (1982), "Computing a Probability Distribution for the Start of the Rains From a Markov Chain Model for Precipitation," *Journal of Applied Meteorology*, 21, 420–423. [210]

Vehtari, A., and Lampinen, J. (2002), "Bayesian Model Assessment and Comparison Using Cross-Validation Predictive Densities," *Neural Computation*, 14, 2439–2468. [208]

Wilks, D., and Wilby, R. (1999), "The Weather Generation Game: A Review of Stochastic Weather Models," *Progress in Physical Geography*, 23, 329–357. [205]

Xing, H., Sun, N., and Chen, Y. (2012), "Credit Rating Dynamics in the Presence of Unknown Structural Breaks," *Journal of Banking & Finance*, 36, 78–89. [206,208]

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. [206,208]

——— (2012), "Model Selection for High-Dimensional, Multi-Sequence Change-Point Problems," *Statistica Sinica*, 22, 1507–1538. [206,208]