Statistica Sinica **25** (2015), 677-708
doi:http://dx.doi.org/10.5705/ss.2012.308

CLUSTERING CURVES BASED ON CHANGE POINT ANALYSIS : A NONPARAMETRIC BAYESIAN APPROACH

Sarat C. Dass[1], Chae Young Lim[2], Tapabrata Maiti[2] and Zhen Zhang[3]

[1] *Universiti Teknologi PETRONAS,* [2] *Michigan State University and* [3] *University of Chicago*

Abstract: Statistical methods for analyzing disease incidence and mortality data over time and geographical regions have gained considerable interest in recent years due to increasing concerns of public health, health disparity and legitimate resource allocation. Trend analysis of cancer incidence and mortality rates is essential for subsequent public health investigations. For example, the National Cancer Institute provides software for fitting statistical models to track changes in cancer curves. Currently available models for detecting trend changes over time are designed for a single curve. When multiple curves are available, current methods could be applied multiple times, however, this may not be efficient in the statistical sense. This paper proposes a statistical model that allows concurrent change-point estimation and grouping for multiple curves while maintaining local variabilities. The Bayesian analysis is carried out by eliciting a Dirichlet process prior on the relevant functional space to model change-points. Improper priors are elicited and the resulting posterior is shown to be valid and proper. The age-adjusted lung cancer mortality rates of U.S. states are analyzed to detect change-points and rates of change as well as clusters of states that share similar trends over time. The procedure is also compared with an approach that group states according to a penalized likelihood criterion.

Key words and phrases: Age-adjusted mortality rates, Bayesian non-parametrics, change-point analysis, clustering, conditional autoregressive Model (CAR), Dirichlet process priors, Markov Chain Monte Carlo methods.

1. Introduction

Cancer is a leading cause of mortality in the United States. The American Cancer Society (ACS, `www.cancer.org`) provides such information on cancer as time trends of age-adjusted cancer morality rates for different cancer types, for different sub-populations defined by geographic and socio-demographic characteristics. It is a fact that there was an increased number of cancer deaths in 2007 as a result of aging and growth of the US population (ACS (2010)). Moreover, the impact of cancer surveillance is not uniformly effective over different U.S. states.

Authors' names are in alphabetical order.

Statistical methods for analyzing disease incidence or mortality rates over time have gained considerable interest in recent years due to increasing concerns of public health, health disparity, and legitimate resource allocation. One important question is whether the trends before and after a change-point are significantly different (statistically speaking) from each other. Several such models (called joinpoint models) have been developed (see, Carlin, Gelfand, and Smith (1992), Kim et al. (2000, 2004), Tiwari et al. (2005) and Ghosh, Ghosh, and Tiwari (2011)) for detecting time points associated with significant changes in the disease trend. The models developed by Kim et al. (2000, 2004), for example, are implemented as software for the National Cancer Institute (NCI) (Ries et al. (2002)). These models are used to fit a single curve of disease rates for detecting joinpoints over time. The joinpoint models assume piecewise linear regression functions connected at the joinpoints. In contrast, we consider piecewise linear regression functions that do not have to be connected at the change-points. We point out that although joinpoint and change-point methods are technically different, they can be applied in the same context for trend analysis. Thus, we consider only change-point detection and clustering, instead of joinpoints, but note that the proposed method can be modified to a joinpoint setting.

When multiple cancer curves are available for trend comparisons, a natural interest is whether there are groups (or, clusters) of curves with similar change-points but with significant variations between and within groups. For example, Figure 1 gives age-adjusted lung cancer mortality rates from 1969 to 2006 for all 48 contiguous states in continental United States (excluding Alaska and Hawaii) and Washington D.C. It is evident that some states such as Ohio and Pennsylvania (red in Figure 1), show similar change-points and rates of change in each time interval. On the other hand, some other states such as Missouri and Iowa (blue in Figure 1, share similar change-points and rates of change but these attributes are different from those for Ohio and Pennsylvania. For a better view, we provide a separate plot for these four states in Figure 2. Figure 1 also demonstrates that each state has different levels of variability over time around the mean trend. The presence of such heterogeneities (different change-points, rates of change and local variabilities) among states is not reflected in the US data. observed in Missouri is not revealed in the US lung cancer mortality curve shown in the lower left corner of Figure 1.

We are interested in developing a model that concurrently detects change-points in trends and clusters multiple curves by similar trends. Such a model may help administrators identify sub-populations (generally a set of states) that are affected by changes (increase or decrease) in risk so that unified surveillance for the group can be done for the prevention of cancer. Concurrent estimation of grouping and detection of change-points cannot be done using existing change-points or joinpoint methodologies that are developed for a single curve analysis.

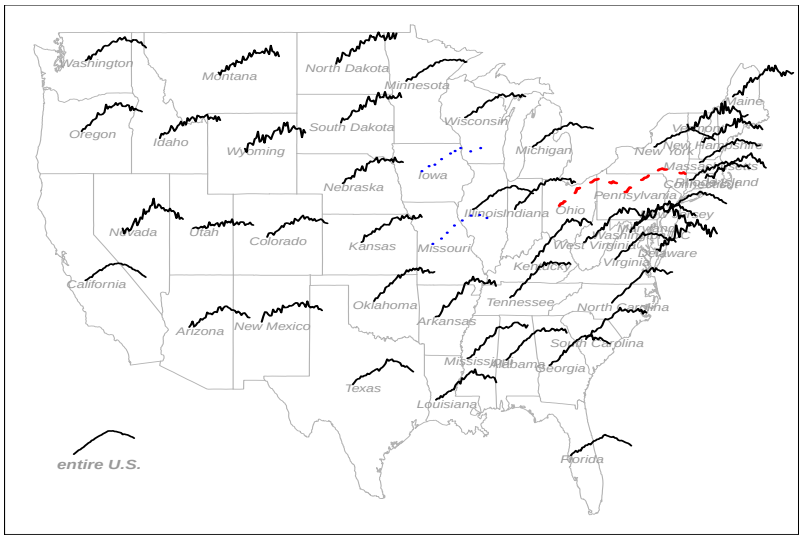

Figure 1. Age-adjusted mortality rates of lung cancer from 1969 to 2006 for all 48 contiguous states, Washington D.C., and the entire US.

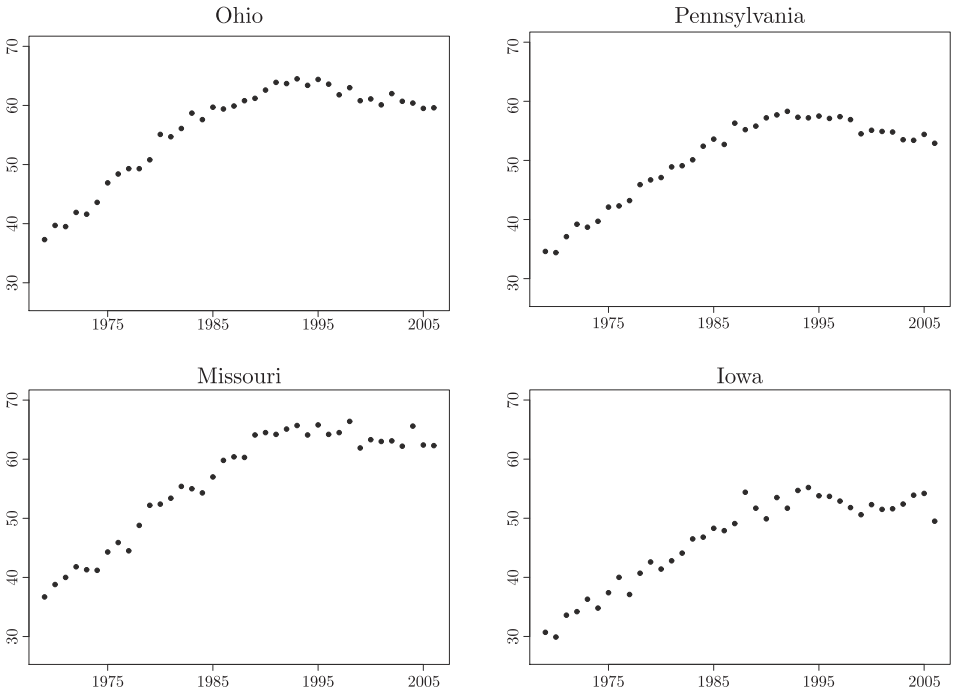

Figure 2. Age-adjusted incidence rates of lung cancer from 1969 to 2006 for four states: Ohio, Pennsylvania, Missouri, Iowa.

One could apply an existing change-point method designed for a single curve to each curve and subsequently apply a clustering algorithm, but existing clustering algorithms are not directly applicable since we want to allow a different number of change-points for each curve. The difficulty here is to come up with a similarity measure since the dimension of feature spaces (of the change-points and rate of changes) can be different between curves. We introduce a similarity measure based on penalized likelihood to account for the varying dimensions and compare the clustering results with the approach we propose here. The comparison is discussed in detail in Section 5.1.

We assume that the mean trend of each curve is piece-wise linear over time. We make use of the Dirichlet Process (DP) methodology (Ferguson (1973, 1974)) in an innovative way to cluster these piecewise linear trends of the different curves. Since change-points are determined by the rates of change in each time interval, the DP prior is developed on the space of piecewise constant functions where each constant level represents the slope of the linear trend in the corresponding time interval. The locations of change-points and the rates of change are random and estimated during the inferential stage.

The rest of the paper is organized as follows. Section 2 presents the proposed change-point model and prior specifications for Bayesian inference. Some prior components are taken to be improper, and therefore, propriety of the posterior is a concern. Section 3 establishes propriety of the posterior distribution under mild conditions. Section 4 gives details of the Bayesian inference: performance measures of change-points and clustering configuration for the comparison. Section 5 present results when the proposed model is applied to simulated data as well as to lung cancer data. We also compare the results of the proposed model with the penalized likelihood approach introduced as an alternative approach. Section 6 gives a summary and discussion for the proposed model and future research.

2. A Change Point Model

The description of characteristics of cancer curves are given first, which motivates the proposed model. Cancer data are obtained from the Surveillance, Epidemiology, and End Results (SEER) program (`seer.cancer.gov`) of the NCI. An age-adjusted mortality rate is the primary measure for monitoring cancer trends over time and over states; it is a weighted average of the age-specific (crude) rates with the proportions of the reference population in the corresponding age groups as weights so that the potential confounding effect of age is reduced. See the SEER program website or Dass, Lim, and Maiti (2011) about this.

We consider age-adjusted lung cancer mortality rates from 1969 to 2006 for the 48 contiguous states (excluding Alaska and Hawaii) and Washington D.C.

(See Figure 1). Thus, mortality rates are observed at 38 time points over 49 locations. For simplicity, we re-index the time as $t = 1, \ldots, 38$. It is clear from Figure 1 that there exists at least one change-point for the most of states. Some states show similar change-points even though the variability around the mean curve differs. To model exponential growth or decay of the age-adjusted rates, we model the logarithm of the age-adjusted rates as a linear function of time, the usual practice in joinpoint analysis literature (see e.g., Kim et al. (2000, 2004), Tiwari et al. (2005), Ghosh, Basu, and Tiwari (2009); Ghosh, Ghosh, and Tiwari (2011)).

2.1. Model specification

Let $Y_{s,t}$ be the logarithm of the age-adjusted mortality rate for state (or, site) s and time t, where $s = 1, 2, \ldots, 49 = N$, and $t = 1, \ldots, 38 = n$. For site s, suppose there are K_s change-points. Then we have $K_s + 1$ time intervals, $[T_0^{(s)}, T_1^{(s)}), [T_1^{(s)}, T_2^{(s)}), \ldots, [T_{K_s-1}^{(s)}, T_{K_s}^{(s)}),\ [T_{K_s}^{(s)}, T_{K_s+1}^{(s)}]$ that form a partition of $\mathcal{T} \equiv \{1, \ldots, n\}$ with $T_0^{(s)} = 1$ and $T_{K_s+1}^{(s)} = n$. For simplicity, we use $[T_{l-1}^{(s)}, T_l^{(s)})$ for the lth time interval for all $l = 1, \ldots, K_s + 1$, with the last interval to include the right-end point. For $t \in [T_{l-1}^{(s)}, T_l^{(s)})$, we consider the model:

$$Y_{s,t} = \alpha_l^{(s)} + t \cdot \beta_l^{(s)} + \epsilon_{s,t}, \tag{2.1}$$

where $l = 1, \ldots, K_s + 1$. We further assume that $\epsilon_{s,t} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_s^2)$.

The model (2.1) assumes K_s change-points for site s and the unknown trend between change-points to be linear. We do not assume that the line segments are connected at the change-points although we observe in practice that the discontinuities between adjacent fitted line segments are negligible. The parameters $\alpha_l^{(s)}$ and $\beta_l^{(s)}$ are the intercept and slope, respectively, for the lth time segment, $l = 1, \ldots, K_s + 1$, for site s. $\epsilon_{s,t}$ represents i.i.d. errors over and above the mean trend $\alpha_l^{(s)} + t \cdot \beta_l^{(s)}$. Since the variability of $Y_{s,t}$ around the mean curves can be different for the different sites, we take $\mathrm{Var}\,(\epsilon_{s,t}) = \sigma_s^2$, the site-specific variance.

2.2. Functional DP prior and other prior specifications

To model clustering of N sites with respect to their change-point locations and corresponding magnitudes, we consider the piecewise constant function,

$$\boldsymbol{\theta}_s(t) = \beta_l^{(s)} \quad \text{if} \quad T_{l-1}^{(s)} \le t \le T_l^{(s)} - 1, \tag{2.2}$$

where $\beta_l^{(s)}$ is the true but unknown slope in the interval $[T_{l-1}^{(s)}, T_l^{(s)})$ for each site s. The piecewise constant function $\boldsymbol{\theta}_s$ contains all information on the number of

change-points, K_s, the locations of change-points, $T_l^{(s)}$, and the slopes (rates of change), $\beta_l^{(s)}$, for the corresponding time segments.

We denote the space of all piecewise constant functions on $\mathcal{T}$ by $\boldsymbol{\Theta}$. The generic element $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ has the form

$$\boldsymbol{\theta}(t) = \beta_l \qquad \text{if} \qquad T_{l-1} \leq t \leq T_l - 1, \tag{2.3}$$

for $l = 1, \ldots, k+1$ with $1 \equiv T_0 < T_1 < \cdots T_k < T_{k+1} \equiv n$. The set of all probability distributions on $\boldsymbol{\Theta}$ is denoted by $\mathcal{P}(\boldsymbol{\Theta})$. The Dirichlet process (DP) prior developed subsequently on $\mathcal{P}(\boldsymbol{\Theta})$ will enable us to cluster the N functions $\boldsymbol{\theta}_s \in \boldsymbol{\Theta}$ related to the sites $s = 1, \ldots, N$.

The traditional DP $\equiv \text{DP}(\alpha_0 G_0)$ depends on two hyper-parameters, α_0 and G_0. G_0 is the baseline (or centering) distribution on $\boldsymbol{\Theta}$ and $\alpha_0 > 0$ is the precision parameter that controls variability around the centering distribution. A randomly generated distribution F from $DP(\alpha_0 G_0)$ is almost surely discrete and admits the representation

$$F = \sum_{i=1}^{\infty} \omega_i \, \delta_{\theta_i}, \tag{2.4}$$

where δ_z denotes a point mass at z, $\omega_1 = \eta_1$, $\omega_i = \eta_i \prod_{k=1}^{i-1}(1-\eta_k)$, for $i = 2, 3, \ldots$ with $\eta_1, \eta_2, \ldots,$ iid $Beta(1, \alpha_0)$ random variables and $\theta_1, \theta_2 \ldots$ i.i.d. from G_0 (Sethuraman (1994)). In the traditional DP formulation, θ_i is assumed to be scalar or vector-valued taking values in R^p. The functional DP as a prior on $\mathcal{P}(\boldsymbol{\Theta})$ conceptually extends the θ_is in (2.4) to piecewise constant functions $\boldsymbol{\theta}_i$ for $i \geq 1$ where each $\boldsymbol{\theta}_i \in \boldsymbol{\Theta}$. Clearly, the random $F \in \mathcal{P}(\boldsymbol{\Theta})$ and the functional DP is a prior on $\mathcal{P}(\boldsymbol{\Theta})$ with centering measure $G_0 \in \mathcal{P}(\boldsymbol{\Theta})$.

Clustering via functional DP is achieved in the same way as in traditional DP. Assuming $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_N$ iid from F and $F \sim DP(\alpha_0 \, G_0)$, we note that, conditional on F, there is a positive probability that each $\boldsymbol{\theta}_s$ will be equal to one of the functions drawn previously since F is discrete. In fact, marginalizing over F, the well-known sequence of conditional distributions given by the Polya urn scheme, $(\boldsymbol{\theta}_s \,|\, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{s-1}) = (G_0(d\boldsymbol{\theta}_s) + \sum_{s'=1}^{s-1} \delta_{\boldsymbol{\theta}_{s'}})/s$ for $s = 1, \ldots, N$, explicitly describes the a-priori clustering distribution: The function $\boldsymbol{\theta}_s$ is either a new piecewise continuous function generated from G_0 with probability $1/s$ or one of the previously generated functions $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_{s-1}$ with probability $(s-1)/s$, for $s = 1, \ldots, N$.

Since we want to cluster curves based on the number and locations of change-points and the slopes, we do not include $\alpha_l^{(s)}$ in the definition of $\boldsymbol{\theta}_s$. Here, the parameters $\alpha_l^{(s)}$ are considered as nuisance parameters for clustering. The proposed methodology enables clustering of piecewise linear curves that have

similar slopes but different intercepts. interested in clustering curves that have similar slopes as well as intercepts, one can extend the definition of $\boldsymbol{\theta}_s$ to include $\alpha_l^{(s)}$ as well. The definition of the space $\boldsymbol{\Theta}$ will change accordingly and the functional DP prior will be defined on a new $\mathcal{P}(\boldsymbol{\Theta})$.

To complete the functional DP prior specification for the proposed model, we introduce a specification of the baseline distribution G_0. The distribution G_0 on $\boldsymbol{\Theta}$ is described in hierarchical fashion.

(i) Distribution on the number of change-points, K: Let K follow a truncated Poisson distribution. We assume that each time interval has at least $w > 0$ units to avoid a zero-length interval. Then, $K \leq k^*$ where

$$k^* \equiv \left[\frac{(n-1)}{w}\right] - 1 \tag{2.5}$$

to ensure $n_0 \equiv n - 1 - (K+1)w > 0$ with probability 1. The corresponding probability when $K = k$ is given by $p(k) = (e^{-\lambda}\lambda^k/k!)/(\sum_{l=0}^{k^*} e^{-\lambda}\lambda^l/l!) = (\lambda^k/k!)/(\sum_{l=0}^{k^*} \lambda^l/l!)$ for $k = 0, 1, \ldots, k^*$ and $E(K) = \lambda(1 - (\lambda^{k^*}/k^*!)/(\sum_{l=0}^{k^*} \lambda^l/l!))$.

(ii) Distribution on the time intervals between change-points: Let n_l be the interval length for the lth time segment after subtracting w. We assume that, conditional on $K = k$,

$$(n_1, \ldots, n_{k+1}) \,\big|\, K = k \;\sim\; \text{Multinomial}\left(n_0, \frac{1}{k+1}, \ldots, \frac{1}{k+1}\right).$$

The times $\{\, T_l,\, l = 1, \ldots, k \,\}$ in (2.3) are recursively obtained as $T_0 = 1$, and $T_l = n_l + T_{l-1} + w$ for $l = 1, \ldots, k$.

(iii) Distribution on the constant levels β_l: Given $K = k$, $\{\, \beta_l,\, l = 1, \ldots, k+1 \,\}$ are generated from the probability density function π_0 on R independently of each other.

(iv) Set $\boldsymbol{\theta}(t)$ according to (2.3) based on the random quantities generated in (i)−(iii).

It then follows that the infinitesimal measure for G_0 is

$$G_0(d\boldsymbol{\theta}) = p(k)\left(\frac{\Gamma(n_0+1)}{\prod_{i=1}^{k}\Gamma(n_i+1)}\left(\frac{1}{k+1}\right)^{n_0}\right)\prod_{l=1}^{k+1}\pi_0(\beta_l)\,d\beta_l, \tag{2.6}$$

where $\boldsymbol{\theta}$ is a randomly generated piecewise constant function. We take

$$\pi_0(\beta_l) \propto 1, \tag{2.7}$$

independently for all $l = 1, \ldots, k+1$. The proposed DP prior involves the hyper-parameters, α_0 and λ. For Bayesian inference, their priors are taken to be π_1 and π_2, respectively with

$$\pi_1(\alpha_0) = \texttt{gamma}(\alpha_0 \,|\, a_{\alpha_0}, b_{\alpha_0}) \text{ and } \pi_2(\lambda) = \texttt{gamma}(\lambda \,|, a_\lambda, b_\lambda). \tag{2.8}$$

Next, we assign priors to model parameters and hyperparameters that are not involved in the functional DP-based clustering. These are the collection of all intercepts $\boldsymbol{\alpha} = (\boldsymbol{\alpha}^{(s)},\, s = 1, \ldots, N)$, where $\boldsymbol{\alpha}^{(s)} = (\alpha_1^{(s)}, \alpha_2^{(s)}, \ldots, \alpha_{K_s+1}^{(s)})^T$, and all site-specific variance parameters $\boldsymbol{\sigma} = (\sigma_1^2, \ldots, \sigma_N^2)$. The prior on $\boldsymbol{\alpha}$ is taken as

$$\pi_3(\boldsymbol{\alpha}) = \prod_{s=1}^{N} \pi_0(\boldsymbol{\alpha}^{(s)} | K_s) = \prod_{s=1}^{N} \prod_{l=1}^{K_s+1} \pi_0(\alpha_l^{(s)}) \tag{2.9}$$

independently for $l = 1, \ldots, K_s + 1$ and $s = 1, \ldots, N$, where π_0 is as defined in (2.7). The site-specific variance parameters $\boldsymbol{\sigma}$ are given independent priors with

$$\pi_4(\boldsymbol{\sigma}) = \prod_{s=1}^{N} \texttt{igamma}(\sigma_s^2 \,|\, a_\sigma, b_\sigma), \tag{2.10}$$

where $\texttt{igamma}(x \,|, a, b)$ is the inverse Gamma probability density function with shape parameter a and scale parameter $1/b$.

Bayesian inference is obtained by implementing the Gibbs sampler for all the unknown parameters involved (see Section 4). The Gamma and inverse Gamma distributions are used here for their conjugacy with the appropriate likelihood components during the Gibbs updating steps. The flat priors on β_l and $\alpha_l^{(s)}$ provide analytical simplifications (and hence, computational efficiency) when calculating their posterior conditional distributions; with them, we are able to obtain explicit expressions for the conditional probability that $\boldsymbol{\theta}_s$ belongs to a new cluster (see the expression $H(n_1, \ldots, n_{k+1})$ in Appendix B, for example). The hyper-parameters for the Gamma and inverse Gamma distributions are chosen to have large variance so that the impact of the prior input is minimal. The specific choices of the hyper-parameters for the various Gamma and inverse Gamma distributions in (2.8)−(2.10) are given in Section 5.

3. Propriety of the Posterior Distribution

Let $\underline{\boldsymbol{\theta}} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N)$ denote the collection of all piecewise constant functions for $s = 1, \ldots, N$. Let $\boldsymbol{K} = (K_1, \ldots, K_N)$ denote the number of change-points for all sites $s = 1, \ldots, N$. We denote by $\boldsymbol{T} = (\boldsymbol{T}^{(s)},\, s = 1, \ldots, N)$ with $\boldsymbol{T}^{(s)} = (T_1^{(s)}, \ldots, T_{K_s}^{(s)})$, to be all the change-point times, and $\boldsymbol{\beta} = (\boldsymbol{\beta}^{(s)},\, s = 1, \ldots, N)$ to denote the collection of all slope parameters with $\boldsymbol{\beta}^{(s)} = (\beta_1^{(s)}, \ldots, \beta_{K_s+1}^{(s)})$. Each

$\boldsymbol{\theta}_s \equiv (\boldsymbol{\beta}^{(s)}, K_s, \boldsymbol{T}^{(s)})$, and therefore, we have $\underline{\boldsymbol{\theta}} \equiv (\boldsymbol{\beta}, \boldsymbol{K}, \boldsymbol{T})$. All observed data is denoted by the vector $\boldsymbol{Y} = (\boldsymbol{Y}^{(s)},\, s = 1, \ldots, N)$ with $\boldsymbol{Y}^{(s)} = (Y_{s,1}, \ldots, Y_{s,n})^T$. We also fix the hyperparameters α_0 and λ involved in the elicitation of G_0 for the moment. The collection of all unknown parameters to be inferred is given by $(\underline{\boldsymbol{\theta}}, \boldsymbol{\alpha}, \boldsymbol{\sigma})$, or equivalently, $(\boldsymbol{\beta}, \boldsymbol{K}, \boldsymbol{T}, \boldsymbol{\alpha}, \boldsymbol{\sigma})$. The priors on $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ are improper according to (2.7) and (2.9). Thus, the propriety of the posterior $\pi(\underline{\boldsymbol{\theta}}, \boldsymbol{\alpha}, \boldsymbol{\sigma} \mid \boldsymbol{Y})$ has to be established before inference can be made from it.

Theorem 1. *Fix α_0 and λ. If the minimum number of observations in each time segment, w, is at least 3, then the posterior $\pi(\underline{\boldsymbol{\theta}}, \boldsymbol{\alpha}, \boldsymbol{\sigma} \mid \boldsymbol{Y})$, resulting from the prior specifications in (2.7) and (2.9)−(2.10), is proper.*

The proof is provided in Appendix A.

Remark 1. By setting $w \geq 3$, we provide an upper bound for the number of change-points for each site s: $K_s \leq k^* \leq [(n-1)/3] - 1$ in (2.5). Concurrently, we also ensure at least three observations in each time segment between change-points. Since the priors on $\beta_l^{(s)}$ and $\alpha_l^{(s)}$ are improper, at least one extra observation in each time interval ensures that the marginal distribution (after integrating out $\alpha_l^{(s)}$ and $\beta_l^{(s)}$) is still finite. The technical details are presented in Appendix A.

Remark 2. By extending the proof in Appendix A, Theorem 1 is easily established for the posterior of $(\underline{\boldsymbol{\theta}}, \boldsymbol{\alpha}, \boldsymbol{\sigma}, \alpha_0, \lambda)$ when the priors π_1 on α_0 and π_2 on λ are proper as in (2.8). An improper prior for the precision parameter, α_0, results in an improper posterior; this is also shown in Appendix A.

4. Bayesian Inference

The validity of the Gibbs updating scheme for $\boldsymbol{\theta}_s$ based on the elicited improper prior components in G_0 as well as for $\boldsymbol{\alpha}$ is established in Theorem 2 in Appendix A. The detailed description of the updating steps for the Bayesian computation is given in Appendix B. In this section, we focus on the posterior analysis for model parameters and clustering configuration. Suppose that we have B Gibbs samples, $(\underline{\boldsymbol{\theta}}^{(b)}, \boldsymbol{\alpha}^{(b)}, \boldsymbol{\sigma}^{(b)}, \alpha_0^{(b)}, \lambda^{(b)})$, for $b = 1, \ldots, B$ after convergence is established. Marginal posterior inference can be carried out for each of these components. For example, to infer $\boldsymbol{\theta}_s(t)$ for a particular site s and time point t, we extract all $\boldsymbol{\theta}_s(t)$ components from each $(\underline{\boldsymbol{\theta}}^{(b)}, \boldsymbol{\alpha}^{(b)}, \boldsymbol{\sigma}^{(b)}, \alpha_0^{(b)}, \lambda^{(b)})$, for $b = 1, \ldots, B$. The B realizations of $\boldsymbol{\theta}_s(t)$ are then used to compute the posterior mean, variance and credible interval. A similar procedure also works for d, the number of distinct clusters, where we can obtain marginal probabilities of d over non-negative integers up to N.

The posterior samples provide the frequency distribution of a number of change-points and all possible combinations of time segments between change-points for each site s. Thus, we can obtain the marginal probability distribution for each $\boldsymbol{T}^{(s)}$ so that $\boldsymbol{T}^{(s)}$ along with K_s is estimated using its posterior mode for each site s. Furthermore, the magnitude of changes can be measured by the difference of two adjacent slopes around the change-point, defined as $\Delta_l^{(s)} = \beta_{l+1}^{(s)} - \beta_l^{(s)}, l = 1, \ldots, K_s$ for site s. Under the same patterns of change-points, larger value of $\Delta_l^{(s)}$ indicates a clear change-point. The associated credible set of $\Delta_l^{(s)}$ provides the amount of uncertainty in such change. On the other hand, the posterior samples also provide the frequency distribution of change-points for each site, which is used to compute the empirical distribution of change-points, say, $p_E(t)$.

A challenging inference problem is to obtain results for the clustering tendencies, for example the "average" clusters from the posterior samples. Since clustering configuration is changing at each Gibbs iteration, posterior analysis (obtaining summary quantity) for clustering configuration is not straightforward. We are considering a distance measure based on how many times each pair of sites belongs to the same cluster and use it for clustering: For every pair of sites (s_1, s_2) in $\{1, \ldots, N\}$, define $\text{dist}_b(s_1, s_2) = 1$ if s_1 and s_2 belong to the same cluster in the bth iteration and 0, otherwise, for $b = 1, \ldots, B$. Subsequently, we construct the average distance measure between the sites s_1 and s_2 using

$$dist(s_1, s_2) = 1 - \sum_{b=1}^{B} \frac{\text{dist}_b(s_1, s_2)}{B}.$$

Based on $dist$, an agglomerative clustering algorithm is performed with $\hat{d}$, the posterior mode of the number of distinct clusters as the threshold for the maximum number of clusters in the algorithm. The clustering configuration from this procedure is matched with our expected scenario. However, posterior estimates of change-points of curves in each cluster of the averaged cluster configuration are not necessarily the same since posterior estimates of change-points are obtained by marginal posterior distribution for each curve.

The proposed model is developed to detect change-points of multiple cancer curves simultaneously, and consequently to allow curve clustering based on locations and magnitudes of change-points. Thus, we consider the following measures for model assessment and comparison.

1. *A measure for comparing change-points*: For the simulation data, we know the true locations of change-points. We define an accuracy rate R_a to assess how good is the change-points detection. Let $R_m(s, b)$ for the bth iteration be

$$\Big[\sum_{l=1}^{K_s} I(T_l^{(s)} \in D_s^{(b)})\Big] - (\mathrm{card}(D_s^{(b)}) - K_s)^+,$$

where $x^+ = x$ if $x > 0$, and 0 otherwise, $D_s^{(b)}$ is the set of detected change-points for the sth site in the bth posterior sample, and $\mathrm{card}(D_s^{(b)})$ is the cardinality of the set $D_s^{(b)}$. The maximum value of $R_m(s,b)$ is K_s iff $D_s^{(b)} = \{T_1^{(s)}, \ldots, T_{K_s}^{(s)}\}$, the set of true change-points. For all other choices of $D_s^{(b)}$, we have integer-valued $R_m(s,b) < K_s$. This measure penalizes a set $D_s^{(b)}$ that is either too large or too small. The accuracy rate, R_a, is taken as the average of the $R_m(s,b)$s,

$$R_a = \mathrm{ave}_s\, \mathrm{ave}_b \left[\frac{R_m(s,b)}{K_s}\right].$$

Here R_a closes to 1 indicates a better capability of detecting the change-points.

2. *A measure for comparing a pair of clustering of sites*: Consider a 2×2 cross-classification table to measure deviation between two cluster configurations, C and C^*, say. Entries of the table are $\{n_{ij}\}, 1 \le i, j \le 2$. n_{11} is the number of pairs of sites, out of all $n_{++} = \binom{N}{2}$ pairs, that are in the same cluster in C as well as C^*. n_{22} is the number of pairs of sites that are not in the same cluster in C as well as C^*. n_{12} is the number of pairs that are in the same cluster in C but not in C^*. Similarly, n_{21} is the number of pairs that are not in the same cluster in C but in the same cluster in C^*. If two cluster configurations C and C^* are the same, we have $n_{11} = n_{+1}$ and $n_{22} = n_{+2}$. The interpretation of n_{+1} is that it is the number of pairs which are in the same cluster in C^*, whereas the interpretation of n_{+2} is that it is the number of pairs that are not in the same cluster in C^*. Thus, the latter quantities n_{+1} and n_{+2} are dependent on C^* only and not on C.
We consider two measures, sensitivity $S_1 = n_{11}/n_{+1}$ and specificity $S_2 = n_{22}/n_{+2}$. The interpretation of S_1 is that it is the proportion of pairs that are also in the same cluster in C given that they are in the same cluster in C^*. Similarly, S_2 is the proportion of pairs that are also not in the same cluster in C given that they are not in the same cluster in C^*. S_1 and S_2 take values between 0 and 1, the ideal value of (S_1, S_2) being $(1,1)$. Also note that for a cluster $C \subset C^*$ (that is, every partition of C is in some partition of C^*), the number of clusters in C is much larger than that of C^*. In this case, $n_{12} = 0$ yielding $S_2 = 1$ but $S_1 \le 1$. On the other hand, when $C^* \subset C$, we have $S_1 = 1$ and $S_2 \le 1$. Thus, deviations from the point $(1,1)$ or from the lines $y = 1$ and $x = 1$ give an idea about the nature of deviations of the clustering C from the clustering C^*.
We have determined the "central" cluster based on *dist*. Thus, (S_1, S_2) can be used to measure deviations between a Gibbs cluster configuration and a

"center" (or true for simulation studies) cluster configuration. Suppose that from the Gibbs output, we have B cluster configurations, $C_1, \ldots, C_B$, of sites $s = 1, \ldots, N$. The *dist* method will give a "central" cluster, say $\bar{C}$. Then, for the bth Gibbs cluster configuration, we can calculate $(S_1(b), S_2(b))$ with $C = C_b$ and $C^* = \bar{C}$. A plot of $(S_1(b), S_2(b))$ will indicate how dispersed the bth Gibbs cluster is with respect to the central cluster configuration $\bar{C}$: If $(S_1(b), S_2(b))$ is concentrated close to $(1, 1)$, this indicates that deviations of C_b from $\bar{C}$ is not much. Points $(S_1(b), S_2(b))$ far from $(1, 1)$ indicate two different types of deviations based on their proximity to the lines $x = 1$ or $y = 1$. Proximity to the line $x = 1$ indicates that among those pairs that are in the same cluster in $\bar{C}$, more pairs are in the same cluster in C_b as well, whereas proximity to $y = 1$ indicates that among those pairs not in the same cluster in $\bar{C}$, more pairs are not in the same cluster in C_b as well. For simulation studies, $\bar{C}$ can be taken to be the true cluster configuration since the true clustering is known.

In the 2×2 cross-classification table, there are two degrees of freedom. The quantities n_{+1} and n_{+2} are fixed for the true (or central) cluster configuration, so we need two free parameters, say n_{11} and n_{22}, to determine all entries in the table completely. Thus, the scatter plot of $\{(S_1(b), S_2(b)), b = 1, \ldots, B\}$ gives a complete picture of variability. As a numerical measure for variability, we consider

$$SS = \frac{1}{B} \sum_{b=1}^{B} (2 - S_1(b) - S_2(b)) \tag{4.1}$$

with smaller SS indicating better performance.

3. *Predictive measure*: For the predictive analysis, $Y^*_{s,t}$, is sampled from the model given model parameters at each Gibbs iteration. The B values of $Y^*_{s,t}$ are then used to construct the 95% credible predictive interval.

5. Analysis of Cancer Curves over States

In this section we introduce an alternative approach based on the penalized likelihood and compare the proposed method with the penalized likelihood approach using simulated data and cancer mortality curve data.

5.1. Alternative two-stage approach

There are several joinpoint models that can detect change-points in a single curve. Thus, to cluster curves based on change-points, one needs to apply clustering methods after fitting each curve. It is not clear how to apply available clustering methods such as a k-mean clustering method when the number of change-points and location of change-points are different for each curve, and

a two-stage approach may not accommodate variability among curves. One way to overcome the different dimensionalities is to use a likelihood based approach. For comparison purpose, we propose a penalized likelihood based approach to detect change-points and cluster curves.

Penalized likelihood approach: For each site s, fix the number of change-points, K_s, and the location of the change-points $\boldsymbol{T}^{(s)} = (T_1^{(s)}, \ldots, T_{K_s}^{(s)})$. Let $\alpha_l^{(s)}$ and $\beta_l^{(s)}$ be, respectively, the intercept and slope of the regression line in the lth time segment, $l = 1, \ldots, (K_s + 1)$. The errors $\epsilon_{s,t}$ follow the normal distribution $N(0, \sigma_s^2)$. Take $\boldsymbol{\vartheta}_s = \{(\alpha_l^{(s)}, \beta_l^{(s)}), l = 1, \ldots, (K_s + 1), \sigma_s^2\}$ to be all the unknown parameters. Denote by $\hat{\boldsymbol{\vartheta}}_s$ the maximum likelihood estimate of $\boldsymbol{\vartheta}_s$ under the normal $N(0, \sigma_s^2)$ model. Consider the penalized log-likelihood function (BIC) given by

$$\begin{aligned} H_s(K_s, \boldsymbol{T}^{(s)}) &= -2 \log \ell(\mathbf{Y}_s \,|\, \hat{\boldsymbol{\vartheta}}_s) + (2K_s + 3)\log(n) \\ &= n \log(\hat{\sigma}_s^2) + (2K_s + 3)\log(n) + \text{const. terms}, \end{aligned} \tag{5.1}$$

since there are $2K_s+3$ parameters consisting of $\alpha_l^{(s)}$ and $\beta_l^{(s)}$ for $l = 1, \ldots, K_s+1$, and the variance parameter σ_s^2. Change-points and parameters are estimated based on minimizing the BIC criteria in (5.1) as the first stage of a two-stage procedure.

In the second stage, clustering is obtained based on a modification of the functions $H_s(K_s, \mathbf{T}^{(s)})$ in (5.1). For sites s_1 and s_2, we get the penalized likelihood measures $H_{s_1}(K_{s_1}, \mathbf{T}^{(s_1)})$ and $H_{s_2}(K_{s_2}, \mathbf{T}^{(s_2)})$. The goal is to see whether s_1 and s_2 can be grouped based on similar locations of change-points and slopes. We consider the combined penalized log-likelihood criteria:

$$\begin{aligned} H_c(K_c, \mathbf{T}^{(c)}) = &-2 \log \ell(\mathbf{Y}_{s_1} \,|\, \hat{\boldsymbol{\vartheta}}_{s_1,c}) - 2 \log \ell(\mathbf{Y}_{s_2} \,|\, \hat{\boldsymbol{\vartheta}}_{s_2,c}) + (3K_c + 5)\log(n) \\ &+\text{const. terms}, \end{aligned}$$

where the slope components of $\boldsymbol{\vartheta}_{s_1,c}$ and $\boldsymbol{\vartheta}_{s_2,c}$ are common to both s_1 and s_2 but the intercept and the variance parameters are different. We propose dissimilarity measure between s_1 and s_2 as

$$d(s_1, s_2) = \inf_{K_c, \mathbf{T}^{(c)}} \left[H_c(K_c, \mathbf{T}^{(c)}) - H_{s_1}(K_c, \mathbf{T}^{(c)}) - H_{s_2}(K_c, \mathbf{T}^{(c)}) \right].$$

If we view $-2 \times$ loglikelihood as a cost function, it is intuitively clear that the cost of not combining is always smaller than the cost of combining because MLEs are obtained over larger sets. The expression

$$-2 \left(\log \ell(\mathbf{Y}_{s_1} \,|\, \hat{\boldsymbol{\vartheta}}_{s_1,c}) + \log \ell(\mathbf{Y}_{s_2} \,|\, \hat{\boldsymbol{\vartheta}}_{s_2,c}) - \log \ell(\mathbf{Y}_{s_1} \,|\, \hat{\boldsymbol{\vartheta}}_{s_1}) - \log \ell(\mathbf{Y}_{s_2} \,|\, \hat{\boldsymbol{\vartheta}}_{s_2}) \right)$$

is, therefore, always non-negative, and 0 only when the cost of combining is as good as the cost of not combining. In this case, it is possible that $d(s_1, s_2) < 0$ due to the addition of the penalty term on the number of parameters, and so we choose $d(s_1, s_2) = 0$ as a rule whenever its value is less than 0. In all other cases, $d(s_1, s_2)$ is typically positive. Once the dissimilarity measures are obtained for all combinations of pairs (s_1, s_2), we can apply an agglomerative clustering algorithm to obtain a 'center' cluster configuration for subsequent analysis. The penalized likelihood approach is applied to the curve for each site s separately in the first stage. Since change-points and parameters are estimated based on minimizing the BIC before the clustering procedure, R_a does not change even if we set different number of clusters in the agglomerative algorithm. The R_a value that we obtain is optimal in the sense that further agglomerative clustering may reduce (or increase) the number of change-points of sites with weak signals when grouped together with other sites with stronger signals. Also, R_a is the averaged value over sites s only since this approach does not produce posterior samples of clusters for variability assessment. Similarly, for SS, the summation in (4.1) involves one term only since there is no summation over b.

5.2. Simulation study

In this section, we apply the proposed method and the penalized likelihood approach to the simulated data and compare their performances. We simulated data with $n = 38$ time points for $N = 49$ sites according to the model in (2.1), which is the same structure as that of U.S. states in the cancer mortality data. We partitioned the sites into $d = 6$ clusters, C_r, $r = 1, \ldots, 6$, with each $s \in C_r$ having common values for $(\boldsymbol{\beta}^{(s)}, K_s, \boldsymbol{T}^{(s)})$, hence denoted by $(\boldsymbol{\beta}^{(r)}, K_r, \boldsymbol{T}^{(r)})$. The piecewise constant function $\boldsymbol{\theta}_s$ was taken to be $\boldsymbol{\theta}_s(t) = \beta_l^{(r)}$ for $t \in [T_{l-1}^{(r)}, T_l^{(r)})$, for $l = 1, \ldots, K_r + 1$. The following choices were made corresponding to each C_r: For $r = 1$, C_1 contained 9 sites ($|C_1| = 9$), with $K_1 = 4$ change-points. We took $\boldsymbol{T}^{(1)} = (1979, 1986, 1993, 2000)$ and $\boldsymbol{\beta}^{(1)} = (0.06, -0.05, 0.04, -0.04, 0.05)$. For $r = 2$, $|C_2| = 9$, $K_2 = 2$, $\boldsymbol{T}^{(2)} = (1976, 1989)$, and $\boldsymbol{\beta}^{(2)} = (0.05, -0.07, 0.06)$. For $r = 3$, $|C_3| = 8$, $K_3 = 3$, $\boldsymbol{T}^{(3)} = (1976, 1986, 1997)$, and $\boldsymbol{\beta}^{(3)} = (0.07, -0.08, 0.07, -0.06)$. For $r = 4$, $|C_4| = 7$, $K_4 = 1$, $\boldsymbol{T}^{(4)} = (1986)$, and $\boldsymbol{\beta}^{(4)} = (0.05, -0.02)$. For $r = 5$, $|C_5| = 7$, $K_5 = 2$, $\boldsymbol{T}^{(5)} = (1980, 1995)$, and $\boldsymbol{\beta}^{(5)} = (0.04, 0.01, -0.03)$. Finally, for $r = 6$, $|C_6| = 9$, $K_6 = 1$, $\boldsymbol{T}^{(6)} = (1989)$, and $\boldsymbol{\beta}^{(6)} = (0.02, -0.01)$. Seven scenarios were made depending on the choices of σ_s^2. For the first four cases, we generated σ_s^2, $s = 1, \ldots, N$, i.i.d. from uniform distributions with the ranges $(0.003, 0.009)$, $(0.009, 0.015)$, $(0.015, 0.021)$, and $(0.003, 0.021)$. The first range (case 1) is similar to the lung cancer mortality data in log scale that we are analyzing. For the last three cases, we generated σ_s^2 from Gamma distribution with shape and scale

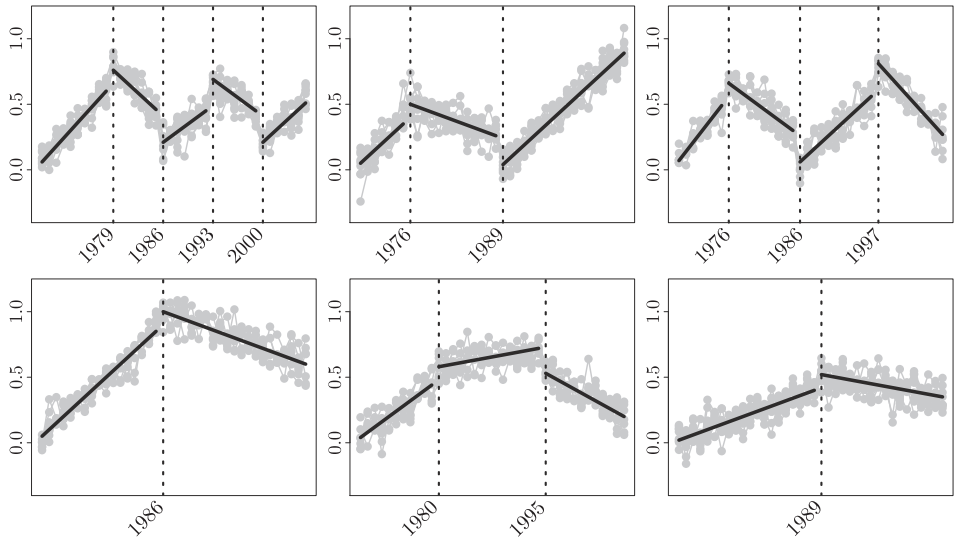

Figure 3. Mean curves (solid lines) for simulated six clusters with data (grey dots) generated under case 1. Data curves (grey dots) within a cluster are plotted together in each panel.

parameters, $(0.5, 0.012)$, $(0.25, 0.024)$, and $(0.2, 0.09)$ so that the σ_s^2 were generated from a skewed distribution. Figure 3 shows the mean curves of six clusters from the above description of clustering and parameter specification with the data generated from the case 1.

The minimum number of time points in each time segment, w, was set as $w = 7$. Theorem 1 since $w \geq 3$. Values of hyper-parameters for the priors of σ_s^2, α_0 and λ were set to make diffused/dispersed priors and the same values were used for the data analysis as well; the specific values are given in the next section. Three Gibbs chains were started from three different settings perturbed from the true setting. We randomly assigned sites into the clusters while we fixed the number of clusters and the number of sites in each cluster. Also, the number of change-points for each site was set to zero at the beginning of the Gibbs chains. The assessment of convergence was carried out based on the methodology of Gelman and Rubin and convergence was achieved after 5,000 iterations. We ran 1,000 iterations and used them for posterior inference. The total running time of 6,000 iterations for each chain was approximately 4 hours.

The results are summarized in Tables 1, 2 and Figure 4 for the simulated data. SS_t is the value of SS calculated based on the Gibbs cluster configurations and the true clustering configuration for the proposed approach. For the penalized likelihood approach, SS_t was calculated based on the resulting clustering configuration from the penalized likelihood approach and the true clustering configuration.

Table 1. Comparison of performances for the proposed and penalized likelihood methods based on simulated data under seven different levels of σ_s^2's: Uniform distributions on $(0.003, 0.009)$, $(0.009, 0.015)$, $(0.015, 0.021)$, $(0.003, 0.021)$, respectively and Gamma distributions with (shape, scale) $(0.5, 0.012)$, $(0.25, 0.024)$, $(0.2, 0.09)$, respectively. SS_t is the SS measure using true cluster configuration. The first two columns are for the proposed method and the last four columns are for the penalized likelihood method. $P5$, $P6$ and $P7$ indicate that the clustering configuration was chosen by setting the number of clusters as $5, 6$ and 7, respectively, for the penalized likelihood method.

Case	R_a	SS_t	$P : R_a$	$P5 : SS_t$	$P6 : SS_t$	$P7 : SS_t$
1	0.886	0.027	0.250	0.049	0.000	0.056
2	0.700	0.073	0.117	0.108	0.052	0.103
3	0.542	0.129	0.145	0.156	0.228	0.307
4	0.929	0.005	0.017	0.049	0.000	0.045
5	0.974	0.030	0.420	0.113	0.082	0.191
6	0.962	0.044	0.384	0.097	0.048	0.161
7	0.955	0.086	0.262	0.302	0.252	0.340

In the simulation settings the mean level of σ_s^2 increases while the variability among the σ_s^2 is the same from Cases 1 to 3 with the uniform distributions. Case 4 has larger variability among the σ_s^2 while the mean level is the same as in Case 2. On the other hand, from Cases 5 to 7 for Gamma distributions, the variability among the σ_s^2 increases.

In Table 1 we expect that the capability of detecting change-points and clustering decreases as we have larger σ_s^2; overall, we see that the values of R_a decrease and the values of SS_t increase for both the proposed and penalized likelihood methods from Cases 1 to 3, and from Cases 5 to 7 (The R_a and SS_t for the proposed method were calculated by averaging over Gibbs samples while they were not for the penalized likelihood method). The proposed method performs better for the detection of change-points as the magnitude of R_a is closer to 1 in each case compared to the penalized likelihood method. The magnitudes of SS_t are comparable for both approaches when the level of σ_s^2 is smaller whereas SS_t is smaller (closer to 0) for the proposed method when the level of σ_s^2 increases. The differences in R_a and SS_t between the approaches is more pronounced when the σ_s^2s are generated from a skewed distribution, which could be the situation with data.

Regarding the performance of change-points detection, the penalized likelihood approach detects change-points for each curve separately. Thus, it is difficult to detect the change-points correctly for curves with larger σ_s^2. The proposed method detects the change-points by borrowing information from other

sites within a cluster, so that it benefits from sites with smaller σ_s^2 within the cluster.

For the Cases 1 and 4, the values of SS_t from the penalized likelihood approach using the true number of clusters are zero (P6 column in the Table 1), which means perfect cluster configuration. Meanwhile, the SS_t for the proposed Bayesian approach are very small but not zero. This is due to the uncertainty in the Gibbs samples. Indeed, SS_t based on the centered clustering configuration obtained by the Gibbs samples and the true clustering configuration are also zero for Cases 1 and 4. Thus, the approaches are comparable. However, we do not know the true number of clusters in practice. When the number of clusters is misspecified (P5 or P6), the values of SS_t for the penalized likelihood method are greater than those for the proposed method.

Cases 1 and 4 show the best performance on the clustering configuration for both approaches. The range for the uniform distribution of Case 4 covers the range for the uniform distribution of Case 1, which implies that some of the σ_s^2 are as small as those in Case 1. This result could be explained by observing that each cluster in their true clustering configuration has the sites with smaller σ_s^2, since the σ_s^2 were uniformly generated and those sites with smaller σ_s^2 (strong signals) were dominating the clustering results.

The degree of uncertainty of clustering configuration for the simulation study are shown by the posterior probabilities of the number of clusters, d, in Table 2 as well as the plots of (Sensitivity, Specificity)= (S_1, S_2) in Figure 4. Table 2 shows the posterior probabilities of the number of clusters are highly concentrated on $d = 6$, the true number of clusters. Although this looks too extreme, the clustering configuration can still vary given correctly estimated d and this can be seen in the plots in Figure 4. The (S_1, S_2) plots Cases 1, 3, 5, and 7 are given. We can see clearly that, when the noise level (σ_s^2) increases, there is more variability. One sees that the penalized likelihood method is sensitive to the assumed number of clusters. While it produces the best result with correctly specified number of clusters, the proposed method gives better result than that of the penalized likelihood method with misspecified number of clusters.

In summary, we find the proposed method more robust with respect to higher noise levels as well as different types of noise distribution compared to the two-stage penalized likelihood approach. Further, the non-parametric Bayesian approach is able to assess the variability of the clustering configurations from a reference configuration based on the SS_t and SS_c criteria developed in (4.1).

5.3. Cancer curves data

We applied the proposed model to the logarithm of cancer mortality curves from 48 U.S. states and Washington D.C. According to Figure 1, we specified the

Table 2. Posterior probabilities of the number of clusters d for the simulation study.

Case	$p(d=4)$	$p(d=5)$	$p(d=6)$	$p(d=7)$	$p(d=8)$
1	0.0000	0.0000	0.9997	0.0003	0.0000
2	0.0000	0.0000	0.9997	0.0003	0.0000
3	0.0000	0.9300	0.0697	0.0003	0.0000
4	0.0000	0.0000	0.9997	0.0003	0.0000
5	0.0000	0.0000	0.9997	0.0003	0.0000
6	0.0000	0.0000	0.9993	0.0007	0.0000
7	0.0000	0.0000	0.9953	0.0047	0.0000

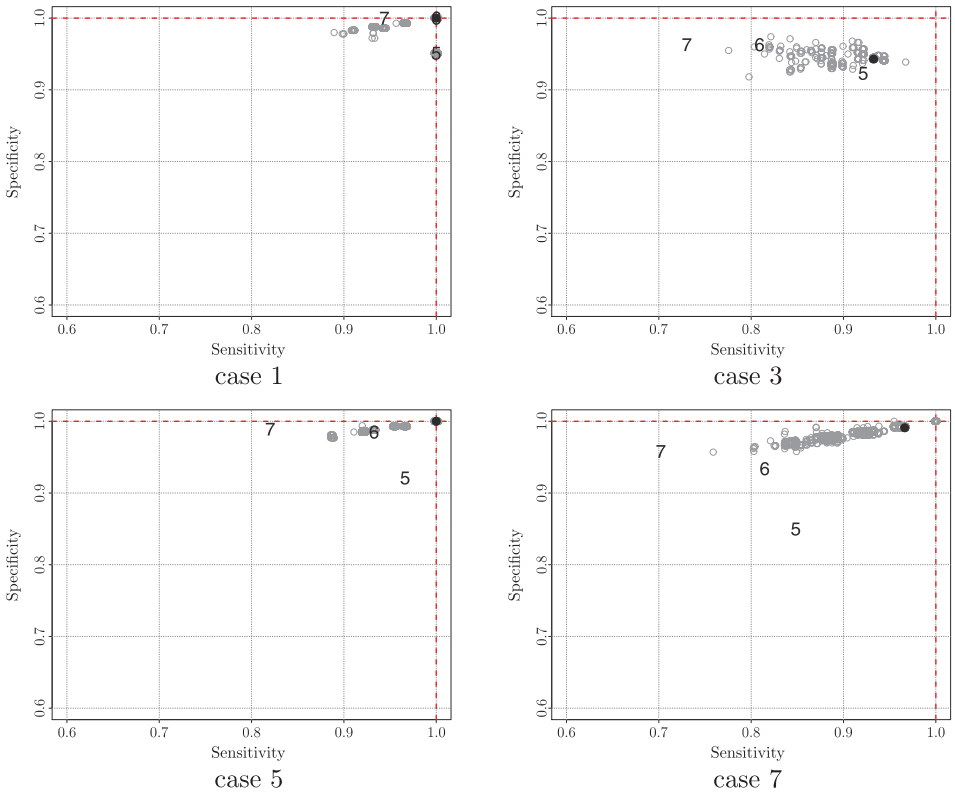

Figure 4. (S_1, S_2) plot of clustering configuration from each Gibbs sample versus the true clustering configuration for simulated data. S_1 is Sensitivity and S_2 is Specificity. The numbers imposed show (S_1, S_2) of clustering configuration under the penalized likelihood approach for d, the number of clusters, fixed at that value, versus the true clustering configuration. The filled dot indicates (S_1, S_2) for central clustering configuration from Gibbs samples versus the true clustering configuration.

Table 3. Posterior probabilities of number of clusters d for fitted model.

$P(d=2)$	$P(d=3)$	$P(d=4)$	$P(d=5)$	$P(d=6)$	$P(d=7)$
0.0000	0.0615	0.9193	0.0193	0.0000	0.0000

minimum number of time points in each time interval as $w = 7$. The resulting maximum number of change-points was $k^* = 4$. The hyper-parameters were specified as follows: for the inverse Gamma prior on σ_s^2, we took $a_\sigma = a_\tau = 2$, $b_\sigma = b_\tau = 100$, so that the variance was infinite; for the Gamma prior on α_0 and λ, we took $a_\alpha = a_\lambda = 2$, $b_\alpha = b_\lambda = 1,000$, so that the priors are well dispersed (mean $= 2 \times 10^3$ and variance $= 2 \times 10^6$). The dispersion parameter α_0 controls the number of clusters and λ controls the number of change-points. From Figure 1, it is reasonable to assume an informative prior which centers around 1 or 2 for λ.

We ran four Gibbs chains. Convergence was established after 15,000 iterations and we took 1,000 samples from each chain after 15,000 iterations so that 4,000 samples were used for further posterior analysis. The number of clusters of states based on the highest posterior probability was found to be $d = 4$; see Table 3 for posterior probabilities. The corresponding dendrograms of clustering results is shown in Figure 5. The posterior inference for parameters and summary statistics is shown in Table 6. The posterior estimates are tremendously different from the prior means, which indicates that the diffuse priors did not affect the posterior behavior. The dendrogram also clearly shows that the four clusters are achieved by the Model.

Cluster (a) had one change-point at year 1990, except for one state. The log-rates for these states decreased a little after increasing until 1990. Cluster (b) also had one change-point, but at year 1988, and the log-rates were steady after increasing until 1988. Cluster (c) represents states with two change-points detected at years 1981 and 1992; these states had log-rates increasing slower after 1981 and dropping after 1992. Cluster (d) had one change-point; the log-rates were decreasing after increasing until 1991, but with decreasing rate larger than that of Cluster (a).

The change rates, and magnitude of change at each detected location for each cluster with 95% credible intervals are summarized in Table 4. State-by-state change-point analysis can also be done. For example, the marginal posterior probabilities corresponding to zero, one, and two change-points for four states are given in Table 5. The marginal posterior probability of having one change-point was the highest for all four states. Table 5 also provides the marginal posterior probabilities of the locations of the change-points. The posterior estimates of the other parameters, α_0 and λ, are given in Table 6. Figures 9−10 show membership

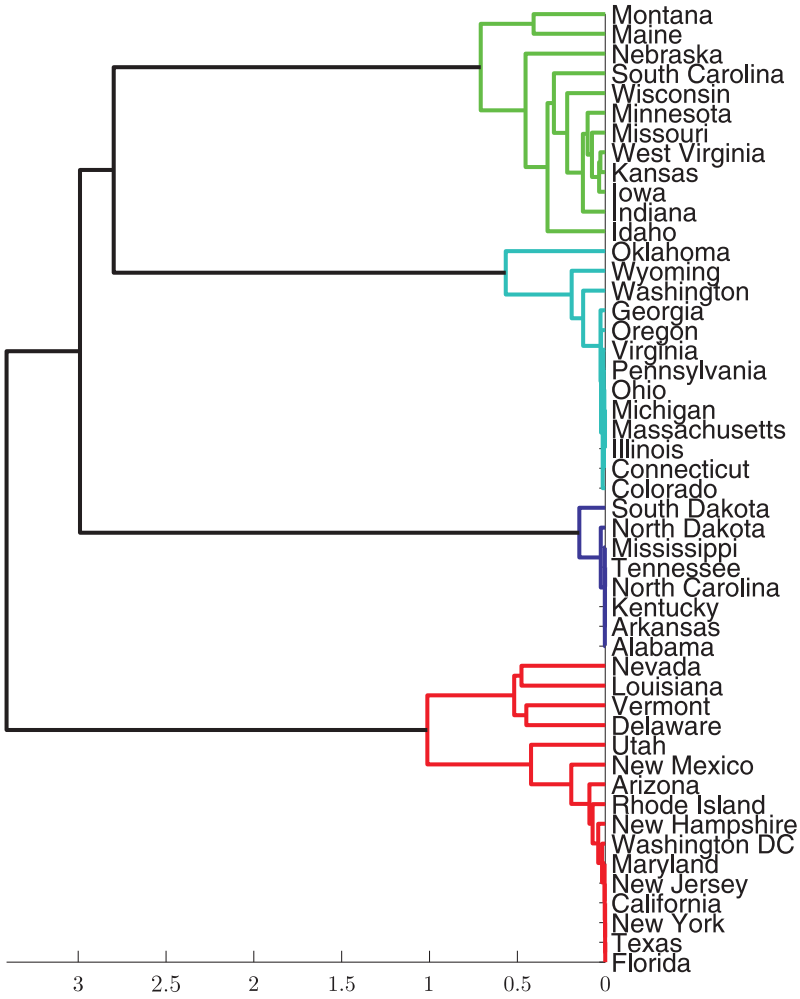

Figure 5. Clustering result based on the dissimilarity measure from the posterior samples.

of each cluster with 95% predictive intervals along with the data curves and the posterior estimate of σ_s^2.

We cannot obtain R_a or SS_t for these data since we do not know the true change-points and clustering configuration. However, we can obtain SS_c and compare the result with the penalized likelihood approach. Figure 6 shows (S_1, S_2) plotted for the data. We plotted (S_1, S_2) values between a clustering configuration from each posterior sample and the central clustering configuration.

For the comparison, we also plotted (S_1, S_2) to compare the central clustering configuration from our proposed method and the clustering configuration of the

Table 4. Summary in terms of posterior means (95% credible intervals) of slopes $\beta_l^d, l = 1, \ldots, K_d+1$ and differences of adjacent slopes $\Delta_l^d = \beta_{l+1}^d - \beta_l^d$, $l = 1, \ldots, K_d$ for cluster obtained by Model 1. K_r is the number of change-points for each cluster. Cluster estimates are averaging over states with representative pattern of change-points.

cluster	K_d	$\hat{\beta}_1^d$	$\hat{\beta}_2^d$	$\hat{\beta}_3^d$	$\hat{\Delta}_1^d$	$\hat{\Delta}_2^d$
(a)	1	0.0281	-0.0020	-	-0.0301	-
		$(0.0267, 0.0291)$	$(-0.0035, -0.0006)$		$(-0.0317, -0.0283)$	
(b)	1	0.0356	0.0002	-	-0.0354	-
		$(0.0341, 0.0369)$	$(-0.0010, 0.0014)$		$(-0.0371, -0.0337)$	
(c)	2	0.0307	0.0164	-0.0077	-0.0144	-0.0241
		$(0.0289, 0.0326)$	$(0.0145, 0.0183)$	$(-0.0090, -0.0058)$	$(-0.0170, -0.0119)$	$(-0.0261, -0.0220)$
(d)	1	0.0197	-0.0131	-	-0.0328	-
		$(0.0186, 0.0211)$	$(-0.0146, -0.0113)$		$(-0.0343, -0.0313)$	

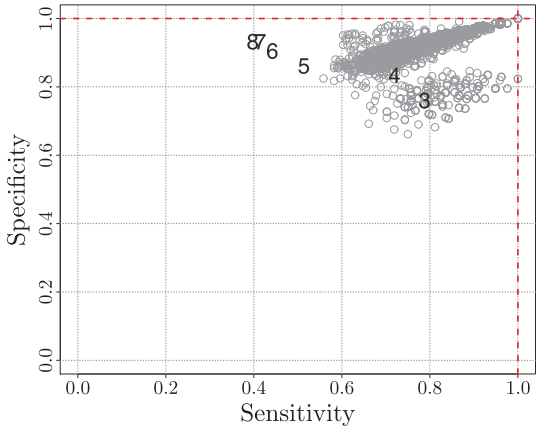

Figure 6. (S_1, S_2) plot for real data: clustering configuration for each posterior sample versus the central clustering configuration. The plotted numbers indicate (S_1, S_2) for the clustering configuration under penalized likelihood approach with the number representing the number of clusters d.

penalized likelihood approach with different numbers of clusters ($d = 3, \ldots, 8$). These are overlaid in Figure 6. SS_c from the proposed method was 0.2951 while SS_c from the penalized likelihood approach were 0.9415, 0.7993, 0.698, 0.7259, 0.7602, and 0.7831 corresponding to $d = 3, \ldots, 8$, respectively. Note that SS_c from the proposed method is smaller than those corresponding to each $d = 3, 4, \ldots, 8$, indicating that if we expect the clustering configuration from the penalized likelihood approach is to be similar to the central clustering configuration obtained by the proposed model, (S_1, S_2) for the penalized likelihood approach should be close to (1,1), but that is not the case.

6. Discussion

Table 5. Marginal posterior probabilities of a change-point for four example states.

Change-Points	Florida	Arizona	Missouri	Indiana
No Change-Points	0	0	0	0
$T_1 = 1992$	0.0068	0.0065	0.0025	0.0008
$T_1 = 1991$	0.717	0.6718	0.0228	0.0143
$T_1 = 1990$	0.276	0.265	0.5125	0.4803
$T_1 = 1989$	0.0003	0.0013	0.3248	0.3335
$T_1 = 1988$	0	0.0003	0.0223	0.033
$T_1 = 1987$	0	0	0	0.002
Two Change-Points	0	0.0553	0.1153	0.1363

Table 6. Summary in terms of posterior means (95% credible intervals) of parameters α_0 and λ.

α_0	λ
2.0247	1.4025
$(0.4568, 3.8889)$	$(1.0311, 1.8047)$

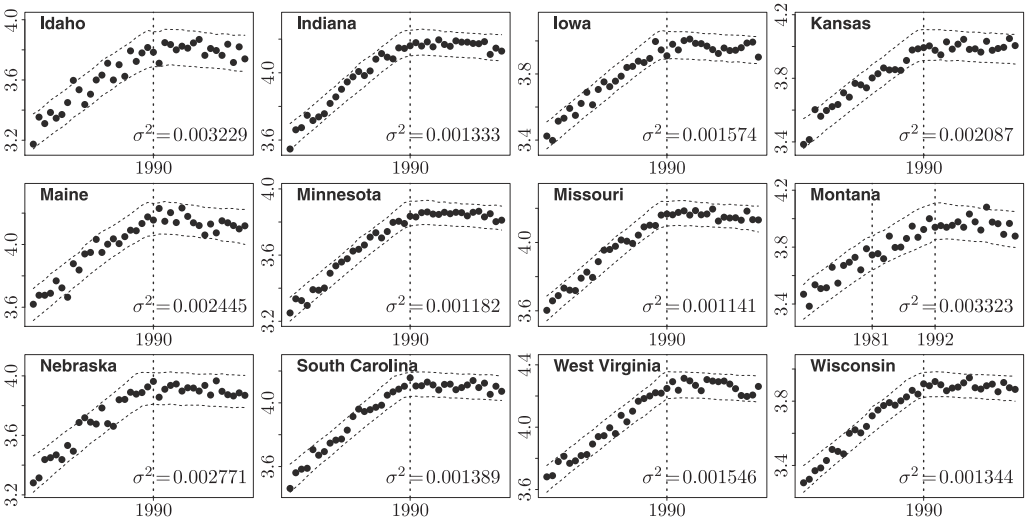

Figure 7. States in cluster (a). We omit the description since it is the same as Figure 9.

We proposed a change-point model that works for multiple curves and concurrent clustering of curves based on their change-points. Clustering is established by introducing a Dirichlet process prior on the space of step functions over time. The model was applied to analyze state-wise log scale age-adjusted cancer mortality rates to find local change-points and clusters that have similar changes.

For the analysis of lung cancer mortality rates, we found that state-level and

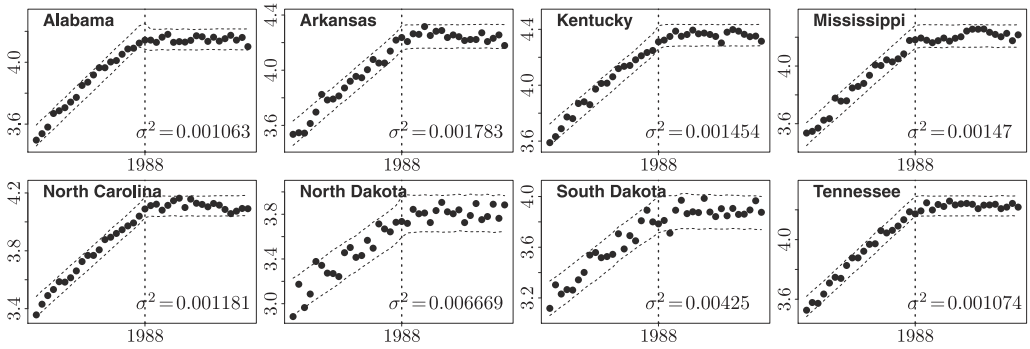

Figure 8. States in cluster (b). We omit the description since it is the same as Figure 9.

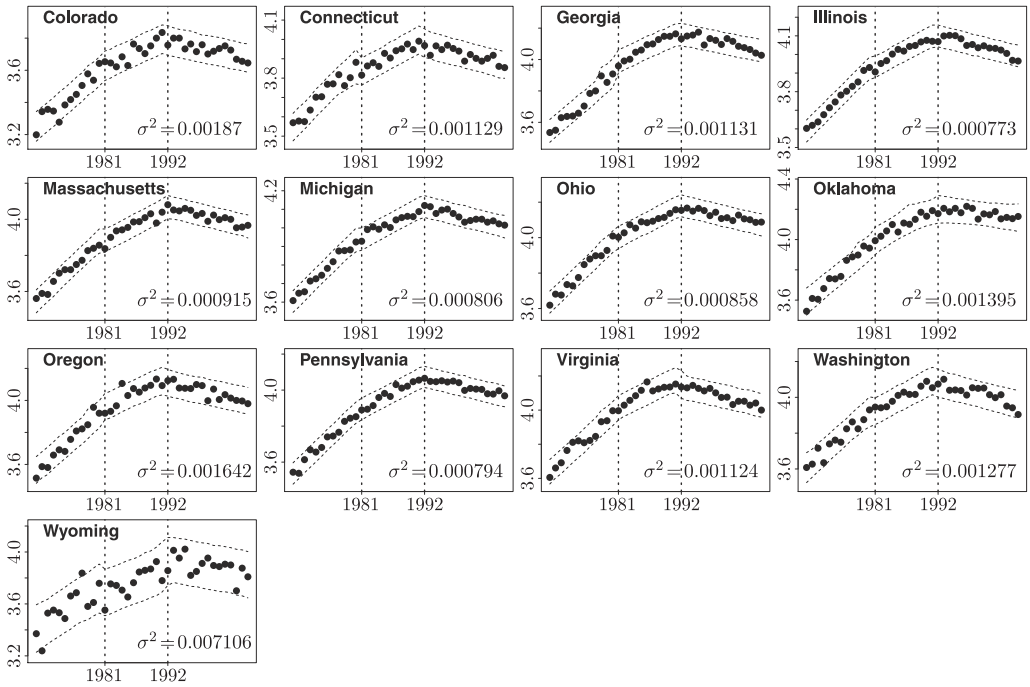

Figure 9. States in cluster (c). Points are cancer mortality data in log scale and the dotted lines along with the data are the 95% predictive intervals. The σ^2 value is the posterior estimate. Vertical dotted lines are change-points locations.

national level age-adjusted lung cancer mortality rates showed a clear change-point around the late 1980s to early 1990s. Some states like Florida and Arizona followed similar patterns as the national level rates, while some states like Missouri and Indiana showed different patterns from the national level rates (see Figure 1). States in Clusters (a) and (b) had smaller rates of change after the

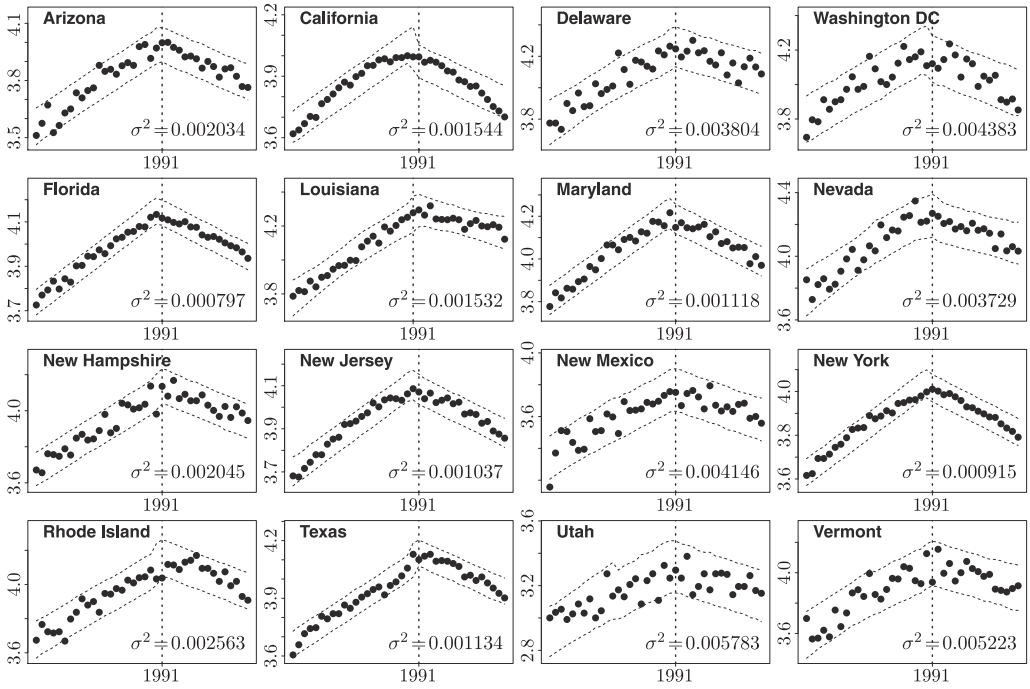

Figure 10. States in cluster (d). We omit the description since it is the same as Figure 9.

change-point compared to Cluster (d) and the national level (see Table 4). We can see that lung cancer mortality rates have not changed much since 1990s for these states, while the national level seems significantly decreased.

Acknowledgement

The authors thank the Editor, an associate editor, and the referees for their valuable suggestions for improving this paper. The authors would like to acknowledge the support of NSF SES-0961649, NSF DMS-1106450, and Universiti Teknologi PETRONAS URIF grant no. 16/2013 while conducting this research.

Appendix A: Propriety of the Posterior and Validity of the Gibbs Updating Scheme

Let $\mathbf{1}_m$ ($\mathbf{0}_m$) be an m-by-1 matrix with all entries 1 (0), and denote by I_m an m-by-m identity matrix. We have $\boldsymbol{Y} = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_N)^T$, with $\boldsymbol{Y}_s = (Y_{s,1}, \ldots, Y_{s,n})^T$, $\boldsymbol{\alpha} = (\boldsymbol{\alpha}^{(1)T}, \ldots, \boldsymbol{\alpha}^{(N)T})^T$ with $\boldsymbol{\alpha}^{(s)} = (\alpha_1^{(s)}, \ldots, \alpha_{K_s+1}^{(s)})^T$, and $\boldsymbol{\beta} = (\boldsymbol{\beta}^{(1)T}, \ldots, \boldsymbol{\beta}^{(N)T})^T$ with $\boldsymbol{\beta}^{(s)} = (\beta_1^{(s)}, \ldots, \beta_{K_s+1}^{(s)})^T$. Take $X_0^{(s)}$ to be the n-by-$K_s + 1$ design matrix that corresponds to $\boldsymbol{\alpha}^{(s)}$, and $X_1^{(s)}$ to be the design

matrix for $\boldsymbol{\beta}^{(s)}$. For all $s \in C_r$, $\boldsymbol{\theta}_s \equiv (\boldsymbol{\beta}^{(s)}, K_s, \boldsymbol{T}^{(s)})$ is common, hence denoted by $\boldsymbol{\theta}_{(r)} \equiv (\boldsymbol{\beta}^{(r)}, K_r, \boldsymbol{T}^{(r)})$. Take $X_0^{(s)} = X_{0,r}$ and $X_1^{(s)} = X_{1,r}$ for $s \in C_r$.

Proof of Theorem 1. Let $S \equiv \{1, \ldots, N\}$ be the set of all N sites. Denote a partition of S by $\boldsymbol{c} = \cup_{r=1}^{d} C_r$, and take $\mathcal{P}$ to be the set of all partitions of S. A randomly generated distribution F from $DP(\alpha_0 G_0)$ admits the Sethuraman representation (Sethuraman (1994)). By integrating out the random measure F, the equivalence between the DP prior (for any general space $\boldsymbol{\Theta}$) and the distribution it induces on $\mathcal{P}$ is well known. Moreover, the probability distribution on $\mathcal{P}$ has the explicit form

$$\pi(\boldsymbol{c}) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_0 + N)} \alpha_0^d \prod_{r=1}^{d} (|C_r| - 1)!, \tag{A.1}$$

where $|C_r|$ is the number of elements in C_r.

To show the propriety of the posterior of $(\underline{\boldsymbol{\theta}}, \boldsymbol{\alpha}, \boldsymbol{\sigma})$, we need to show

$$\sum_{\boldsymbol{c} \in \mathcal{P}} \int_{\boldsymbol{\Theta}^d} \int_{\boldsymbol{\sigma}} \int_{\boldsymbol{\alpha}} f(\boldsymbol{Y} \,|\, \underline{\boldsymbol{\theta}}, \boldsymbol{c}, \boldsymbol{\alpha}, \boldsymbol{\sigma}) \pi(\boldsymbol{\alpha} \,|\, \boldsymbol{c}) \pi(\boldsymbol{\sigma}) \pi(\boldsymbol{c})\, d\boldsymbol{\alpha} d\boldsymbol{\sigma} G_0(d\underline{\boldsymbol{\theta}}) < \infty, \tag{A.2}$$

where

$$f(\boldsymbol{Y} \,|\, \underline{\boldsymbol{\theta}}, \boldsymbol{\alpha}, \boldsymbol{c}, \boldsymbol{\sigma}) = \prod_{r=1}^{d} \prod_{s \in C_r} \frac{1}{(2\pi\sigma_s^2)^{\frac{n}{2}}} \exp\left\{ -\frac{1}{2\sigma_s^2} \Delta_s^T \Delta_s \right\} \tag{A.3}$$

with $\Delta_s = \boldsymbol{Y}_s - X_{0,r}\boldsymbol{\alpha}^{(s)} - X_{1,r}\boldsymbol{\beta}^{(r)}$. Then, it is enough to show for each $r = 1, \ldots, d$,

$$\begin{aligned} M^r(K_r, \boldsymbol{T}^{(r)}, C_r) = \prod_{s \in C_r} \int_{\sigma_s^2} \int_{\boldsymbol{\beta}^{(r)}} \int_{\boldsymbol{\alpha}^{(s)}} & f(\boldsymbol{Y}_s \,|\, \boldsymbol{\alpha}^{(s)}, \boldsymbol{\beta}^{(r)}, \sigma_s^2, \boldsymbol{c}, K_r, T^{(r)}) \pi(\boldsymbol{\alpha}^{(s)} \,|\, \boldsymbol{c}) \\ & d\boldsymbol{\alpha}^{(s)} \pi(\boldsymbol{\beta}^{(r)} \,|\, \boldsymbol{c}) d\boldsymbol{\beta}^{(r)} \pi(\sigma_s^2) d\sigma_s^2 < \infty, \end{aligned} \tag{A.4}$$

for every realization of $\boldsymbol{K}, \boldsymbol{T}, \boldsymbol{c}$. This is so since $\pi(\boldsymbol{K} \,|\, \boldsymbol{c})$, $\pi(\boldsymbol{T} \,|\, \boldsymbol{c})$, and $\pi(\boldsymbol{c})$ are discrete distributions over finite possible realizations and (A.2) is obtained by finite sums with respect to $\boldsymbol{K}, \boldsymbol{T}, \boldsymbol{c}$.

We integrate with respect to $\boldsymbol{\alpha}^{(s)}$ for each fixed $s \in C_r$:

$$\begin{aligned} & \int_{\boldsymbol{\alpha}^{(s)}} f(\boldsymbol{Y}_s \,|\, \boldsymbol{\alpha}^{(s)}, \boldsymbol{\beta}^{(r)}, \sigma_s^2, \boldsymbol{c}, K_r, T^{(r)}) \pi(\boldsymbol{\alpha}^{(s)} \,|\, \boldsymbol{c})\, d\boldsymbol{\alpha}^{(s)} \\ & \quad = \frac{|X_{0,r}^T X_{0,r}|^{-1/2}}{(2\pi\sigma_s^2)^{(n-K_r-1)/2}} \exp\left\{ -\frac{1}{2\sigma_s^2} Z_s^T P_{X_{0,r}} Z_s \right\}, \end{aligned} \tag{A.5}$$

where $Z_s = \boldsymbol{Y}_s - X_{1,r}\boldsymbol{\beta}^{(r)}$ and $P_{X_{0,r}} = I_n - X_{0,r}\left(X_{0,r}^T X_{0,r}\right)^{-1} X_{0,r}^T$. Here $P_{X_{0,r}}$

has rank $n-(K_r+1)$. Consider the integration of $\boldsymbol{\beta}^{(r)}$:

$$\begin{aligned}
&\int_{\boldsymbol{\beta}^{(r)}}\int_{\boldsymbol{\alpha}^{(s)}} f(\boldsymbol{Y}_s \,|\, \boldsymbol{\alpha}^{(s)}, \boldsymbol{\beta}^{(r)}, \sigma_s^2, \boldsymbol{c}, K_r, T^{(r)})\pi(\boldsymbol{\alpha}^{(s)} \,|\, \boldsymbol{c})\, d\boldsymbol{\alpha}^{(s)}\pi(\boldsymbol{\beta}^{(r)} \,|\, \boldsymbol{c})\, d\boldsymbol{\beta}^{(r)} \\
&\quad = \int_{\boldsymbol{\beta}^{(r)}} \frac{|X_{0,r}^T X_{0,r}|^{-1/2}}{(2\pi\sigma_s^2)^{(n-K_r-1)/2}} \exp\left\{-\frac{1}{2\sigma_s^2} Z_s^T P_{X_{0,r}} Z_s\right\} \pi(\boldsymbol{\beta}^{(r)} \,|\, \boldsymbol{c})\, d\boldsymbol{\beta}^{(r)} \\
&\quad = \frac{|X_{0,r}^T X_{0,r}|^{-1/2}|X_{1,r}^T P_{X_{0,r}} X_{1,r}|^{-1/2}}{(2\pi\sigma_s^2)^{(n-2(K_r-1))/2}} \exp\{-Q_{s,r}\} \\
&\quad \le \frac{|X_{0,r}^T X_{0,r}|^{-1/2}|X_{1,r}^T P_{X_{0,r}} X_{1,r}|^{-1/2}}{(2\pi\sigma_s^2)^{(n-2(K_r-1))/2}}, \qquad \text{(A.6)}
\end{aligned}$$

where $Q_{s,r} = (1/2\sigma_s^2)\boldsymbol{Y}_s^T \left(P_{X_{0,r}} - P_{X_{0,r}} X_{1,r}(X_{1,r}^T P_{X_{0,r}} X_{1,r})^{-1} X_{1,r}^T P_{X_{0,r}}\right) \boldsymbol{Y}_s$. The second equality holds since $X_{1,r}^T P_{X_{0,r}} X_{1,r}$ is a non-singular matrix by assuming the rank of $X_{1,r}$ is less than the rank of $P_{X_{0,r}}$, or $K_r + 1 \le n - (K_r + 1)$. This holds since $w \ge 3$. The last inequality follows from $Q_{s,r} \ge 0$, since $P_{X_{0,r}} - P_{X_{0,r}} X_{1,r}(X_{1,r}^T P_{X_{0,r}} X_{1,r})^{-1} X_{1,r}^T P_{X_{0,r}}$ is a projection matrix.

Finally, we have

$$\begin{aligned}
&M^r(K_r, \boldsymbol{T}^{(r)}, C_r) \\
&\le \prod_{s\in C_r} \int_{\sigma_s^2} \frac{|X_{0,r}^T X_{0,r}|^{-1/2}|X_{1,r}^T P_{X_{0,r}} X_{1,r}|^{-1/2}}{(2\pi\sigma_s^2)^{(n-2(K_r-1))/2}} \pi(\sigma_s^2) d\sigma_s^2 \\
&= \frac{|X_{0,r}^T X_{0,r}|^{-1/2}|X_{1,r}^T P_{X_{0,r}} X_{1,r}|^{-1/2}}{(2\pi)^{(n-2(K_r-1))/2}} \frac{\Gamma(a_\sigma + (n-2(K_r-1))/2)}{\Gamma(a_\sigma)} b_\sigma^{(n-2(K_r-1))/2} \\
&< \infty, \qquad \text{(A.7)}
\end{aligned}$$

which completes the proof.

To validate the Gibbs updating scheme for $\boldsymbol{\theta}_s$, we introduce expressions for various conditional densities to be used for the proof. Note that $\pi(\underline{\boldsymbol{\theta}} \,|\, \boldsymbol{c}, \boldsymbol{Y})$ can be written as

$$\pi(\underline{\boldsymbol{\theta}} \,|\, \boldsymbol{c}, \boldsymbol{Y}) = \prod_{r=1}^{d} \frac{\prod_{s\in C_r} \ell_s(\boldsymbol{Y}_s \,|\, \boldsymbol{\theta}_{(r)})\, G_0(d\boldsymbol{\theta}_{(r)})}{\int_{\boldsymbol{\Theta}} \prod_{s\in C_r} \ell_s(\boldsymbol{Y}_s \,|\, \boldsymbol{\theta}_{(r)})\, G_0(d\boldsymbol{\theta}_{(r)})}, \qquad \text{(A.8)}$$

where $\ell_s(\boldsymbol{Y}_s \,|\, \boldsymbol{\theta}_{(r)}) \propto \int_0^\infty f(\boldsymbol{Y} \,|\, \boldsymbol{\theta}_{(r)}, \sigma_s^2)\pi(\sigma_s^2) d\sigma_s^2$ and $\boldsymbol{\theta}_{(r)}$ is a change-point function for the rth cluster. Note that $\boldsymbol{\theta}_s \equiv \boldsymbol{\theta}_{(r)}$ if $s \in C_r$. Let $\boldsymbol{\theta}_{-s} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_{s-1}, \boldsymbol{\theta}_{s+1}, \ldots, \boldsymbol{\theta}_N)$ to be the collection of all $\underline{\boldsymbol{\theta}}$-components except $\boldsymbol{\theta}_s$. Let $\boldsymbol{c}_{-s}$ be the partition of $S \setminus \{s\}$ determined by $\boldsymbol{\theta}_{-s}$; the identical components of $\boldsymbol{\theta}_{-s}$ uniquely determine the partition of $\boldsymbol{c}_{-s}$. Suppose that $\boldsymbol{c}_{-s} = \cup_{r=1}^{N^*} E_r$. For a partition $\boldsymbol{c}$, $\pi(\boldsymbol{c} \,|\, \boldsymbol{\theta}_{-s}, \boldsymbol{Y})$ can be obtained from the fact that $\pi(\boldsymbol{c} \,|\, \boldsymbol{\theta}_{-s}, \boldsymbol{Y}) \propto \pi(\boldsymbol{c}, \boldsymbol{\theta}_{-s} \,|\, \boldsymbol{Y})$

with $\boldsymbol{\theta}_{-s}$ and $\boldsymbol{c}_{-s}$ treated as fixed. Thus, $\pi(\boldsymbol{c}\,|\,\boldsymbol{\theta}_{-s},\,\boldsymbol{Y}) = K_1(\boldsymbol{\theta}_{-s},\boldsymbol{c}_{-s},\boldsymbol{Y})$ $K_2(\boldsymbol{c},\boldsymbol{\theta}_{-s},\boldsymbol{Y})$, where $K_2(\boldsymbol{c},\boldsymbol{\theta}_{-s},\boldsymbol{Y})$ has the expression

$$K_2(\boldsymbol{c},\boldsymbol{\theta}_{-s},\boldsymbol{Y}) = \underbrace{\Big(\prod_{r=1}^{N^*}\prod_{j\in E_r}\ell_j(\boldsymbol{Y}_j\,|\,\boldsymbol{\theta}_{(r)})\,G_0(d\boldsymbol{\theta}_{(r)})\Big)}_{(*)}\Big(\int_{\boldsymbol{\Theta}}\ell_s(\boldsymbol{Y}_s\,|\,\boldsymbol{\theta}_s)\,G_0(d\boldsymbol{\theta}_s)\Big)\pi(\boldsymbol{c}\,|\,\boldsymbol{c}_{-s}) \quad \text{(A.9)}$$

if $\boldsymbol{c} = \{s\}\cup\boldsymbol{c}_{-s}$, and

$$K_2(\boldsymbol{c},\boldsymbol{\theta}_{-s},\boldsymbol{Y}) \quad \text{(A.10)}$$
$$= \Big(\prod_{\substack{r=1\\ r\neq r_0}}^{N^*}\prod_{j\in E_r}\ell_j(\boldsymbol{Y}_j\,|\,\boldsymbol{\theta}_j)G_0(d\boldsymbol{\theta}_{(r)})\Big)\Big(\prod_{j\in E_{r_0}\cup\{s\}}\ell_j(\boldsymbol{Y}_j\,|\,\boldsymbol{\theta}_{(r_0)})G_0(d\boldsymbol{\theta}_{(r_0)})\Big)\pi(\boldsymbol{c}\,|\,\boldsymbol{c}_{-s})$$

if $\boldsymbol{c} = (E_1, E_2, \ldots, E_{r_0}\cup\{s\}, \ldots, E_{N^*})$ for $r_0 = 1,\ldots,N^*$; in the above,

$$K_1(\boldsymbol{\theta}_{-s},\boldsymbol{c}_{-s},\boldsymbol{Y}) = \Big(\sum_{\boldsymbol{c}} K_2(\boldsymbol{c},\boldsymbol{\theta}_{-s},\boldsymbol{Y})\Big)^{-1}$$

is the normalizing constant with sum ranging over the appropriate (N^*+1) partitions of $\boldsymbol{c}$, and $\pi(\boldsymbol{c}\,|\,\boldsymbol{c}_{-s})$ is the conditional probability of the partition $\boldsymbol{c}$ given partition $\boldsymbol{c}_{-s}$ for $S\setminus\{s\}$.

Theorem 2. *Let G_0 be a prior (either proper or improper) on $\boldsymbol{\Theta}$ for which the posterior is proper. Then,*

$$\pi(\boldsymbol{\theta}_s\,|\,\boldsymbol{\theta}_{-s},\,\boldsymbol{Y}) = \frac{q_0 G_0^*(d\boldsymbol{\theta}_s) + \sum_{r=1}^{N^*} q_r\,\delta_{\boldsymbol{\theta}_{(r)}}}{q_0 + \sum_{r=1}^{N^*} q_r} \quad \text{(A.11)}$$

is a valid mixture distribution with mixing weights $q_0, q_1, \ldots, q_{N^*}$, where

$$G_0^*(d\boldsymbol{\theta}_s) = \frac{\ell_s(\boldsymbol{Y}_s\,|\,\boldsymbol{\theta}_s)\,G_0(d\boldsymbol{\theta}_s)}{\int_{\boldsymbol{\Theta}}\ell_s(\boldsymbol{Y}_s\,|\,\boldsymbol{\theta}_s)\,G_0(d\boldsymbol{\theta}_s)} \quad \text{(A.12)}$$

is the posterior distribution of $\boldsymbol{\theta}_s$ given that a new cluster is formed by site s,

$$q_0 = \left[\int_{\boldsymbol{\Theta}}\ell_s(\boldsymbol{Y}_s\,|\,\boldsymbol{\theta}_s)\,G_0(d\boldsymbol{\theta}_s)\right]\pi(\boldsymbol{c}\,|\,\boldsymbol{c}_{-s}) \quad \textit{and} \quad q_r = \ell_s(\boldsymbol{Y}_s\,|\,\boldsymbol{\theta}_{(r)})\,\pi(\boldsymbol{c}\,|\,\boldsymbol{c}_{-s}), \quad \text{(A.13)}$$

respectively, for $\boldsymbol{c} = \{s\}\cup\boldsymbol{c}_{-s}$ and $\boldsymbol{c} = (E_1, E_2, \ldots, E_r\cup\{s\}, \ldots, E_{N^*})$ for $r = 1,\ldots,N^*$.

Proof. We have $\pi(\boldsymbol{\theta}_s \,|\, \boldsymbol{\theta}_{-s},\, \boldsymbol{Y}) = \pi(\boldsymbol{\theta}_s \,|\, \boldsymbol{c}, \boldsymbol{\theta}_{-s},\, \boldsymbol{Y})\, \pi(\boldsymbol{c} \,|\, \boldsymbol{\theta}_{-s},\, \boldsymbol{Y})$. If $\boldsymbol{c} = \{s\} \cup \boldsymbol{c}_{-s}$, it follows that

$$\pi(\boldsymbol{\theta}_s \,|\, \boldsymbol{c}, \boldsymbol{\theta}_{-s},\, \boldsymbol{Y}) = \frac{\ell_s(\boldsymbol{Y}_s \,|\, \boldsymbol{\theta}_s)\, G_0(d\boldsymbol{\theta}_s)}{\int_{\boldsymbol{\Theta}} \ell_s(\boldsymbol{Y}_s \,|\, \boldsymbol{\theta}_s)\, G_0(d\boldsymbol{\theta}_s)}$$

by (A.8) and independence. Also, $\pi(\boldsymbol{c} \,|\, \boldsymbol{\theta}_{-s},\, \boldsymbol{Y}) = K_1(\boldsymbol{\theta}_{-s}, \boldsymbol{c}_{-s}, \boldsymbol{Y})\, K_2(\boldsymbol{c}, \boldsymbol{\theta}_{-s}, \boldsymbol{Y})$, where the first term (*) on the right hand side of (A.9) depends only on $\boldsymbol{\theta}_{-s}$ and hence is fixed. If $\boldsymbol{c} = (E_1, E_2, \ldots, E_r \cup \{s\}, \ldots, E_{N^*})$, we can write $K_2(\boldsymbol{c}, \boldsymbol{\theta}_{-s}, \boldsymbol{Y})$ as

$$K_2(\boldsymbol{c}, \boldsymbol{\theta}_{-s}, \boldsymbol{Y}) = \underbrace{\Big(\prod_{r=1}^{N^*} \prod_{j \in E_r} \ell_j(\boldsymbol{Y}_j \,|\, \boldsymbol{\theta}_{(r)})\, G_0(d\boldsymbol{\theta}_{(r)}) \Big)}_{(A)} \ell_s(\boldsymbol{Y}_s \,|\, \boldsymbol{\theta}_{(r)})\, \pi(\boldsymbol{c} \,|\, \boldsymbol{c}_{-s}),$$

where (A) is identical to the first term (*) in (A.9) of $K_2(\boldsymbol{c}, \boldsymbol{\theta}_{-s}, \boldsymbol{Y})$. Also, $\pi(\boldsymbol{\theta}_s \,|\, \boldsymbol{c}, \boldsymbol{\theta}_{-s},\, \boldsymbol{Y}) = \delta_{\boldsymbol{\theta}_r}$ where $\delta_{\boldsymbol{\theta}}$ is the point mass at $\boldsymbol{\theta}$. Summing these expressions and normalizing gives the form of the distribution of $\pi(\boldsymbol{\theta}_s \,|\, \boldsymbol{\theta}_{-s}, \boldsymbol{Y})$. We have suppressed the subscript s in q_0 for simplicity.

Choice of improper prior on α_0 in Remark 2: When an improper prior is used for α_0, it can be shown that the resulting marginal distribution for $\boldsymbol{Y}$, $m(\boldsymbol{Y})$, is improper. Let $\pi(\boldsymbol{c} \,|\, \alpha_0) \equiv \pi(\boldsymbol{c})$ in (A.1) and π^* be the improper prior on α_0. The impropriety of π^* results from $\int_0^\epsilon \pi^*(\alpha_0)\, d\alpha_0 = \infty$ for a sufficiently small ϵ, $\int_M^\infty \pi^*(\alpha_0)\, d\alpha_0 = \infty$ for a large M. In case $\alpha_0 \to 0$, $\pi(\boldsymbol{c}) \to 1$ for $\boldsymbol{c} = S$. Thus, $m(\boldsymbol{Y} \,|\, \alpha_0) \equiv \sum_{\boldsymbol{c}} \pi(\boldsymbol{Y} \,|\, \boldsymbol{c})\, \pi(\boldsymbol{c}) \geq \pi(\boldsymbol{Y} \,|\, S)\, \pi(S) \to k_1$ as $\alpha_0 \to 0$ and $m(\boldsymbol{Y}) = \int_0^\infty m(\boldsymbol{Y} \,|\, \alpha_0)\, \pi^*(\alpha_0)\, d\alpha_0 = \infty$. Similarly, if $\alpha_0 \to \infty$, $\pi(\boldsymbol{c}) \to 1$ for the singleton partition sets, $\boldsymbol{c}_0$ say. Then $m(\boldsymbol{Y} \,|\, \alpha_0) \equiv \sum_{\boldsymbol{c}} \pi(\boldsymbol{Y} \,|\, \boldsymbol{c})\, \pi(\boldsymbol{c}) \geq \pi(\boldsymbol{Y} \,|\, \boldsymbol{c}_0)\, \pi(\boldsymbol{c}_0) \to k_2$ as $\alpha_0 \to \infty$, resulting in $m(\boldsymbol{Y}) = \infty$.

Appendix B: Gibbs Updating Steps

(1) Update $\boldsymbol{\theta}_s$: The posterior of $\boldsymbol{\theta}_s$ conditional on $\boldsymbol{\theta}_{-s}$ is

$$\pi(\boldsymbol{\theta}_s \,|\, \boldsymbol{\theta}_{-s},\, \boldsymbol{Y},\, \boldsymbol{u}) = \frac{q_{s,0} G_0^*(d\boldsymbol{\theta}_s) + \sum_{r=1}^{N^*} q_{s,r}\, \delta_{\boldsymbol{\theta}_{(r)}}}{q_{s,0} + \sum_{r=1}^{N^*} q_{s,r}}, \tag{B.1}$$

where $q_{s,0}$, $q_{s,r}$, and $G_0^*(d\boldsymbol{\theta}_s)$ are given in (A.12) and (A.13). The conditional distribution $\pi(\boldsymbol{c} \,|\, \boldsymbol{c}_{-s})$ based on (A.1) is $\pi(\boldsymbol{c} \,|\, \boldsymbol{c}_{-s}) = \alpha_0/(\alpha_0 + N - 1)$ if $\boldsymbol{c} = \{s\} \cup \boldsymbol{c}_{-s}$ and $\pi(\boldsymbol{c} \,|\, \boldsymbol{c}_{-s}) = N_r/(\alpha_0 + N - 1)$, with N_r denoting the number of elements in E_r, where $\boldsymbol{c}_{-s} = \cup_{j=1}^{N^*} E_j$. Thus, we have

$$q_{s,0} = \alpha_0 \sum_{k=0}^{k^*} \sum_{(n_1, \ldots, n_{k+1})} \exp\left[H(n_1, \ldots, n_{k+1})\right] \frac{n_0!}{n_1! \cdots n_{k+1}!} \left(\frac{1}{k+1}\right)^{n_0} p(k), \tag{B.2}$$

where

$$H(n_1,\dots,n_{k+1}) = \log\left\{\int_{\mathbb{R}^{K_s+1}}\int_0^\infty f(\boldsymbol{Y}_s\,|\,\boldsymbol{\beta}^{(s)},\sigma_s^2)\pi(\sigma_s^2)\,d\sigma_s^2 d\boldsymbol{\beta}^{(s)}\right\}$$
$$= \frac{n-(2K_s+2)}{2}\log\left(\frac{b_\sigma}{2\pi}\right)+\log\Gamma\left(\frac{n^*}{2}\right)-\log\Gamma(a_\sigma)-\frac{1}{2}\log\left|X_0^{(s)T}X_0^{(s)}\right|$$
$$+\frac{K_s+1}{2}\log n^*-\frac{1}{2}\log\left|X_1^{(s)T}\Sigma_s X_1^{(s)}\right|$$
$$-\frac{n^*}{2}\log\left(1+(\boldsymbol{Y}_s-u_s\mathbf{1}_n)^T V^*(\boldsymbol{Y}_s-u_s\mathbf{1}_n)\right),$$

with $n^* = 2a_\sigma + n - 2K_s - 2$ and

$$V^* = \frac{1}{2}b_\sigma P_{X_{0,s}}(I_n - X_1^{(s)}(X_1^{(s)T}P_{X_{0,s}}X_1^{(s)})^{-1}X_1^{(s)T}P_{X_{0,s}}).$$

Expression (B.1) explicitly demonstrates the clustering capability of the functional DP prior. The current value of $\boldsymbol{\theta}_s$ can be selected to be one of the distinct $\boldsymbol{\theta}_{(r)}$ functions with probability $\sum_{r=1}^{N^*} q_{s,r}/(q_{s,0}+\sum_{r=1}^{N^*} q_{s,r})$, this positive probability being the reason for possible clustering of sites in terms of $\boldsymbol{\theta}_s$. Expression (B.1) also allows for a new $\boldsymbol{\theta}_s$ to be generated from the posterior distribution G_0^*. For a new $\boldsymbol{\theta}_s$, we generate $\boldsymbol{\alpha}^{(s)}$ accordingly from the posterior of $\boldsymbol{\alpha}^{(s)}$ given the change-points structure of $\boldsymbol{\theta}_s$. Similarly, when $\boldsymbol{\theta}_s$ is assigned to one of existing $\boldsymbol{\theta}_{(r)}$, we update $\boldsymbol{\alpha}^{(s)}$ accordingly from the posterior distribution given the change-points structure of $\boldsymbol{\theta}_{(r)}$.

(2) Update $\boldsymbol{\sigma}$: The update of σ_s^2 is carried out once $\boldsymbol{\theta}_s$ is obtained via (B.1). Regardless of whether $\boldsymbol{\theta}_s$ is a new value or an existing $\boldsymbol{\theta}_{(r)}$, for each site $s = 1,\dots,N$, the conditional posterior distribution of σ_s^2 given other parameters is

$$\pi(\sigma_s^2\,|\,\cdots) = \texttt{igamma}(a,b),$$

with $a = n/2 + a_\sigma$ and $b = ((1/2)(\boldsymbol{Y}_s - X_0^{(s)}\boldsymbol{\alpha}^{(s)} - X_1^{(s)}\boldsymbol{\beta}^{(s)})^T(\boldsymbol{Y}_s - X_0^{(s)}\boldsymbol{\alpha}^{(s)} - X_1^{(s)}\boldsymbol{\beta}^{(s)}) + b_\sigma^{-1})^{-1}$.

Here, $\underline{\boldsymbol{\theta}}$ uniquely determines the collection of parameters $(\boldsymbol{\beta},\boldsymbol{K},\boldsymbol{T})$. Since $\underline{\boldsymbol{\theta}}$ contains several identical components, it follows that the corresponding components of $(\boldsymbol{\beta},\boldsymbol{K},\boldsymbol{T})$ are identical to each other. We present the updating steps for the d distinct components of $(\boldsymbol{\beta},\boldsymbol{K},\boldsymbol{T})$: $(\boldsymbol{\beta}^{(r)},K_r,\boldsymbol{T}^{(r)})$ for $r = 1,\dots,d$. Let $\cup_{r=1}^d C_r$ be the partition of $\{1,\dots,N\}$ at the current update of the Gibbs sampler (thus, d is the number of distinct clusters).

(3) Update $(\boldsymbol{\beta}^{(r)},K_r,\boldsymbol{T}^{(r)})$: We first update K_r from the posterior marginal of K_r, and then update $\boldsymbol{T}^{(r)}\,|\,K_r$, and finally $\boldsymbol{\beta}^{(r)}\,|\,\boldsymbol{T}^{(r)}$, K_r from their respective conditional distributions. The posterior marginal probability of $K_r = k$ is proportional to

$$p(k) \sum_{(n_1,\ldots,n_{k+1})} \nu(n_1,\ldots,n_{k+1}), \tag{B.3}$$

with

$$\nu(n_1,\ldots,n_{k+1}) = \exp\{\tilde{H}(n_1,\ldots,n_{k+1})\} \frac{\Gamma(n_0+1)}{\prod_{l=1}^{k+1}\Gamma(n_l+1)} \left(\frac{1}{k+1}\right)^{n_0}, \tag{B.4}$$

where

$$\begin{aligned}
\tilde{H}(n_1,\ldots,n_{k+1}) &= \log\left\{\int_{\mathbb{R}^{k+1}} \prod_{s\in C_r} f(\boldsymbol{Y}_s \,|\, \boldsymbol{\beta}^{(r)}, \sigma_s^2)\, d\boldsymbol{\beta}^{(r)}\right\} \\
&= -\frac{(n-(k+1))|C_r|-(k+1))}{2}\log(2\pi) - \frac{n-k-1}{2}\sum_{s\in C_r}\log\sigma_s^2 - \frac{|C_r|}{2}\log\left|X_{0,r}^T X_{0,r}\right| \\
&\quad -\frac{1}{2}\log\left|X_{1,r}^T P_{X_{0,r}} X_{1,r}\right| - \frac{k+1}{2}\log\Big(\sum_{s\in C_r}\sigma_s^{-2}\Big) - \frac{1}{2}\sum_{s\in C_r}\sigma_s^{-2}\boldsymbol{Y}_s^T P_{X_{0,r}}\boldsymbol{Y}_s \\
&\quad +\frac{1}{2}\Big(\sum_{s\in C_r}\sigma_s^{-2}X_{1,r}^T P_{X_{0,r}}\boldsymbol{Y}_s\Big)^T\Big(\sum_{s\in C_r}\sigma_s^{-2}X_{1,r}^T P_{X_{0,r}} X_{1,r}\Big)^{-1}\Big(\sum_{s\in C_r}\sigma_s^{-2}X_{1,r}^T P_{X_{0,r}}\boldsymbol{Y}_s\Big).
\end{aligned}$$

The summation in (B.3) is over all non-negative integers $n_1, n_2, \ldots, n_{k+1}$ such that $\sum_{l=1}^{k+1} n_l = n_0 \equiv n-1-(k+1)w$. Obtaining the posterior probability of $K_r = k$ requires evaluation of (B.4) for each value of $k \geq 0$. This could potentially require a significant amount of computational time and drastically reduce the efficiency of the Gibbs chain, but this did not occur in our application due to the closed-form expression of $\tilde{H}$ using the flat prior π_0.

To update $\boldsymbol{T}^{(r)}$ given $K_r = k$, note that this is equivalent to updating $(n_1,\ldots,n_{k+1})$ with probabilities $p(n_1,\ldots,n_{k+1}) \propto v(n_1,n_2,\ldots,n_{k+1})$. This is carried out by exhaustively listing all such combinations and numerically computing the corresponding probabilities. The update of $\boldsymbol{\beta}^{(r)}$ given $\boldsymbol{T}^{(r)}$ and $K_r = k$ is done based on the conditional distribution $\mathcal{N}(\mu_r, \Sigma_r)$ with

$$\Sigma_r = \left(X_{1,r}^T P_{X_{0,r}} X_{1,r}\right)^{-1}\left(\sum_{s\in C_r}\sigma_s^{-2}\right)^{-1} \tag{B.5}$$

and

$$\mu_r = \Sigma_r\left(X_{1,r}^T P_{X_{0,r}}\sum_{s\in C_r}\sigma_s^{-2}\boldsymbol{Y}_s\right) \tag{B.6}$$

with the $k+1$ components of $\boldsymbol{\beta}^{(r)}$ generated independently of each other from their respective component densities.

(4) Update λ: λ is updated using

$$\pi(\lambda \mid \cdots) \propto \pi_2(\lambda) \prod_{s=1}^{N} p(K_s) \propto \frac{\lambda^{a_\lambda^*-1} e^{-\lambda/b_\lambda}}{(\sum_{l=0}^{k^*} \frac{\lambda^l}{l!})^N}, \tag{B.7}$$

with $a_\lambda^* = a_\lambda + \sum_{r=1}^{N^*} N_r^* k_r$, where k_r is the number of change-points corresponding to $\boldsymbol{\theta}_{(r)}$ in cluster C_r, and N_r^* is the number of sites in cluster C_r for $r = 1, \ldots, N^*$.

(5) Update α_0: For updating α_0 with $\pi_1(\alpha_0) \propto \alpha_0^{a_\alpha - 1} e^{-\alpha_0/b_\alpha}$, we utilize the two-step procedure of Escobar and West (1998): at the bth iteration: (1) draw κ from the Beta distribution $\texttt{beta}\left(\alpha_0^{(b-1)} + 1, N\right)$, and (2), draw $\alpha_0^{(b)}$ from the mixture density of two Gamma distributions

$$\pi_\kappa \texttt{gamma}(a_\alpha + N^*, (\frac{1}{b_\alpha} - \log(\kappa))^{-1}) + (1 - \pi_\kappa) \texttt{gamma}(a_\alpha + N^* - 1, (\frac{1}{b_\alpha} - \log(\kappa))^{-1}),$$

where N^* is the latest number of clusters and the membership probability is

$$\pi_\kappa = \frac{a_\alpha + N^* - 1}{N(1/b_\alpha - \log(\kappa))}.$$

(6) Update $\boldsymbol{\alpha}^{(s)}$: The update of $\boldsymbol{\alpha}^{(s)}$, given other parameters, is based on the conditional distribution $\mathcal{N}(\mu_{\boldsymbol{\alpha}}, \Sigma_{\boldsymbol{\alpha}})$ with

$$\Sigma_{\boldsymbol{\alpha}} = \sigma_s^2 \left(X_{0,r}^T X_{0,r}\right)^{-1} \quad \text{and} \quad \mu_{\boldsymbol{\alpha}} = \Sigma_{\boldsymbol{\alpha}} \sigma_s^{-2} X_{0,r}^T \boldsymbol{Y}_s.$$

References

American Cancer Society (2010). *Cancer Facts & Figures* 2010. American Cancer Society, Atlanta, GA.

Carlin, B. P., Gelfand, A. E., and Smith, A. F. M. (1992). Hierarchical Bayesian analysis of changepoint problems. *Appl. Statist.* **41**, 389-405.

Dass, S. C., Lim, C. and Maiti, T. (2011). Change point analysis of cancer mortality rates for US states using functional Dirichlet processes. Technical Report RM 690, Department of Statistics and Probability, Michigan State University.

Escobar, M. D. and West, M. (1998). Computing Nonparametric Hierarchical Models. In *Practical Nonparametric and Semiparametric Bayesian Statistics* (Edited by D. Dey, P. Muller and D. Sinha).

Ferguson, T. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1**, 209-230.

Ferguson, T. (1974). Prior distributions on spaces of probability measures. *Ann. Statist.* **2**, 615-629.

Ghosh, P., Basu, S. and Tiwari, R. C. (2009). Bayesian analsysis of cancer rates from SEER Program using parametric and semiparametric jointpoint regression models. *J. Amer. Statist. Assoc.* **104**, 439-452.

Ghosh, P., Ghosh, K. and Tiwari, R. (2011). Bayesian approach to cancer-trend analysis using age-stratified Poisson regression models. *Statist. Med.* **30**, 127-139.

Kim, H. J., Fay, M. P., Feuer, E. J., and Midthune, D. N. (2000). Permutation tests for joinpoint regression with applications to cancer rates. *Statist. Med.* **19**, 335-351.

Kim, H. J., Fay, M. P., Yu, B., Barrett, M. J., and Feuer, E. J. (2004). Comparability of segmented line regression models. *Biometrics* **60**, 1005-1014.

Ries, L. A. G., Eisner, M. P., Kosary, C. L., Hanley, B. F., Miller, B. A., Clegg, L. and Edwards, B. K. (2002). SEER Cancer Statistics Review. National Cancer Institute, Bethesda, MD. Available at `http://seer.cancer.gov/csr/1973-19995`.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statist. Sinica* **4**, 639-650.

Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) SEER*Stat Database, "Mortality - All COD, Aggregated With State, Total U.S. (1969-2007) <Katrina/Rita Population Adjustment>", National Cancer Institute, DCCPS, Surveillance Research Program, Cancer Statistics Branch, released June 2010. Underlying mortality data provided by NCHS (www.cdc.gov/nchs).

Tiwari, R. C., Cronin, K. A., Davis, W., Feuer, E. J., Yu, B. and Chib, S. (2005). Bayesian model selection for joinpoint regression with application to age-adjusted cancer rates. *J. Roy. Statist. Soc. Ser. C* **54**, 919-939.

Department of Fundamental & Applied Sciences, Universiti Teknologi PETRONAS, Bandar Seri Iskandar, 31750 Tronoh, Perak, Malaysia.

E-mail: saratcdass@gmail.com

Department of Statistics & Probability, Michigan State University, East Lansing, MI 48824, USA.

E-mail: lim@stt.msu.edu

Department of Statistics & Probability, Michigan State University, East Lansing, MI 48824, USA.

E-mail: maiti@stt.msu.edu

Department of Statistics, University of Chicago, Chicago, IL 60637, USA.

E-mail: zhangz19@galton.uchicago.edu

(Received October 2012; accepted April 2014)