Published in final edited form as:

Stat Med. 2014 October 15; 33(23): 4087–4103. doi:10.1002/sim.6221.

Clustering of trend data using joinpoint regression models

Hyune-Ju Kim,

Department of Mathematics, Syracuse University, Syracuse, New York 13244, U.S.A.

Jun Luo,

Information Management Services, Inc., 3901 Calverton Blvd., Suite 200, Calverton, MD 20705, U.S.A.

Jeankyung Kim,

Department of Statistics, Inha University, 253 Yonghyundoing, Namgu, Incheon, 402-751, Republic of Korea

Huann-Sheng Chen, and

Division of Cancer Control and Population Sciences, National Cancer Institute, 9609 Medical Center Drive MSC 9765 Room 4E348 Bethesda, MD 20892 - 9765

Eric J. Feuer

Division of Cancer Control and Population Sciences, National Cancer Institute, 9609 Medical Center Drive MSC 9765 Room 4E534 Bethesda, MD 20892 - 9765

Abstract

In this paper, we propose methods to cluster groups of two-dimensional data whose mean functions are piecewise linear into several clusters with common characteristics such as the same slopes. To fit segmented line regression models with common features for each possible cluster, we use a restricted least squares method. In implementing the restricted least squares method, we estimate the maximum number of segments in each cluster by using both the permutation test method and the Bayes Information Criterion (BIC) method, and then propose to use the BIC to determine the number of clusters. For a more effective implementation of the clustering algorithm, we propose a measure of the minimum distance worth detecting, and illustrate its use in two examples. We summarize simulation results to study properties of the proposed methods and also prove the consistency of the cluster grouping estimated with a given number of clusters. The presentation and examples in this paper focus on the segmented line regression model with the ordered values of the independent variable, which has been the model of interest in cancer trend analysis, but the proposed method can be applied to a general model with design points either ordered or un-ordered.

Keywords

Joinpoint regression; Clustering; Permutation test; Bayes information criterion; Minimum distance worth detecting

1 Introduction

Statistical similarity of objects has been studied in many different contexts by using various statistical procedures. A situation that received a considerable attention in the past decade is where each object is composed with a series of two-dimensional data points that can be described by a time-series model or a regression model, and clustering such objects has been of interest in many applications. An important issue in clustering time-series data has been to identify a good measure of (dis)similarity between the two or more time-series data sets, and a wide range of similarity measures were proposed in literature [1, 2, 3]. Using the similarity measure of choice, then, a classical clustering algorithm can be applied to a group of objects. Another approach proposed in literature is to use a model based clustering, more specifically to include a cluster membership variable in the model and estimate the cluster membership distribution using an EM algorithm or a Bayesian posterior distribution. For example, Qin and Self [4] proposed a model-based clustering method to cluster genes with a similar regression relationship to the covariate(s).

This paper is motivated from the question on how to identify and combine the neighboring age groups for which cancer incidence or mortality rates follow the same trend. For cancer trend analysis, Kim et al. [5] proposed a joinpoint regression model to describe trend changes, and it has been successfully applied in various trend analyses. The joinpoint regression model assumes that its regression mean function is piecewise linear and the segments are continuously connected at unknown change-points. For a non-linear model such as a polynomial model, the slope of the trend changes continuously and it makes the interpretation of when trend changes occur more difficult, compared to the joinpoint regression model. With an aim to cluster age groups with similar cancer incidence/mortality trends together and to summarize cluster characteristics using joinpoint regression models, our focus in this paper would be on how to cluster joinpoint regression models, but the general idea of this paper can be applied to various other models with the independent variable ordered or un-ordered.

To describe cancer rate trend changes, Kim et al. [5] used a grid search to obtain a segmented line fit for a model with a given number of change-points and proposed a permutation test procedure to select the number of change-points. They implemented the procedure in Joinpoint software available at the National Cancer Institute website (<http://surveillance.cancer.gov/joinpoint>), and enhanced features such as a continuous fitting, sequential stopping rules, Bayes Information Criterion, etc. have been added since its first release in 1998. Kim et al. [6] extended this method to compare two groups of data that follow joinpoint regression models and discussed how to apply two group comparability tests to more than two group situations.

Our specific aim in this paper is to cluster groups of two-dimensional data points, whose mean functions are expressed as joinpoint regression mean functions, into several clusters sharing common characteristics such as the same slopes. We first consider the model with a fixed number of clusters and fit the model by using a restricted least squares method and determine the best grouping. Using similar arguments of Kim and Kim [7], we prove that the estimated grouping converges to the true grouping. In doing so, we present detailed

arguments for the case where the objects are in order such as five-year age groups in cancer rate analysis and one wishes to cluster neighboring age groups together, and extend the result to a general situation with un-ordered objects. We then discuss how to select the number of clusters and how to apply the proposed method in practice.

This paper is organized as follows. In Section 2, we present the mathematical model and formulate the restricted least squares estimates. Section 3 includes the asymptotic results. In Sections 4 and 5, we present methods to select the number of clusters, summarize the simulation results to examine the performance of the proposed procedure, and propose a practical method to produce a more parsimonious grouping. Section 6 includes examples, and concluding remarks are given in Section 7.

2 Model and Fitting

Suppose that G objects come from M populations and that each object is presented as a multidimensional random vector that follows a segmented line regression model. The objects in each population are assumed to share some common characteristics, and our interest in this paper is on the cases where the objects in each population have the identical or parallel mean functions. That is, we observe n_g pairs of data points $\{(x_{g,1}, y_{g,1}), \dots, (x_{g,n_g}, y_{g,n_g})\}$ for $g = 1, 2, \dots, G$ and assume that $\mu_{g,i} = E(y_{g,i}|x_{g,i}) = \beta_{g,0} + \beta_{g,1}x_{g,i} + \delta_{g,1}(x_{g,i} - \tau_{g,1})^+ + \dots + \delta_{g,\kappa_g}(x_{g,i} - \tau_{g,\kappa_g})^+$ for $i = 1, \dots, n_g$, where $a^+ = \max(0, a)$ and κ_g is the unknown number of change-points for the object g . We also assume that $y_{g,i} = \mu_{g,i} + \epsilon_{g,i}$, where the $\epsilon_{g,i}$'s are independent with variance σ_g^2 and have a symmetric distribution. Since the G objects are from M populations, there exists an M -partition of $\{1, 2, \dots, G\}$, $\pi = \{\pi_1, \pi_2, \dots, \pi_M\}$, such that

- i. $\pi_m \subset \{1, 2, \dots, G\}$ for $m = 1, \dots, M$,
- ii. $\pi_i \cap \pi_j = \emptyset$ for $i \neq j$, and
- iii. $\cup_{m=1}^M \pi_m = \{1, 2, \dots, G\}$.

In order to fit the model at a given M , consider the set of all possible partitions of $\{1, 2, \dots, G\}$ and call it $\Omega = \{\pi_1, \pi_2, \dots, \pi_L\}$, where each element of Ω partitions the G objects into M clusters and L is the number of all possible partitions. Then, we estimate the model parameters including π by minimizing the object function,

$$\sum_{m=1}^M \sum_{j \in \pi_m} w_{m,j} \left\| \mathbf{y}_{m,j} - \boldsymbol{\mu}_{m,j} \right\|^2,$$

where $\pi = \{\pi_1, \dots, \pi_M\}$, the $w_{m,j}$ denote the weights such that $\text{Var}(w_{m,j}y_{m,j}) = \sigma^2 I$ for all m and j , $\mathbf{y}_{m,j}$ and $\boldsymbol{\mu}_{m,j}$ denote the observation and its mean vectors for the j -th object in the m -th cluster of the partition π , respectively, and $\|\cdot\|$ is the Euclidean norm. How to assign $w_{m,j}$ will be discussed later. If our goal is to cluster G objects into M clusters where each cluster contains the objects whose mean functions are equal, then at each $m = 1, \dots, M$,

$$(\kappa_g, \tau_{g,1}, \dots, \tau_{g,\kappa_g}, \beta_{g,0}, \beta_{g,1}, \delta_{g,1}, \dots, \delta_{g,\kappa_g}) \equiv (\kappa^{(m)}, \tau_1^{(m)}, \dots, \tau_{\kappa^{(m)}}^{(m)}, \beta_0^{(m)}, \beta_1^{(m)}, \delta_1^{(m)}, \dots, \delta_{\kappa^{(m)}}^{(m)})$$

for all $g \in \pi_m$. If we are interested in clustering together the objects whose mean functions are parallel, then at each $m = 1, \dots, M$,

$$(\kappa_g, \tau_{g,1}, \dots, \tau_{g,\kappa_g}, \beta_{g,1}, \beta_{g,1}, \delta_{g,1}, \dots, \delta_{g,\kappa_g}) \equiv (\kappa^{(m)}, \tau_1^{(m)}, \dots, \tau_{\kappa^{(m)}}^{(m)}, \beta_1^{(m)}, \delta_1^{(m)}, \dots, \delta_{\kappa^{(m)}}^{(m)})$$

for all $g \in \pi_m$. Figure 1 illustrates a situation with two clusters where two objects in each cluster share the same trend, that is, the parallel mean functions with the same change-point and slopes, which is usually of our interest in cancer trend analysis. The left panel includes the four objects where the objects denoted by $\circ$ and $*$ are from one population with

$\kappa^{(1)}=1, \tau_1^{(1)}=6, \beta_1^{(1)}=0.15$, and $\delta_1^{(1)} = -0.18$, and the two objects denoted by $\bullet$ and $\circ$ come

from a different population with $\kappa^{(2)}=1, \tau_1^{(2)}=8, \beta_1^{(2)}=0.15$, and $\delta_1^{(2)} = -0.13$. The right panel shows the data generated with the same mean functions, but with a smaller value of σ_g , which shows a difference between the two clusters more clearly. Data sets in the left panel will be used later in our simulation study to assess the performance of the proposed method, and it will be observed that the proposed method performs very well for a data set with noise of this size.

For the parametrization describe above, we now fit the model using a restricted least squares method. With $\hat{\mu}_{m,j}(\pi_l)$ denoting the mean vector fitted by the weighted least squares (WLS) method with the weight, $w_{m,j}$, at a partition $\pi_l = \{\pi_{l,1}, \dots, \pi_{l,M}\}$, we correspondingly estimate the cluster grouping as $\hat{\pi} = (\hat{\pi}_1, \dots, \hat{\pi}_M)'$ such that

$$Q(\hat{\pi}) = \min_{l=1, \dots, L} \sum_{m=1}^M \sum_{j \in \pi_{l,m}} w_{m,j} \|\mathbf{y}_{m,j} - \hat{\mu}_{m,j}(\pi_l)\|^2 = \sum_{m=1}^M \sum_{j \in \hat{\pi}_m} w_{m,j} \|\mathbf{y}_{m,j} - \hat{\mu}_{m,j}\|^2.$$

A situation that we are particularly interested in and will pursue first is the case where G objects are in order and clusters are to be formed over the neighboring objects. For example, in cancer trend studies, data are collected in different age groups: $0-4, 5-9, \dots, 85+$, and the trends in nearby age groups tend to be more similar and can be grouped into one cluster. For such ordered partition, $\pi = \{\{1, \dots, \rho_1\}, \{\rho_1 + 1, \dots, \rho_2\}, \dots, \{\rho_{M-1} + 1, \dots, G\}\}$, where $\rho = (\rho_1, \dots, \rho_{M-1})'$ denotes the breaking points where the first $M-1$ clusters end. Our aim is then to estimate the breaking point parameters, $\rho = (\rho_1, \dots, \rho_{M-1})'$, as $\hat{\rho}$ such that

$$Q(\hat{\rho}) = Q(\hat{\rho}_1, \hat{\rho}_2, \dots, \hat{\rho}_{M-1}) = \min_{d \in \{d = (d_1, \dots, d_{M-1}) : d_1 < d_2 < \dots < d_{M-1}\}} Q(d),$$

where the object function is

$$Q(\mathbf{d}) = Q(d_1, d_2, \dots, d_{M-1}) = \sum_{m=1}^M \sum_{j=d_{m-1}+1}^{d_m} w_{m,j} \|\mathbf{y}_{m,j} - \hat{\boldsymbol{\mu}}_{m,j}(\mathbf{d})\|^2$$

with $d_0 = 0$ and $d_M = G$. To estimate $\boldsymbol{\rho}$, we note that $Q(\mathbf{d})$ decreases as the number of change-points $\kappa^{(m)}$ increases, so we find the restricted least squares estimates of $\kappa^{(m)}$ as $kmax^{(m)}$, where the $kmax^{(m)}$ are predetermined constants such that $\kappa^{(m)} \leq kmax^{(m)}$. In the next section, we show that $\hat{\boldsymbol{\rho}}$ estimated under the condition that $\kappa^{(m)} \leq kmax^{(m)}$ or equivalently $\hat{\kappa}^{(m)} = kmax^{(m)}$ is consistent. We also discuss how to choose $kmax^{(m)}$ in Section 4.

Remark 1

In practice, the weight w_g is often unknown, and we may use w_g such that

$\frac{1}{w_g} = \|\mathbf{y}_g - \hat{\boldsymbol{\mu}}_g(kmax_g)\|^2 / (n - q)$, where $\hat{\boldsymbol{\mu}}_g(kmax_g)$ is the mean vector fitted with $\kappa_g = kmax_g$ for an appropriately chosen $kmax_g$ and q is the number of unknown free parameters.

For the j -th object in the m -th cluster at a given partition $\boldsymbol{\pi}$, we may also use

$w_{m,j} = \lim_{r \rightarrow \infty} w_{m,j}^{(r)}(kmax^{(m)})$, if the limit exists, such that for $r = 1, 2, \dots$,

$$\frac{1}{w_{m,j}^{(r)}(kmax^{(m)})} = \|\mathbf{y}_{m,j} - \tilde{\boldsymbol{\mu}}_{m,j}^{(r-1)}(kmax^{(m)})\|^2 / (n - q),$$

where $\tilde{\boldsymbol{\mu}}_{m,j}^{(0)}(kmax^{(m)})$ denotes the mean vector estimated by the ordinary least squares method assuming $\kappa^{(m)} = kmax^{(m)}$ and the $\tilde{\boldsymbol{\mu}}_{m,j}^{(r)}(kmax^{(m)})$ are iteratively updated using $w_{m,j}^{(r)}(kmax^{(m)})$.

Remark 2

$Var(\mathbf{y}_1, \dots, \mathbf{y}_G) = \begin{pmatrix} \sigma_1^2 V_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_G^2 V_G \end{pmatrix}$ where $V_g = I_{n_g \times n_g}$ for $g = 1, \dots, G$, the weighted least squares fit can be obtained by using a weight matrix W which can be estimated following a similar iterative method described in Remark 1.

3 Asymptotics

Assume that there are M clusters with M given, and for $m = 1, \dots, M$, let $\boldsymbol{\theta}^{(m)}$ and $\boldsymbol{\tau}^{(m)}$ be the regression coefficients and the change-points for the m -th cluster, respectively. Recall that the partition, $\boldsymbol{\pi} = \{\pi_1, \dots, \pi_M\}$, indicates which cluster the object g belongs to. We first present the results for the case where G -objects are in order, mainly for notational simplicity, and then will discuss its generalization to un-ordered situations. In the ordered situation where $\pi_m = \{\rho_{m-1} + 1, \rho_{m-1} + 2, \dots, \rho_m\}$ with $\rho_0 = 1$ and $\rho_M = G$, we let

$\theta^{(m)} = (\beta_0^{(m)}, \beta_1^{(m)}, \delta_1^{(m)}, \dots, \delta_{\kappa^{(m)}}^{(m)})'$ for the model with the identical mean function in a cluster, $\theta^{(m)} = (\beta_{\rho_{m-1}+1,0}, \dots, \beta_{\rho_m,0}, \beta_1^{(m)}, \delta_1^{(m)}, \dots, \delta_{\kappa^{(m)}}^{(m)})'$ for the objects in the m -th cluster, and $\tau^{(m)} = (\tau_1^{(m)}, \tau_2^{(m)}, \dots, \tau_{\kappa^{(m)}}^{(m)})'$ for the model with parallel mean functions.

Let

$$\theta = (\theta^{(1)'}, \theta^{(2)'}, \dots, \theta^{(M)'})', \tau = (\tau^{(1)'}, \tau^{(2)'}, \dots, \tau^{(M)'})', \quad \text{and} \quad \gamma = (\theta', \tau')'.$$

We also let $\kappa = (\kappa^{(1)}, \kappa^{(2)}, \dots, \kappa^{(M)})'$ and express the parameter set as a parameter triple (ρ, κ, γ) . Note that γ depends on $\rho = (\rho_1, \dots, \rho_{M-1})$ and κ .

To study asymptotic properties of the least squares estimators of ρ and γ , we assume the following conditions for the independent variable x . Note that $\tau^{(0)}$ denote the true change-points, and we also assume that the x 's are scaled such that $\min_{g,i} x_{g,i} = 0$, $\max_{g,i} x_{g,i} = 1$, without loss of generality, and the same number of data points for any object g , $n_g = n$.

Assumption 1

For each object g , the independent variable x_g has a positive and continuous density function in any small neighborhoods of 0, 1 and true change-points of all objects. Also x_g is independent of the error, ϵ_g .

For the case of nonrandom design points, the following assumption can replace Assumption 1. Note that the data spacing should be of order $O(1/n)$ to satisfy this assumption.

Assumption 1'

Let $\{h_n\}$ be a sequence of constants for which $O(1/n) \leq h_n$. For each g , the number of data points in any small neighborhoods with volume h_n of 0, 1 and true change-points of all objects is of order at least nh_n .

When the G -objects are in order, denote the mean function for the parallel model as

$$\mu_g(\rho, \kappa, \gamma, x) = \sum_{m=1}^M \left\{ \beta_{g,0} + \beta_1^{(m)} x + \delta_1^{(m)} (x - \tau_1^{(m)})^+ \cdots + \delta_{\kappa^{(m)}}^{(m)} (x - \tau_{\kappa^{(m)}}^{(m)})^+ \right\} I(\rho_{m-1} + 1 \leq g \leq \rho_m),$$

where $\rho_0 = 0$, $\rho_M = G$, $\tau_{g,0} = 0$, $\tau_{g,\kappa^{(m)}+1} = 1$, and $I(A) = 1$ if A is true and zero otherwise. The $\mu_g(\rho, \kappa, \gamma, x)$ for the equal mean function model are defined similarly.

Let

$$Q_n(\rho, \kappa, \gamma, w) = \frac{1}{nG} \sum_{g=1}^G w_g \sum_{i=1}^n (y_{g,i} - \mu_g(\rho, \kappa, \gamma, x_{g,i}))^2,$$

and consider

$$G_n(\boldsymbol{\rho}, \boldsymbol{\kappa}, \boldsymbol{\gamma}, \boldsymbol{w}) = Q_n(\boldsymbol{\rho}, \boldsymbol{\kappa}, \boldsymbol{\gamma}, \boldsymbol{w}) - Q_n(\boldsymbol{\rho}^{(0)}, \boldsymbol{\kappa}^{(0)}, \boldsymbol{\gamma}^{(0)}, \boldsymbol{w}),$$

where $\boldsymbol{w} = (w_1, \dots, w_G)'$ and $(\boldsymbol{\rho}^{(0)}, \boldsymbol{\kappa}^{(0)}, \boldsymbol{\gamma}^{(0)})$ denote the true parameter vector. Let $\hat{\boldsymbol{\rho}} = \hat{\boldsymbol{\rho}}_{kmax}$ minimize $\inf_{\boldsymbol{\rho}} Q_n(\boldsymbol{\rho}, kmax, \boldsymbol{\gamma}, \boldsymbol{w})$, where $kmax = (kmax^{(1)}, kmax^{(2)}, \dots, kmax^{(M)})'$, are predetermined constants such that $kmax^{(m)} \rightarrow \kappa^{(m), (0)}$ for each $m = 1, \dots, M$.

To establish the consistency of $\hat{\boldsymbol{\rho}}$, we first prove the following lemma.

Lemma—Under Assumption 1 (or Assumption 1'), for any $\delta > 0$,

$$\lim_{n \rightarrow \infty} \inf_{\boldsymbol{\rho} \neq \boldsymbol{\rho}^{(0)}} \inf_{\boldsymbol{\gamma}} EG_n(\boldsymbol{\rho}, kmax, \boldsymbol{\gamma}, \boldsymbol{w}) > 0.$$

Theorem 1—Under Assumption 1 (or Assumption 1'),

$$\lim_{n \rightarrow \infty} P(\hat{\boldsymbol{\rho}} = \boldsymbol{\rho}^{(0)}) = 1.$$

See Appendix for the proofs of Lemma and Theorem 1.

Remark 1

Lemma and Theorem 1 hold with estimated weights if $E(w_{g \in g, i}) = 0$ and $1/w_g$ converges to σ_g^2 in probability. These two conditions are satisfied with w_g in Remark 1 of Section 1 such that $\frac{1}{w_g} = \|\boldsymbol{y}_g - \hat{\boldsymbol{\mu}}_g(kmax_g)\|^2 / (n - q)$.

Remark 2

In the case of un-ordered objects, Theorem 1 can be re-written for $\hat{\boldsymbol{\pi}}$ such that

$$\lim_{n \rightarrow \infty} P(\hat{\boldsymbol{\pi}} = \boldsymbol{\pi}^{(0)}) = 1.$$

Remark 3

Assumption 1' holds in the equally spaced time point case, which is our interest with cancer rate analysis. Letting n get bigger in these asymptotic results means that observations are measured more frequently, and the large sample results will hold for reasonably frequently measured data points, which could be yearly measurements sometimes or monthly measurements in some other cases.

4 Selection of M and Simulations

In Sections 2 and 3, we used predetermined values of $kmax^{(m)}$ and assumed that M is given, which is usually not known in practice. In this section, we propose a method to select M as well as discuss how to choose $kmax^{(m)}$, and conduct simulations to compare accuracies of

various selection procedures. First, we propose a data-driven method to find $kmax^{(m)}$ following the suggestion made in Kim et al. [6]. In the ordered object case with $(\rho_1, \dots, \rho_{M-1}) = (d_1, \dots, d_{M-1})$, we use the maximum number of change-points for the m -th cluster estimated as

$$kmax^{(m)} = \max \left(\hat{\kappa}_{d_{m-1}+1}, \hat{\kappa}_{d_{m-1}+2}, \dots, \hat{\kappa}_{d_m}, \hat{\kappa}_m^{(0)} \right),$$

where $\hat{\kappa}_g$ is the number of change-points selected for object g and $\hat{\kappa}_m^{(0)}$ is the common number of change-points estimated for all of the $d_m - d_{m-1}$ objects in the m -th cluster.

To determine the number of change-points and the number of clusters, we considered both the hypothesis testing method based on permutation tests and the Bayes Information Criterion (BIC) selection method. The permutation procedure to select the number of change-points, κ , is described in Kim et al. [5] and its idea is summarized here. To estimate the number of change-points, κ , we start with testing the null hypothesis of $H_0 : \kappa = k_0$ against the alternative hypothesis of $H_1 : \kappa = k_1 (> k_0)$. When H_0 is rejected and $k_1 > k_0 + 1$, we test for $\kappa = k_0 + 1$ versus $\kappa = k_1$, while we proceed to the test of $\kappa = k_0$ versus $\kappa = k_1 - 1$ when H_0 is not rejected and $k_1 - 1 > k_0$. We repeat testing until we test the null hypothesis of $H_0 : \kappa = k$ against the alternative of $H_1 : \kappa = k + 1$ for some k ($k_0 < k < k_1$). Due to a difficulty to find an analytic distribution of the test statistic, Kim et al. [5] proposed to use the permutation distribution of the test statistic to estimate its p-value at each step, and the significance level at each step is adjusted to control the over-fitting probability to a conventional level of α . This procedure is directly applied to the clustering context to estimate the number of clusters M using the hypothesis testing method. For the BIC, we used the following functions and chose k and m to minimize $BIC(k)$ and $\widetilde{BIC}(m)$, respectively:

$$\begin{aligned} BIC(k) &= \log \left(\frac{RSS_k}{n} \right) + \frac{2k}{n} \log n, \\ \widetilde{BIC}(m) &= \log \left(\frac{\widetilde{RSS}_m}{N} \right) + \frac{p}{N} \log N, \end{aligned}$$

where RSS_k denotes the residual sum of squares for an object of size n fitted for a joinpoint model with k change-points, $\widetilde{RSS}_m$ denotes the total sum of squared errors for G objects separated into m clusters, $N = \sum_{g=1}^G n_g$, and p is defined as

$$p = (m-1) + \sum_{j=1}^n \{2k_j + 1 + (d_j - d_{j-1})\} = (m-1) + 2 \sum_{j=1}^m k_j + m + G,$$

with $(\hat{\rho}_1, \dots, \hat{\rho}_{M-1}) = (d_1, \dots, d_{M-1})$, $d_0 = 0$ and $d_m = G$. We denote the values of k and m that minimize $BIC(k)$ and $\widetilde{BIC}(m)$ as $\hat{\kappa}$ and $\hat{M}$, respectively.

It has been found in Kim et al. [8] that the permutation test tends to be conservative compared to the BIC method in finding $\hat{\kappa}$, and similar results are expected for the selection

of $kmax^{(m)}$ and the number of clusters, M . Although selecting a most parsimonious model is usually our goal in cancer trend analysis, the use of permutation tests both for finding the maximum number of change-points $kmax^{(m)}$ and the number of clusters M (JPerm-CPerm method) is too time consuming and its performance was not very satisfactory in our preliminary simulation study. Combinations of model selection methods such as JPerm-CBIC, JBIC-CPerm, and JBIC-CBIC to find $kmax^{(m)}$ and $\hat{M}$ have been also considered in our preliminary simulation study, and we observed that the combinations of JPerm-CBIC and JBIC-CBIC perform a lot better with higher probabilities of correctly selecting M and correctly identifying the clusters than the JPerm-CPerm and the JBIC-CPerm combinations. This indicates that the permutation test to select the number of clusters may not have high power, although the method to select the number of change-points has a less significant effect on the overall performance, which could be explained by the proposed method to use $kmax^{(m)}$. The JPerm-CBIC and JBIC-CBIC combinations produced comparable probabilities of correct selections in most cases, and we used both the JPerm-CBIC and JBIC-CBIC combinations for the detailed study summarized in the tables below.

We considered the log-linear model with normally distributed errors:

$$y_{g,i} = \log r_{g,i} = \mu_{g,i} + \epsilon_{g,i},$$

where the r 's denote the original rates and the ϵ 's are independent and normally distributed with mean zero and variance σ_g^2 for $g = 1, \dots, G$, and $i = 1, \dots, 30$. In the simulation study, we estimated the probability of correctly selecting the number of clusters, M , and also the probability of correctly identifying clusters, $\{\pi_1, \dots, \pi_M\} = \{\{1, \dots, \rho_1\}, \{\rho_1 + 1, \dots, \rho_2\}, \dots, \{\rho_{M-1} + 1, \dots, G\}\}$. When the permutation method was used to choose $kmax^{(m)}$, we used the number of simulations, $N_{sim} = 619$, and the number of permutations, $N_{perm} = 199$, chosen following the suggestion of Boos and Zhang [9]. For the simulations with the JBIC-CBIC method, we used $N_{sim} = 619$.

Tables 1, 2, 3, and 4 summarize the simulation results for one, two, and three cluster situations, with various choices of σ_g and various cluster mean functions. Simulations are conducted to estimate P_M and P_C , the probabilities of correctly selecting the number of clusters, M , and of correctly identifying the clusters, respectively. In the tables, we considered the joinpoint regression mean functions motivated from actual cancer trends, Hodgkin's lymphoma for Table 1, brain cancer for Tables 2 and 3, and prostate cancer for Table 4. In Table 1, we see that three choices of σ_g from 0.1 to 0.5 do not influence P_M and P_C , for the case of $G = 6$ and $M = 1$. However, σ_g does influence P_M and P_C when $M = 2$ as shown in Cases II-1, II-2, II-3 and II-4 of Table 2. Table 2 summarizes the simulation results for $G = 6$ and $M = 2$, and boldface numbers indicate simulation parameter values different from the reference parameter values in Case II-1. Table 3 summarizes simulation results when the first cluster mean function is a joinpoint regression model with one change-point, which is the only difference from Case II-2 in Table 2. The boldface numbers in Table 3 indicate simulation parameter values of the first cluster different from those of the second cluster in the reference case of Case II-2. The mean functions of the two clusters in Cases II-2-1, II-2-2, II-2-3, and II-2-4 allow changes both in the slope parameters and the location

of the change-point, while the remaining cases in Table 3 considers the situations where two cluster mean functions are different only in the change-point or in the slope parameter. In Cases II-2-5 and II-2-6, two cluster mean functions share the same change-point, but the slopes after the change-point are different. Cases II-2-7 and II-2-8 consider the situation where two cluster mean functions share the same slope parameters, but the locations of the change-point are different. The simulation results indicate that the proposed method identifies clusters whose mean functions are different only in the change-point or only in the slope parameters as well, and its performance is reasonably good if the effect size is relatively large.

In Tables 2 and 3, we observe that P_M and P_C increase as the difference between the clusters gets big, and we note that P_M and P_C can be presented as monotonic functions of an effect size,

$$\Delta = \frac{(\boldsymbol{\eta}^{(1)} - \boldsymbol{\eta}^{(2)})' (\boldsymbol{\eta}^{(1)} - \boldsymbol{\eta}^{(2)})}{\frac{1}{\rho_1^2} \sum_{g=1}^{\rho_1} \sigma_g^2 + \frac{1}{(G-\rho_1)^2} \sum_{g=\rho_1+1}^G \sigma_g^2},$$

where $\boldsymbol{\eta}^{(m)} = \boldsymbol{\mu}^{(m)} - \bar{\boldsymbol{\mu}}^{(m)}$ with $\boldsymbol{\mu}^{(m)} = (\mu_m(x_1), \dots, \mu_m(x_n))'$, the m -th cluster mean vector, and $\bar{\boldsymbol{\mu}}^{(m)} = \frac{\sum_{i=1}^n \mu_m(x_i)}{n} (1, \dots, 1)'$. We also note in Table 3 that P_M and P_C are relatively lower than those in Table 2 although they still increase as Δ increases. This seems to be because the true mean function for the first cluster in Table 3 has one change-point, which may not have been correctly identified by the permutation or BIC procedure, while the probability of correctly specifying no-change-point model for the first cluster in Table 2 could be relatively high. Table 4 reports the simulation results for a three cluster situation.

5 Minimum Difference Worth Detecting (MDWD)

When we applied the clustering method proposed in this article to real data, we observed that the method is quite powerful in detecting even small differences among the clusters. To make the clustering more meaningful for practical use, we introduce a measure to provide a threshold for distinguishing two clusters. We call this measure the minimum difference worth detecting (MDWD). The MDWD given below is developed mainly to cluster the models with same slopes together, and a similar measure can be used when one's goal is to cluster the model with identical mean functions together. Specifically, the clustering is accomplished by the following procedure. For each pair of consecutive clusters, C_1 and C_2 , we first remove the estimated change-points found in each cluster from the original data points if they coincide with original data x -values, and denote the remaining data points $\{\tilde{x}_1, \dots, \tilde{x}_p\}$. That is,

$$\{\tilde{x}_1, \dots, \tilde{x}_p\} = \{x_1, \dots, x_n\} \setminus \{\hat{\tau}_{1,1}, \dots, \hat{\tau}_{1,k_1}, \hat{\tau}_{2,1}, \dots, \hat{\tau}_{2,k_2}\},$$

where $\{x_1, \dots, x_n\}$ are the data points and $\hat{\tau}_{l,j}$ is the j th change-point estimated for the cluster C_l , $l = 1, 2$. We then find the estimated annual percent change (APC) for the segment where $\tilde{x}_j$ lies for each object, and denote them $\hat{\zeta}_j^{(1)}$ and $\hat{\zeta}_j^{(2)}$ respectively. Note that the estimated APCs are $\hat{\zeta}_j^{(l)} = 100 \left(e^{\hat{\beta}_{(l)}} - 1 \right)$, where $\hat{\beta}_{(l)}$ is the estimated slope of the segment where $\tilde{x}_j$ lies for the l -th cluster. We now define the distance between the two clusters as

$$d(C_1, C_2) = \frac{\sum_{j=1}^p |\hat{\zeta}_j^{(1)} - \hat{\zeta}_j^{(2)}|}{p}.$$

If the distance $d(C_1, C_2)$ is smaller than the pre-determined MDWD, the two clusters C_1 and C_2 are deemed not meeting the MDWD criteria. More specifically, we adopt the following procedure that mixes clustering method and the MDWD criteria. For a given set of G objects,

Step 1. Set a maximum number of clusters, $M_{\max}$. For each $m \in \{1, \dots, M_{\max}\}$, apply the proposed clustering method to determine the best grouping of m clusters. Then, use the BIC method to estimate the number of clusters, which is denoted by $\hat{M}$.

Step 2. Set the MDWD, say at ϕ , for example. For $m = \hat{M}$, examine if the estimated adjacent clusters meet the MDWD criteria. That is, examine if $d(C_j, C_{j+1}) > \phi$ for all $j = 1, \dots, m - 1$.

Step 3. If $m = 1$ or all the neighboring pairs meet the MDWD criteria, then stop.

Step 4. If there are at least one neighboring pairs of the estimated clusters not meeting the MDWD criteria, then reduce m by one to $\hat{M} - 1$. Repeat Steps 2-4.

The procedure requires all distances of neighboring pairs of suggested clusters bigger than the MDWD so that the suggested one forms more practical clusters. Clustering is conducted minimizing the restricted sum of squared errors by fitting an identical/parallel mean function model and using a permutation procedure or the BIC selection method, and thus the individual mean functions of the objects in one cluster do not differ using a criteria of statistical significance. The size of these differences may be small, moderate, or large depending on the statistical power to detect differences. However, in cases where there is substantial statistical power to detect even small differences, the MDWD criterion may override the statistical criteria when the differences are smaller than an ad hoc cutoff value set based on practical consideration. Note that the implementation of the MDWD seeks for an optimal grouping of the objects at each m , and thus there is a possibility that the objects combined into one cluster for the situation with m clusters may get separated into different clusters when Steps 2-3 are repeated for the situation with $m-1$ clusters after Step 4. By repeating Steps 2-3 after finding a difference that does not meet the MDWD criterion, we are assured that the resulting clusters represent an optimal grouping, albeit under a smaller m .

6 Example

To demonstrate the proposed methods, we analyze cancer incidence using 9 registries from the National Cancer Institute's Surveillance Epidemiology and End Results (SEER) program representing 10% of the US population, and U.S. mortality data collected from the states by the National Center for Health Statistics. We will consider the US prostate cancer mortality rates and thyroid cancer incidence rates both from 1975-2009 for illustration of our methods.

Prostate cancer is the second most common cancer cause of death in the U.S. among men with an estimated 28,170 deaths in 2012 [10]. Figure 2 shows the usual age-adjusted trend for all combined. Mortality was rising slowly in the 1970's until the late 1980's when it started rising rapidly, leveled off briefly, and then has fallen since 1994. The decline in mortality is associated with advances in treatment, and possibly PSA screening for prostate cancer, which started in the early 1990's (See Etzioni et al. [11]). The rise and fall in prostate cancer mortality around the time of the rapid introduction of PSA testing in the general population are consistent with a hypothesis that a fixed percent of the rising and falling pool of recently diagnosed patients who die of other causes may be mislabeled as dying of prostate cancer [12].

To determine if the trends are consistent across all ages we applied the clustering algorithm proposed. Since prostate cancer is a disease of older men and there are zero counts in some of the younger age groups, we collapsed 19 age groups 0 – 4, 5 – 9, ..., 80 – 84, and 85+ into 11 age groups by combining ages 0 – 39 into one group. We then used JPerm-CBIC method, the permutation procedure to fit the joinpoint models and the BIC to select the number of clusters. Using the clustering procedure, the 11 age groups are grouped into 6 clusters: ages 0–54, 55–64, 65 – 74, 75 – 79, 80 – 84, and 85+. Figure 3 shows the fitted Joinpoint trends as well as the observed mortality rates in each age group. With so many prostate cancer deaths occurring each year there is sufficient statistical power to detect even small differences in trends. While the trends differ somewhat in each of the 6 clusters, it was felt that the differences between some of the clusters were small and would be difficult to interpret, so we applied a MDWD criteria.

We considered a few different MDWD's. For MDWD = 0.5%, the age groups are clustered into the same six clusters as in MDWD = 0% , while with a MDWD of 1.0% and 1.5% there were two clusters (ages 0 – 84 and 85+), and with a MDWD of 2% all of the age groups were combined. Based on our knowledge of cancer trends and what is considered a "large difference," we decided that a threshold of 1% was most reasonable, and the trends based on two clusters are shown in Figure 4. The trends in the two clusters are not parallel, as the age 0 – 84 group has a smaller APC until 1987. Since 1987, both clusters show a sudden increase, but the trend for 0 – 84 age group started turning downward since 1991 and decreased even further since 1994, while the downward trend in age 85+ group did not appear until 1993.

The second example is thyroid cancer incidence rates for white females. Thyroid cancer is the 5th most common cancer among women in 2012 [10] with an estimated 43,201 cases in the US. It is the fastest rising cancer in the U.S. although it is unclear whether thyroid cancer

is really on the rise or whether doctors were just getting better at detecting smaller tumors that may have gone unnoticed in previous decades (See Davies and Welch [13]). Figure 5 shows the trend for the overall age-adjusted rates, which shows a rise of over 5% a year from 1993 to 1999 and over 7% a year from 1999 onward. We use the clustering methodology to determine if the rise differs across age groups. Since SEER 9 represents 10% of the U.S. population, there are just over 4,000 cases per year in these registries (with considerably less statistical power to detect differences than for the 28,000 prostate cancer deaths). In the original 19 age groups, some of the younger age groups have zero counts, thus we lumped them into one age group. In particular, age groups 0 – 4, 5 – 9, and 10 – 14 are combined into one group 0 – 14, with 16 age groups remaining. We found that there are only 4 distinguishable clusters: 0 – 24, 25 – 29, 30 – 79, and 80+. The trends for each cluster are shown in Figure 6. In the three clusters 0 – 24, 25 – 29, and 80+, only one joinpoint is selected. For age group 0 – 24, the change-point is located at 2007, where a sudden dramatic increase in cancer incidence is observed. For age group 25 – 29, the incidence trend is stable or slightly decreasing until 1987 when a slow upward trend is shown. For age group 80+, the incidence rate is stable until 1997 when the incidence rate increases at 5% per year. The age group 30 – 79 shows an increasing trend from 1980. Because this is the age group where the bulk of the cases are, there is more statistical power to detect changes in trend. There are 5 change-points found in this age group. From 1980 to 1993 there was moderate increase of 2.76% per year, followed a large increase of over 5% per year from 1993 to 2000, a large increase of over 11% per year from 2000 to 2003, and an slightly smaller increase of almost 7% per year from 2003 onward. Running a MDWD for 0.5%, 1.0%, 1.5% and 2.0% still yielded 4 clusters. Although overall the trends across these four age clusters look somewhat similar, the locations of the change-points and the annual percent changes between the clusters are very different. Analysis of this type can help isolate potential causes of the rise in thyroid cancer.

In these two examples, the MDWD performed as expected. In the case of prostate cancer, small differences between clusters (which were statistically significant) were collapsed using the MDWD criteria. For female thyroid cancer, the smaller sample size means that only reasonably large differences can be detected, and thus the MDWD criteria did not yield any further collapsing of the age groups. For each area of application, what is an appropriate MDWD is a subjective determination, and it is gained through experience in analyzing the data. For cancer we feel that a reasonable MDWD is approximately 1%, although in specific analysis situations this may differ.

7 Discussion

In this paper, we proposed a method to cluster groups of observations that follow joinpoint regression models into several clusters that share similar characteristics. We implemented various methods to select the common number of change-points for each set of data and also to select the number of clusters, and suggested to use the MDWD for more practical clustering. For these model selection problems, we considered combinations of the permutation test procedure and BIC, and recommend the methods of JPerm-CBIC or JBIC-CBIC as a faster way to cluster, which use the permutation procedure or the BIC to select the number of change-points, respectively, and the BIC for the number of clusters. Although

the presentation and examples of this paper focused on cancer trend analysis where the design points are the ordered time-points, the method works for situations with un-ordered design points, x , and the asymptotic results in Section 3 also consider such situations. In some situations, one may wish to cluster objects only according to the locations of change-points or based on the slope parameters regardless of the locations of change-points, and then the proposed method can be applied with a parametrization, different from those considered in Section 2 but incorporating the question of one's interests. In cancer trend analysis, however, it would be rarely of our interests to consider the clustering problem only based on the change-points or on the slope parameter. Also, note that the idea of the proposed method can be applied to other models such as a general linear/non-linear regression model as long as the mean vector for the j -th object in the m -th cluster is estimated and thus the residual sum of squares can be calculated, but the asymptotic results presented in this paper may not work in general.

Another method proposed in Kim et al. [6] to compare multiple objects is to use multiple pairwise tests, where a pairwise comparability test is used for each pair of objects to test if the two joinpoint mean functions are parallel or identical and then the significance levels are adjusted to incorporate multiple comparisons conducted. In our preliminary simulation study, we conducted simulations to compare the exhaustive clustering proposed in this paper with the multiple pairwise comparison with a sequentially rejective Bonferroni procedure proposed by Holm [14] to improve its power. We observed that the exhaustive clustering of this paper has a higher probability of correctly classifying multiple objects into clusters than the multiple pairwise tests, and we expect a better performance of the exhaustive clustering in general since it incorporates all available data for clustering.

However, the exhaustive clustering proposed in this paper is time consuming, and its application may not be practical when G is large and/or one wants to cluster un-ordered objects. An ad-hoc approach is to avoid estimating the number of change-points, κ , and/or the locations of the change-points, τ , and this can be done either by using a polynomial regression model or fitting a segmented linear model with the change-points assumed to be at some pre-selected values. As indicated in Kim et al. [6], however, the over-specification or mis-specification of the model may lower the accuracy of the clustering procedure. Similarly, under-performance is expected when clustering is conducted by using a general distance/similarity measure ignoring the underlying model structure, as proposed in the context of mining time-series data. Another approach to replace the exhaustive clustering is to treat the cluster membership information as a missing information and use an EM algorithm with additional assumptions on the distribution as in Qin and Self [4], who proposed a method to cluster linear models and linear mixed models. In our future research, we plan to investigate if these alternative procedures could improve computational efficiency of the procedure proposed in this paper, without losing too much of accuracy and/or limiting the applicability of the proposed method to a class of specific distributions. Another future research problem is on large sample properties of $\hat{M}$ and effects of the data-driven choice of $kmax^{(m)}$ on the consistency of $\hat{\pi}$ presented in Section 3. The use of a fixed $kmax^{(m)}$ can be implemented with the possible maximum number of change-points at given n , although the use of such $kmax^{(m)}$ may take impractically long time, and we also anticipate

the BIC or a modified version of the BIC to provide a consistent estimator of M , and thus the consistency of $\hat{\pi}$. In any case, these asymptotic results imply that when we have enough number of data points around the true change-points, the estimated grouping and estimated regression parameters would be close to the true grouping and true parameter values.

The clustering algorithm will be included in a future version of Joinpoint and interested readers are recommended to visit the Joinpoint website for updated information. (<http://surveillance.cancer.gov/joinpoint>).

Acknowledgments

H.-J. Kim's research was partially supported by NIH Contracts HHSN261201000509P and HHSN261200900681P. J. Kim's research was supported by a research grant from Inha University.

APPENDIX

Lemma: Sketch of Proof

Note that

$$\begin{aligned} G_n(\boldsymbol{\rho}, \boldsymbol{\kappa}, \boldsymbol{\gamma}, \boldsymbol{w}) &= \frac{1}{nG} \sum_{g=1}^G w_g \sum_{i=1}^n \left(\mu_g(\boldsymbol{\rho}, \boldsymbol{\kappa}, \boldsymbol{\gamma}, \mathbf{x}_{g,i}) - \mu_g(\boldsymbol{\rho}^{(0)}, \boldsymbol{\kappa}^{(0)}, \boldsymbol{\gamma}^{(0)}, x_{g,i}) \right)^2 \\ &\quad - \frac{2}{nG} \sum_{g=1}^G w_g \sum_{i=1}^n \epsilon_{g,i} \left(\mu_g(\boldsymbol{\rho}, \boldsymbol{\kappa}, \boldsymbol{\gamma}, \mathbf{x}_{g,i}) - \mu_g(\boldsymbol{\rho}^{(0)}, \boldsymbol{\kappa}^{(0)}, \boldsymbol{\gamma}^{(0)}, x_{g,i}) \right). \end{aligned}$$

We sketch a proof for a simple situation with $M = 2$, $G = 4$, $\rho^{(0)} = 2$ and $\boldsymbol{\kappa}^{(0)} = (1, 1)'$ and the case of nonrandom design data points. Let $\rho = \rho^{(0)} + 1 = 3$. We first note that $E(\epsilon_{g,i}) = 0$ since ϵ_i and $-\epsilon_i$ have the same distribution. Then, it is true that for any $\boldsymbol{\gamma}$, $EG_n(\boldsymbol{\rho}, \boldsymbol{\kappa}^{max}, \boldsymbol{\gamma}, \boldsymbol{w})$ is greater than or equal to

$$\frac{1}{nG} \sum_{g=2,3} w_g \sum_{i=1}^n \left(\mu_g(3, \boldsymbol{\kappa}^{max}, \boldsymbol{\gamma}, \mathbf{x}_{g,i}) - \mu_g(2, \boldsymbol{\kappa}^{(0)}, \boldsymbol{\gamma}^{(0)}, x_{g,i}) \right)^2. \quad (1)$$

Let

$$\mu(\beta_0, \beta_1, \boldsymbol{\delta}, \boldsymbol{\tau}, \kappa, \mathbf{x}) = \beta_0 + \beta_1 x + \delta_1(x - \tau_1)^+ + \cdots + \delta_\kappa(x - \tau_\kappa)^+,$$

where $\boldsymbol{\delta} = (\delta_1, \dots, \delta_\kappa)'$ and $\boldsymbol{\tau} = (\tau_1, \dots, \tau_\kappa)'$. Then

$$\begin{aligned} \mu_2(2, \boldsymbol{\kappa}^{(0)}, \boldsymbol{\gamma}^{(0)}, x) &= \mu(\beta_{2,0}^{(0)}, \beta_1^{(1),(0)}, \delta_1^{(1),(0)}, \tau_1^{(1),(0)}, 1, x), \\ \mu_3(2, \boldsymbol{\kappa}^{(0)}, \boldsymbol{\gamma}^{(0)}, x) &= \mu(\beta_{3,0}^{(0)}, \beta_1^{(2),(0)}, \delta_1^{(2),(0)}, \tau_1^{(2),(0)}, 1, x). \end{aligned}$$

It is natural to assume that for any constant c there exist positive constants c_1 and c_2 such that

$$|\mu_2(2, \kappa^{(0)}, \gamma^{(0)}, x) - \mu_3(2, \kappa^{(0)}, \gamma^{(0)}, x) + c| > c_1$$

for x in a neighborhood with volume c_2 of 0, 1 or $\tau_1^{(j), (0)}$ for $j = 1, 2$.

Since for $g = 2, 3$,

$$\mu_g(3, \mathbf{kmax}, \gamma, \mathbf{x}) = \mu(\beta_{g,0}, \beta_1^{(1)}, \delta^{(1)}, \tau^{(1)}, \mathbf{kmax}^{(1)}, x),$$

it is true that $\mu_2(3, \mathbf{kmax}, \gamma, x) = \mu_3(3, \mathbf{kmax}, \gamma, x) + c_0$ for $c_0 = \beta_{2,0} - \beta_{3,0}$. Therefore, either it is true that

$$|\mu_2(3, \mathbf{kmax}, \gamma, \mathbf{x}) - \mu_2(2, \kappa^{(0)}, \gamma^{(0)}, x)| > \frac{c_1}{2}$$

for x in a neighborhood with volume $c_2/2$ of 0, 1 or $\tau_1^{(j), (0)}$ for $j = 1, 2$, or the following is correct:

$$|\mu_3(3, \mathbf{kmax}, \gamma, \mathbf{x}) - \mu_3(2, \kappa^{(0)}, \gamma^{(0)}, x)| > \frac{c_1}{2}$$

for x in a neighborhood with volume $c_2/2$ of 0, 1 or $\tau_1^{(j), (0)}$ for $j = 1, 2$.

By the Assumption 1', it can be shown that for large n , (1) is greater than or equal to a positive constant which does not depend on γ .

Proof of Theorem 1

Since

$$G_n(\rho, \mathbf{kmax}, \gamma, w) = EG_n(\rho, \mathbf{kmax}, \gamma, w) + (G_n(\rho, \mathbf{kmax}, \gamma, w) - EG_n(\rho, \mathbf{kmax}, \gamma, w)),$$

we have

$$\inf_{|\rho - \rho^{(0)}| > \delta} \inf_{\gamma} G_n(\rho, \mathbf{kmax}, \gamma, w) \geq \inf_{|\rho - \rho^{(0)}| > \delta} \inf_{\gamma} EG_n(\rho, \mathbf{kmax}, \gamma, w) - \sup_{|\rho - \rho^{(0)}| > \delta} \sup_{\gamma} |G_n(\rho, \mathbf{kmax}, \gamma, w) - EG_n(\rho, \mathbf{kmax}, \gamma, w)|.$$

By Lemma, $\inf_{|\rho - \rho^{(0)}| > \delta} \inf_{\gamma} EG_n(\rho, \mathbf{kmax}, \gamma, w) > 0$ for large n . And

$$\sup_{|\rho - \rho^{(0)}| > \delta} \sup_{\gamma} |G_n(\rho, \mathbf{kmax}, \gamma, w) - EG_n(\rho, \mathbf{kmax}, \gamma, w)| = o_p(1),$$

by Lemma 2.1 of Kim and Kim [7]. Thus, with probability approaching one,

$$\begin{aligned}
\inf_{|\rho - \rho^{(0)}| > \delta} \inf_{\gamma} G_n(\rho, kmax, \gamma, w) > 0 &= G_n(\rho^{(0)}, \kappa^{(0)}, \gamma^{(0)}, w) \\
&\geq \inf_{\gamma} G_n(\rho^{(0)}, \kappa^{(0)}, \gamma, w) \\
&\geq \inf_{\gamma} G_n(\rho^{(0)}, kmax, \gamma, w)
\end{aligned}$$

as $n \rightarrow \infty$. This completes the proof.

References

- [1]. Keogh EJ, Kasetty S. On the need for time series data mining benchmarks: a survey and empirical demonstration. *Data Mining and Knowledge Discovery*. 2003; 7(4):349–371.
- [2]. Lei H, Govindaraju V. Generalized regression model for sequence matching and clustering. *Knowledge and Information Systems*. 2007; 12(1):77–94.
- [3]. Ding H, Trajcevski G, Scheuermann P, Wang X, Keogh E. Querying and mining of time series data: experimental comparison of representations and distant measures. *Proceedings of the VLDB Endowment*. 2008; 1(2):1542–1552.
- [4]. Qin LX, Self SG. The clustering of regression models method with applications in gene expression data. *Biometrics*. 2006; 62(2):526–533. [PubMed: 16918917]
- [5]. Kim HJ, Fay MP, Feuer EJ, Midthune DN. Permutation tests for joinpoint regression with applications in cancer rates. *Statistics in Medicine*. 2000; 19(3):335–351. [PubMed: 10649300] 2001; 20(4):655. correction:
- [6]. Kim HJ, Fay MP, Yu B, Barrett MJ, Feuer EJ. Comparability of segmented line regression models. *Biometrics*. 2004; 60(4):1005–1014. [PubMed: 15606421]
- [7]. Kim J, Kim HJ. Asymptotic results in segmented multiple regression. *Journal of Multivariate Analysis*. 2008; 99(9):2016–2038.
- [8]. Kim HJ, Yu B, Feuer EJ. Selecting the number of change-points in segmented line regression. *Statistica Sinica*. 2009; 19(2):597–609. [PubMed: 19738935]
- [9]. Boos DD, Zhang J. Monte Carlo evaluation of resampling-based hypothesis tests. *Journal of the American Statistical Association*. 2000; 95:486–492.
- [10]. American Cancer Society. *Cancer Facts & Figures 2012*. American Cancer Society; Atlanta, GA:
- [11]. Etzioni R, Gulati R, Tsodikov A, Wever EM, Penson DF, Heijnsdijk EA, Katcher J, Draisma G, Feuer EJ, de Koning HJ, Mariotto AB. The prostate cancer conundrum revisited: treatment changes and prostate cancer mortality declines. *Cancer*. 2012; 118(23):5955–5963. [PubMed: 22605665]
- [12]. Feuer EJ, Merrill RM, Hankey BF. Cancer surveillance series: interpreting trends in prostate cancer—part II: Cause of death misclassification and the recent rise and fall in prostate cancer mortality. *Journal of the National Cancer Institute*. 1999; 91(12):1025–1032. [PubMed: 10379965]
- [13]. Davies L, Welch HG. Increasing incidence of thyroid cancer in the United States, 1973–2002. *Journal of the American Medical Association*. 2006; 295(18):2164–2167. [PubMed: 16684987]
- [14]. Holm S. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*. 1979; 6(2):65–70.

Figure 1.
Example Data Sets

Figure 2.
Prostate Cancer Mortality

Figure 3.
Prostate Cancer Mortality with 6 Clusters

Figure 4.
Prostate Cancer Mortality with 2 Clusters at MDWD=1%

Figure 5.
Thyroid Cancer Incidence for White Females

Figure 6.
Thyroid Cancer Incidence with 4 Clusters

Table 1Case I with $G = 6$ and $M = 1$

Case	σ_g	$\mu_g(x)$	JPerm-CBIC		JBIC-CBIC	
			P_M	P_C	P_M	P_C
I-1	$\sigma_g = 0.10$ ($g = 1, \dots, 6$)	$\mu_1(x) = 0.3 - 0.002x$	0.9693	0.9693	0.9515	0.9515
		$\mu_2(x) = 0.6 - 0.002x$				
		$\mu_3(x) = 1.0 - 0.002x$				
		$\mu_4(x) = 1.3 - 0.002x$				
		$\mu_5(x) = 1.6 - 0.002x$				
		$\mu_6(x) = 2.0 - 0.002x$				
I-2	$\sigma_1 = 0.30$ $\sigma_2 = 0.30$ $\sigma_3 = 0.20$ $\sigma_4 = 0.20$ $\sigma_5 = 0.10$ $\sigma_6 = 0.10$	same as μ_g in Case I-1	0.9661	0.9661	0.9435	0.9435
I-3	$\sigma_g = 0.50$ ($g = 1, \dots, 6$)	same as μ_g in Case I-1	0.9693	0.9693	0.9515	0.9515

Table 2Case II with $G = 6$ and $M = 2$

Case	ρ_1	σ_g	$\mu_g(x)$	JPerm-CBIC		JBIC-CBIC		
				P_M	P_C	P_M	P_C	
II-1	3	$\sigma_g = 0.10$ ($g = 1, \dots, 6$)	$\mu_1(x) = 1.5 + 0.004x$ $\mu_2(x) = 2.0 + 0.004x$ $\mu_3(x) = 2.5 + 0.004x$ $\mu_4(x) = 1.5 + 0.15x - 0.13(x - 8)^+$ $\mu_5(x) = 2.0 + 0.15x - 0.13(x - 8)^+$ $\mu_6(x) = 2.5 + 0.15x - 0.13(x - 8)^+$	0.9871	0.9871	0.9725	0.9725	593.36
II-2	3	$\sigma_g = \mathbf{0.50}$ ($g = 1, \dots, 6$)	same as μ_g in Case II-1	0.8691	0.8207	0.8756	0.8239	23.73
II-3	3	$\sigma_g = \mathbf{0.30}$ ($g = 1, 2, 3$) $\sigma_g = \mathbf{0.80}$ ($g = 4, 5, 6$)	same as μ_g in Case II-1	0.6801	0.5961	0.6931	0.6058	16.26
II-4	3	$\sigma_g = \mathbf{1.0}$ ($g = 1, \dots, 6$)	same as μ_g in Case II-1	0.2068	0.1470	0.2520	0.1712	5.93
II-5	3	$\sigma_g = \mathbf{0.50}$ ($g = 1, \dots, 6$)	$\mu_1(x) = 1.5 + 0.004x$ $\mu_2(x) = 2.0 + 0.004x$ $\mu_3(x) = 2.5 + 0.004x$ $\mu_4(x) = 1.5 + 0.15x - 0.13(\mathbf{x} - \mathbf{10})^+$ $\mu_5(x) = 2.0 + 0.15x - 0.13(\mathbf{x} - \mathbf{10})^+$ $\mu_6(x) = 2.5 + 0.15x - 0.13(\mathbf{x} - \mathbf{10})^+$	0.9758	0.9548	0.9467	0.9289	38.80
II-6	3	$\sigma_g = \mathbf{0.50}$ ($g = 1, \dots, 6$)	$\mu_1(x) = 1.5 + 0.004x$ $\mu_2(x) = 2.0 + 0.004x$ $\mu_3(x) = 2.5 + 0.004x$ $\mu_4(x) = 1.5 + 0.15x - 0.13(\mathbf{x} - \mathbf{12})^+$ $\mu_5(x) = 2.0 + 0.15x - 0.13(\mathbf{x} - \mathbf{12})^+$ $\mu_6(x) = 2.5 + 0.15x - 0.13(\mathbf{x} - \mathbf{12})^+$	0.9855	0.9822	0.9693	0.9677	58.43
II-7	2	$\sigma_i = \mathbf{0.50}$ ($i = 1, \dots, 6$)	$\mu_1(x) = 1.5 + 0.004x$ $\mu_2(x) = 2.0 + 0.004x$ $\mu_3(x) = 1.5 + 0.15x - 0.13(x - 8)^+$ $\mu_4(x) = 2.0 + 0.15x - 0.13(x - 8)^+$ $\mu_5(x) = 2.5 + 0.15x - 0.13(x - 8)^+$	0.8643	0.8029	0.8514	0.7835	21.10

Case	ρ_1	σ_g	$\mu_g(x)$	JPerm-CBIC		JBIC-CBIC		
				P_M	P_C	P_M	P_C	
			$\mu_6(x) = 3.0 + 0.15x - 0.13(x - 8)^+$					
II-8	3	$\sigma_g = 0.10$ ($g = 1, \dots, 6$)	$\mu_1(x) = 1.5 + 0.004x$ $\mu_2(x) = 2.0 + 0.004x$ $\mu_3(x) = 2.5 + 0.004x$ $\mu_4(x) = 1.5 + \mathbf{0.004}x - 0.13(x - 8)^+$ $\mu_5(x) = 2.0 + \mathbf{0.004}x - 0.13(x - 8)^+$ $\mu_6(x) = 2.5 + \mathbf{0.004}x - 0.13(x - 8)^+$	0.9871	0.9871	0.9725	0.9725	4211.56
II-9	3	$\sigma_g = \mathbf{0.50}$ ($g = 1, \dots, 6$)	$\mu_1(x) = 1.5 + 0.004x$ $\mu_2(x) = 2.0 + 0.004x$ $\mu_3(x) = 2.5 + 0.004x$ $\mu_4(x) = 1.5 + \mathbf{0.004}x - 0.13(x - 8)^+$ $\mu_5(x) = 2.0 + \mathbf{0.004}x - 0.13(x - 8)^+$ $\mu_6(x) = 2.5 + \mathbf{0.004}x - 0.13(x - 8)^+$	0.9871	0.9871	0.9693	0.9693	168.46
II-10	3	$\sigma_g = \mathbf{0.50}$ ($g = 1, \dots, 6$)	$\mu_1(x) = 1.5 + 0.004x$ $\mu_2(x) = 2.0 + 0.004x$ $\mu_3(x) = 2.5 + 0.004x$ $\mu_4(x) = 1.5 + \mathbf{0.004}x - 0.13(\mathbf{x} - \mathbf{10})^+$ $\mu_5(x) = 2.0 + \mathbf{0.004}x - 0.13(\mathbf{x} - \mathbf{10})^+$ $\mu_6(x) = 2.5 + \mathbf{0.004}x - 0.13(\mathbf{x} - \mathbf{10})^+$	0.9855	0.9855	0.9677	0.9677	141.96

Table 3

Case II with $G = 6$ and $M = 2$

Case	ρ_1	σ_g	$\mu_g(x)$	JPerm-CBIC		JBIC-CBIC		
				P_M	P_C	P_M	P_C	
II-2	3	$\sigma_g = 0.50$ ($g = 1, \dots, 6$)	$\mu_1(x) = 1.5 + 0.004x$ $\mu_2(x) = 2.0 + 0.004x$ $\mu_3(x) = 2.5 + 0.004x$ $\mu_4(x) = 1.5 + 0.15x - 0.13(x - 8)^+$ $\mu_5(x) = 2.0 + 0.15x - 0.13(x - 8)^+$ $\mu_6(x) = 2.5 + 0.15x - 0.13(x - 8)^+$	0.8691	0.8207	0.8756	0.8239	23.73
II-2-1	3	$\sigma_g = 0.50$ ($g = 1, \dots, 6$)	$\mu_1(x) = 1.5 + 0.15x - \mathbf{0.17}(x - 6)^+$ $\mu_2(x) = 2.0 + 0.15x - \mathbf{0.17}(x - 6)^+$ $\mu_3(x) = 2.5 + 0.15x - \mathbf{0.17}(x - 6)^+$ $\mu_4(x) = 1.5 + 0.15x - 0.13(x - 8)^+$ $\mu_5(x) = 2.0 + 0.15x - 0.13(x - 8)^+$ $\mu_6(x) = 2.5 + 0.15x - 0.13(x - 8)^+$	0.8514	0.7884	0.8465	0.8078	28.25
II-2-2	3	$\sigma_g = 0.50$ ($g = 1, \dots, 6$)	$\mu_1(x) = 1.5 + 0.15x - \mathbf{0.17}(x - 10)^+$ $\mu_2(x) = 2.0 + 0.15x - \mathbf{0.17}(x - 10)^+$ $\mu_3(x) = 2.5 + 0.15x - \mathbf{0.17}(x - 10)^+$ $\mu_4(x) = 1.5 + 0.15x - 0.13(x - 8)^+$ $\mu_5(x) = 2.0 + 0.15x - 0.13(x - 8)^+$ $\mu_6(x) = 2.5 + 0.15x - 0.13(x - 8)^+$	0.1422	0.0969	0.1163	0.0921	8.38
II-2-3	3	$\sigma_g = 0.50$ ($g = 1, \dots, 6$)	$\mu_1(x) = 1.5 + 0.15x - \mathbf{0.18}(x - 6)^+$ $\mu_2(x) = 2.0 + 0.15x - \mathbf{0.18}(x - 6)^+$ $\mu_3(x) = 2.5 + 0.15x - \mathbf{0.18}(x - 6)^+$ $\mu_4(x) = 1.5 + 0.15x - 0.13(x - 8)^+$ $\mu_5(x) = 2.0 + 0.15x - 0.13(x - 8)^+$ $\mu_6(x) = 2.5 + 0.15x - 0.13(x - 8)^+$	0.9354	0.9095	0.9402	0.9257	40.53
II-2-4	3	$\sigma_g = 0.50$ ($g = 1, \dots, 6$)	$\mu_1(x) = 1.5 + 0.15x - \mathbf{0.18}(x - 10)^+$ $\mu_2(x) = 2.0 + 0.15x - \mathbf{0.18}(x - 10)^+$ $\mu_3(x) = 2.5 + 0.15x - \mathbf{0.18}(x - 10)^+$ $\mu_4(x) = 1.5 + 0.15x - 0.13(x - 8)^+$ $\mu_5(x) = 2.0 + 0.15x - 0.13(x - 8)^+$ $\mu_6(x) = 2.5 + 0.15x - 0.13(x - 8)^+$	0.3764	0.3021	0.3796	0.3263	14.09
			$\mu_1(x) = 1.5 + 0.15x - \mathbf{0.17}(x - 8)^+$					

Case	ρ_1	σ_g	$\mu_g(x)$	JPerm-CBIC		JBIC-CBIC		
				P_M	P_C	P_M	P_C	
II-2-5	3	$\sigma_g = 0.50$ ($g = 1, \dots, 6$)	$\mu_2(x) = 2.0 + 0.15x - \mathbf{0.17}(x - 8)^+$ $\mu_3(x) = 2.5 + 0.15x - \mathbf{0.17}(x - 8)^+$ $\mu_4(x) = 1.5 + 0.15x - 0.13(x - 8)^+$ $\mu_5(x) = 2.0 + 0.15x - 0.13(x - 8)^+$ $\mu_6(x) = 2.5 + 0.15x - 0.13(x - 8)^+$	0.4766	0.4087	0.4782	0.4233	15.95
II-2-6	3	$\sigma_g = 0.50$ ($g = 1, \dots, 6$)	$\mu_1(x) = 1.5 + 0.15x - \mathbf{0.18}(x - 8)^+$ $\mu_2(x) = 2.0 + 0.15x - \mathbf{0.18}(x - 8)^+$ $\mu_3(x) = 2.5 + 0.15x - \mathbf{0.18}(x - 8)^+$ $\mu_4(x) = 1.5 + 0.15x - 0.13(x - 8)^+$ $\mu_5(x) = 2.0 + 0.15x - 0.13(x - 8)^+$ $\mu_6(x) = 2.5 + 0.15x - 0.13(x - 8)^+$	0.7431	0.6931	0.7625	0.7173	24.92
II-2-7	3	$\sigma_g = 0.50$ ($g = 1, \dots, 6$)	$\mu_1(x) = 1.5 + 0.15x - 0.13(x - \mathbf{6})^+$ $\mu_2(x) = 2.0 + 0.15x - 0.13(x - \mathbf{6})^+$ $\mu_3(x) = 2.5 + 0.15x - 0.13(x - \mathbf{6})^+$ $\mu_4(x) = 1.5 + 0.15x - 0.13(x - 8)^+$ $\mu_5(x) = 2.0 + 0.15x - 0.13(x - 8)^+$ $\mu_6(x) = 2.5 + 0.15x - 0.13(x - 8)^+$	0.0565	0.0372	0.0614	0.0339	1.96
II-2-8	3	$\sigma_g = 0.50$ ($g = 1, \dots, 6$)	$\mu_1(x) = 1.5 + 0.15x - 0.13(x - \mathbf{15})^+$ $\mu_2(x) = 2.0 + 0.15x - 0.13(x - \mathbf{15})^+$ $\mu_3(x) = 2.5 + 0.15x - 0.13(x - \mathbf{15})^+$ $\mu_4(x) = 1.5 + 0.15x - 0.13(x - 8)^+$ $\mu_5(x) = 2.0 + 0.15x - 0.13(x - 8)^+$ $\mu_6(x) = 2.5 + 0.15x - 0.13(x - 8)^+$	0.8401	0.7835	0.8481	0.7964	28.94

Table 4Case III with $G = 10$ and $M = 3$

Case	ρ	σ_g	$\mu_g(x)$	JPerm-CBIC		JBIC-CBIC	
				P_M	P_C	P_M	P_C
III-1	$\rho_1 = 3$ $\rho_2 = 6$	$\sigma_g = 0.05$ $(g = 1, \dots, 10)$	$\mu_1(x) = 1 + 0.03x + 0.12(x - 10)^+ - 0.10(x - 20)^+$	0.8546	0.8401	0.7932	0.7819
			$\mu_2(x) = 1.5 + 0.03x + 0.12(x - 10)^+ - 0.10(x - 20)^+$				
			$\mu_3(x) = 2 + 0.03x + 0.12(x - 10)^+ - 0.10(x - 20)^+$				
			$\mu_4(x) = 3 + 0.03x + 0.10(x - 10)^+ + 0.05(x - 14)^+ - 0.14(x - 19)^+$				
			$\mu_5(x) = 3.5 + 0.03x + 0.10(x - 10)^+ + 0.05(x - 14)^+ - 0.14(x - 19)^+$				
			$\mu_6(x) = 4 + 0.03x + 0.10(x - 10)^+ + 0.05(x - 14)^+ - 0.14(x - 19)^+$				
			$\mu_7(x) = 4.5 + 0.02x + 0.10(x - 10)^+ - 0.02(x - 16)^+ - 0.05(x - 20)^+$				
			$\mu_8(x) = 5 + 0.02x + 0.10(x - 10)^+ - 0.02(x - 16)^+ - 0.05(x - 20)^+$				
			$\mu_9(x) = 5.5 + 0.02x + 0.10(x - 10)^+ - 0.02(x - 16)^+ - 0.05(x - 20)^+$				
			$\mu_{10}(x) = 6 + 0.02x + 0.10(x - 10)^+ - 0.02(x - 16)^+ - 0.05(x - 20)^+$				

Table 5

Clustering results with different values of MDWD for U.S. Prostate Cancer Mortality from 1975 - 2009.

MDWD (ϕ)	0	0.5%	1%	1.5%	2%
$\hat{M}^\varphi$	6	6	2	2	1
Age groups	0 – 54, 55 – 64, 65 – 74, 75 – 79, 80 – 84, 85 ⁺	0 – 54, 55 – 64, 65 – 74, 75 – 79, 80 – 84, 85 ⁺	0 – 84, 85 ⁺	0 – 84, 85 ⁺	0 – 85 ⁺

Table 6

Clustering results with different values of MDWD for U.S. female thyroid cancer incidence from 1975 - 2009.

MDWD (ϕ)	0	0.5%	1%	1.5%	2%
$\hat{M}^\varphi$	4	4	4	4	4
Age groups	0 – 24, 25 – 29, 30 – 79, 80 ⁺	0 – 24, 25 – 29, 30 – 79, 80 ⁺	0 – 24, 25 – 29, 30 – 79, 80 ⁺	0 – 24, 25 – 29, 30 – 79, 80 ⁺	0 – 24, 25 – 29, 30 – 79, 80 ⁺