Appl. Statist. (2005)
54, *Part* 5, *pp.* 919–939

Bayesian model selection for join point regression with application to age-adjusted cancer rates

Ram C. Tiwari, Kathleen A. Cronin, William Davis and Eric J. Feuer,

National Cancer Institute, Rockville, USA

Binbing Yu

Information Management Services, Silver Spring, USA

and Siddhartha Chib

Washington University, St Louis, USA

[Received June 2004. Final revision January 2005]

Summary. The method of Bayesian model selection for join point regression models is developed. Given a set of $K+1$ join point models $M_0, M_1, \ldots, M_K$ with $0, 1, \ldots, K$ join points respectively, the posterior distributions of the parameters and competing models M_k are computed by Markov chain Monte Carlo simulations. The Bayes information criterion BIC is used to select the model M_k with the smallest value of BIC as the best model. Another approach based on the Bayes factor selects the model M_k with the largest posterior probability as the best model when the prior distribution of M_k is discrete uniform. Both methods are applied to analyse the observed US cancer incidence rates for some selected cancer sites. The graphs of the join point models fitted to the data are produced by using the methods proposed and compared with the method of Kim and co-workers that is based on a series of permutation tests. The analyses show that the Bayes factor is sensitive to the prior specification of the variance σ^2, and that the model which is selected by BIC fits the data as well as the model that is selected by the permutation test and has the advantage of producing the posterior distribution for the join points. The Bayesian join point model and model selection method that are presented here will be integrated in the National Cancer Institute's join point software (http://www.srab.cancer.gov/joinpoint/) and will be available to the public.

Keywords: Annual percentage change; Bayes factor; Bayes information criterion; Markov chain Monte Carlo methods; Permutation test

1. Introduction

A question that is of particular interest when analysing cancer incidence and mortality rates is whether or not there has been a change in the trend over time and, if there has been a change, when it occurred. Questions of this type play an important role in measuring progress against cancer and in assessing the effect of population intervention on the outcome of disease. For example, a change in trend for lung cancer incidence may reflect the population effect of anti-tobacco programmes or changes in the trend for cancer mortality may be the result of new screening modalities. In addition to helping to explain what factors influence trends, identifying when changes occur also plays a role in defining the current trend. In past cancer statistics publications (Ries *et al.*, 2002) the current trend has been defined by fitting a line through a

Address for correspondence: Binbing Yu, Information Management Services, Inc., Suite 200, 12501 Prosperity Drive, Silver Spring, MD 20904, USA.
E-mail: yub@imsweb.com

© 2005 Royal Statistical Society 0035–9254/05/54919

Downloaded from https://academic.oup.com/jrssc/article/54/5/919/7113011 by guest on 19 September 2024

prespecified number of years usually at the end of the observed data. Although this may be useful for easily summarizing the most recent trend over a large number of cancer sites for a fixed period, it may not properly characterize the trend. For this reason, we have found that a log-linear model with random changepoints has been quite useful in modelling and interpreting cancer trends. Since the models define a changepoint as a change in slope, but do not allow a jump in the level at a change, we refer to these types of models as join point models. The current method for fitting join point models in the annual report to the nation on the status of cancer (Edwards *et al.*, 2002) and the National Cancer Institute's cancer statistics review (Ries *et al.*, 2002) is described by Kim *et al.* (2000) and briefly reviewed below. The purpose of this paper is to develop Bayesian model selection methods by using criteria, namely the Bayes factor (BF) and Bayes information criterion BIC, both to fit a join point regression model to age-adjusted cancer rates and to provide a measure of uncertainty related to the number of join points in a data series. The performances of these methods are compared with the permutation-test-based (PTB) method for fitting a join point model that was developed by Kim *et al.* (2000).

1.1. Join point model

Let d_{ij} and n_{ij} denote respectively the cancer counts and population size at time x_i, and for age group j, $i = 1, \ldots, n,\ j = 1, \ldots, J$. The age-adjusted rates are

$$r_i = \sum_{j=1}^{J} \frac{c_j d_{ij}}{n_{ij}}, \qquad i = 1, \ldots, n,$$

where c_js are the known standards and $\Sigma_{j=1}^{J} c_j = 1$. Let $y_i = \log(r_i)$ denote the logarithm of the observed age-adjusted rates at time x_i, $i = 1, \ldots, n$. Under the assumption that the d_{ij}s are independent Poisson random variables with means $n_{ij}\lambda_{ij}$, an estimate of $\mathrm{var}(y_i)$ was given by Kim *et al.* (2000):

$$w_i = \mathrm{var}(y_i) = \sum_{j=1}^{J} \frac{c_j^2 d_{ij}}{n_{ij}^2} \Big/ \left(\sum_{j=1}^{J} \frac{c_j d_{ij}}{n_{ij}} \right)^2. \tag{1}$$

A join point model M_k, with k join points for fitting the observed data $\{(x_i, y_i) : x_1 < \ldots < x_n; i = 1, \ldots, n\}$, was given by Lerman (1980) and Kim *et al.* (2000):

$$y_i = \beta_0 + \beta_1 x_i + \sum_{r=1}^{k} \delta_r s_r(x_i) + \varepsilon_i, \tag{2}$$

where $s_r(x) = (x - \tau_r)^+$, and $a^+ = a$ if $a > 0$, and $a^+ = 0$, otherwise, $\boldsymbol{\beta}_k^{\mathrm{T}} = (\beta_0, \beta_1, \delta_1, \ldots, \delta_k)$ are the regression parameters and $\boldsymbol{\tau}_k^{\mathrm{T}} = (\tau_1, \ldots, \tau_k)$ are the join points, and $\varepsilon_1, \ldots, \varepsilon_n$ are zero-mean random errors. Here, for any vector or matrix v, v^{T} denotes the transpose of v. The annual percentage change APC of the age-adjusted rates between τ_k and τ_{k+1} (i.e. for the $(k+1)$th segment) is given by

$$\mathrm{APC}_k = 100\{\exp(\beta_1 + \delta_1 + \ldots + \delta_k) - 1\}.$$

Model (2) is also known as a spline model with $s_r(x)$ as the rth basis function evaluated at x, τ_r as the corresponding knot and δ_r as the corresponding coefficient. For $k = 0$, the join point model (2), corresponding to a zero join point, is the simple linear regression model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$. A more general form of model (2), which allows a pth- ($p \geqslant 1$) degree polynomial on each interval between two consecutive join points, which has $p - 1$ continuous derivatives everywhere, can be also considered. However, for analysing the age-adjusted cancer incidence

Downloaded from https://academic.oup.com/jrsssc/article/54/5/919/7113011 by guest on 19 September 2024

or mortality rates, where we are interested in APC, the single-join-point regression model (2) is more appropriate.

Several approaches to join point regression models have been considered in the literature; for example, for a nonparametric formulation see Hinkley (1971), Pettitt (1980) and Kim *et al.* (2000), for a likelihood formulation see Hinkley (1970) and for a Bayesian formulation see Smith (1975, 1980), Carlin *et al.* (1992), Stephens (1994), Green (1995) and Denison *et al.* (1998), among others.

1.2. Review of permutation-test-based approach of Kim et al. (2000)

Kim *et al.* (2000) developed a nonparametric PTB approach to fit the join point model (2) to the data $\mathbf{y}_n = \{(x_i, y_i) : x_1 < \ldots < x_n; i = 1, \ldots, n\}$, first assuming that $\mathrm{var}(\varepsilon_i) = \sigma^2$, for all i, and then extending the methodology to handle the case when $\mathrm{var}(\varepsilon_i) = w_i$ are specified constants. The model selection based on a series of permutation tests is briefly described as follows. First, a maximum number k_1 and a minimum number k_0 of possible join points are selected. Usually $k_0 = 0$ and $k_1 = 3$ or $k_1 = 4$, depending on the length and complexity of the data series. It begins with testing the null hypothesis of k_0 join points against the alternative of k_1 join points, where $0 \leqslant k_0 < k_1$. If the null hypothesis is rejected at level α_1 and $k_1 - k_0 \geqslant 2$, then we test H_0: there are $k_0 + 1$ join points against H_0: there are k_1 join points. If the null hypothesis of k_0 join points is not rejected and $k_1 - k_0 \geqslant 2$, then we test H_0: there are k_0 join points against H_1: there are $k_1 - 1$ join points. The testing procedure continues until testing the null hypothesis of k join points against the alternative of $k + 1$ join points for some $k_0 \leqslant k < k_1$. The estimated number of join points is $k + 1$ if the final null hypothesis is rejected and k otherwise. The level of each test is adjusted to $\alpha_1 = \alpha/(k_1 - k_0)$ by using the Bonferroni correction to reach the overall significance level of α. The test in each step is carried out by permutation of the residuals. An F-type statistic $F(\mathbf{y}_n)$ is calculated from the original data $\mathbf{y}_n$. The residuals $\hat{\varepsilon}_n$ and predicted value $\hat{\mathbf{y}}_n$ are obtained by fitting the k_0-join-point model under the null hypothesis. The permutations of the residuals, $\hat{\varepsilon}_{p(n)}$, are added back to $\hat{\mathbf{y}}_n$ to create permuted samples $\mathbf{y}_{p(n)}$ and the test statistic $F(\mathbf{y}_{p(n)})$ is calculated for the permuted sample. The p-value of the test is the proportion of times that $F(\mathbf{y}_{p(n)}) > F(\mathbf{y}_n)$ over a large number of permutations. The National Cancer Institute has developed the join point regression software for the analysis of trends by using join point models (see the Web site `http://srab.cancer.gov/joinpoint/index.html`).

1.3. Method proposed

We first assume that the errors in model (2) are independent and identically distributed (IID) with $\varepsilon_i \sim N(0, \sigma^2)$, a normal distribution with mean 0 and variance σ^2. Conditional on a fixed maximum number of the join points, K, we develop a Bayesian model selection procedure for comparing the $K + 1$ join point models, $\{M_0, M_1, \ldots, M_K\}$. For each $k = 0, 1, \ldots, K$, model M_k is characterized by the parameter vector $\boldsymbol{\theta}_k = (\boldsymbol{\beta}_k^{\mathrm{T}}, \boldsymbol{\tau}_k^{\mathrm{T}}, \sigma^2)^{\mathrm{T}}$, where $\boldsymbol{\beta}_k \in R^{(k+2)}$ with R denoting the real line, $\sigma^2 > 0$ and the join points $\tau_1, \ldots, \tau_k$ take values in $\{x_1, \ldots, x_n\}$.

Let $\pi(\boldsymbol{\theta}_k | M_k)$ be the prior of $\boldsymbol{\theta}_k$ under model M_k and $L(\boldsymbol{\theta}_k | \mathbf{y}_n)$ be the likelihood function given observed data $\mathbf{y}_n$. The posterior distribution of $\boldsymbol{\theta}_k$ is given by

$$\pi(\boldsymbol{\theta}_k | M_k, \mathbf{y}_n) \propto \pi(\boldsymbol{\theta}_k | M_k)\, L(\boldsymbol{\theta}_k | \mathbf{y}_n).$$

One model selection approach is based on Schwarz's Bayes information criterion BIC (Schwarz, 1978; Pauler, 1998; Kass and Wasserman, 1995) for model M_k, i.e.

$$\mathrm{BIC}(M_k) = \frac{-2\, \log\{L(\hat{\boldsymbol{\theta}}_k | \mathbf{y}_n)\}}{n} + \frac{p}{n} \log(n), \tag{3}$$

Downloaded from https://academic.oup.com/jrsssc/article/54/5/919/7113011 by guest on 19 September 2024

where $\hat{\theta}_k$ is usually the maximum likelihood estimate (MLE) of θ_k and p is the number of parameters. The BIC approach selects model M_k with the minimum value of BIC as the best model. Note that $\log\{L(\theta_k|\mathbf{y}_n)\} \propto \log\{\pi(\theta_k|M_k, \mathbf{y}_n)\} - \log\{\pi(\theta_k|M_k)\}$ and, if $\pi(\theta_k|M_k)$ is a unimodal function, then, for large n, maximizing $\log\{L(\theta_k|\mathbf{y}_n)\}$ is equivalent to maximizing $\log\{\pi(\theta_k|M_k, \mathbf{y}_n)\}$. Thus, an approximation to BIC(M_k) can be obtained by replacing $\hat{\theta}_k$ in equation (3) by the mode of the posterior distribution (Tan *et al.*, 2003). However, as mentioned in Section 2.2, the BIC computation works for any choice of θ_k, such as the mean, median or mode, as long as it is a high density point. For example, Spiegelhalter *et al.* (2002) used the posterior mean to compute the deviance information criterion DIC, which is related to BIC. As a remark, we mention that the Bayesian version of BIC based on the posterior mode rather than the MLE is not necessary if our goal is just to select the best join point model. However, since we are also interested in the posterior distribution of the join point locations, we use the 'Bayesian' version of BIC.

The second approach of model selection compares the BF and selects the model with the largest value. The BF (Berger, 1985) for comparing a pair of models M_k and M_l is defined by

$$B_{kl} = \frac{P(M_k|\mathbf{y}_n)/P(M_l|\mathbf{y}_n)}{P(M_k)/P(M_l)}. \tag{4}$$

The term $P(M_k|\mathbf{y}_n)/P(M_l|\mathbf{y}_n)$, in the numerator of equation (4), is the posterior odds ratio, and the term in the denominator, $P(M_k)/P(M_l)$, is the prior odds ratio. Under the uniform prior distribution over the set $\{M_0, M_1, \ldots, M_K\}$, i.e. $P(M_k) = 1/(K+1)$, $k = 0, 1, \ldots, K$, the posterior probability of M_k, given the data $\mathbf{y}_n$, is given by

$$P(M_k|\mathbf{y}_n) = m(\mathbf{y}_n|M_k) \Big/ \sum_{r=0}^{K} m(\mathbf{y}_n|M_r), \tag{5}$$

where $m(\mathbf{y}_n|M_k)$ is the marginal likelihood function of model M_k:

$$m(\mathbf{y}_n|M_k) = \int f(\mathbf{y}_n|M_k, \theta_k)\, \pi(\theta_k|M_k)\, \mathrm{d}\theta_k, \tag{6}$$

with $f(\mathbf{y}_n|M_k, \theta_k)$ and $\pi(\theta_k|M_k)$ denoting respectively the likelihood function and the prior. Hence, under the uniform prior for the M_ks, the approach based on the BF is the same as comparing the models by using posterior probabilities $P(M_k|y_n)$ of k join points ($k = 0, 1, \ldots, K$). See Kass and Raftery (1995) for more discussion on the BF. Note that the second approach is highly prior dependent, whereas the first approach depends on $L(\hat{\theta}_k|y_n)$ and its delta approximation.

Next, in model (2) we relax the assumption of IID errors by assuming that ε_i, $i = 1, \ldots, n$, are independent normal $N(0, \omega_i\sigma^2)$ with known weights ω_i. Assume that the spacings $\Delta = x_{i+1} - x_i$ between two data points are constant. We further relax the assumption that the join points occur at the data points x_i, $i = 1, \ldots, n$. We augment the data $\{x_1, \ldots, x_n\}$ by inserting $m-1$ equally spaced points

$$x_{i,u} = x_i + u\delta, \qquad u = 1, \ldots, m-1,$$

in the interval (x_i, x_{i+1}), where $\delta = \Delta/m$. The cancer trend data will be analysed under three model assumptions: IID errors, non-IID errors and augmented data.

1.4. Outline of paper

The rest of this paper is organized as follows. The prior–posterior analysis of the join point regression model M_k in equation (2) is carried out in Section 2. We assume independent normal priors for the regression parameters, and an inverted gamma prior for the error variances. For

Downloaded from https://academic.oup.com/jrsssc/article/54/5/919/7113011 by guest on 19 September 2024

the join points, we assume that the join points are discrete random variables taking values in $\{x_1, \ldots, x_n\}$. The Bayesian model selection methods (BIC and the BF) are developed in Section 3. The estimation of the marginal likelihoods in equation (6) is based on the Markov chain Monte Carlo (MCMC) method that was described in Chib (1995). The results of Section 2 are used, in Section 3, to analyse the observed age-adjusted cancer incidence rates for the period from 1973 to 1999 for the USA and to identify the changes in the trend for colorectal cancer, prostate cancer, breast cancer in white women and breast cancer in black women. The data were collected by the National Cancer Institute's 'Surveillance, epidemiology and end result' (SEER) programme (`http://seer.cancer.gov`) (1999). The elucidation of prior distributions of the model parameters is discussed. In particular, the prior for σ^2 is assessed through the weights $\{w_i\}$ as given in equation (1). For comparison, the analyses using the permutation test for the selected cancer sites are also carried out in this section. Sensitivity of the prior for σ^2 and some extensions of model (2) are discussed in Section 4. Finally, the conclusions are stated in Section 5.

The data that are analysed in the paper can be obtained from

`http://www.blackwellpublishing.com/rss`

2. Bayesian model selection

2.1. Prior and posterior distributions

Under model M_k, let the joint prior $\pi(\boldsymbol{\theta}_k|M_k)$ of $\boldsymbol{\theta}_k = (\boldsymbol{\beta}_k^{\mathrm{T}}, \boldsymbol{\tau}_k^{\mathrm{T}}, \sigma^2)^{\mathrm{T}}$ be specified as follows. Let $\boldsymbol{\beta}_k$, $\boldsymbol{\tau}_k$ and σ^2 be independent and be distributed as

$$\left.\begin{aligned} \boldsymbol{\beta}_k|M_k &\sim N_{k+2}(\boldsymbol{\beta}_{0k}, B_{0k}), \\ \sigma^2|M_k &\sim \mathrm{IG}\left(\frac{\nu_0}{2}, \frac{\delta_0}{2}\right), \\ \pi(\tau_1|M_k) &\propto \frac{1}{n-(2l+k-1)}, \qquad \tau_1 \in \{x_{l+1}, \ldots, x_{n-l-k+1}\}, \quad l \geqslant 0, \\ \pi(\tau_u|M_k, \tau_{u-1} = l', \tau_{u+1} = l'') &\propto \frac{1}{l''-l'-1}, \qquad \tau_u \in \{l'+1, \ldots, l''-1\}, \quad u = 2, \ldots, k, \end{aligned}\right\} \tag{7}$$

where $N_m(\boldsymbol{\mu}, \Sigma)$ denotes an m-dimensional normal distribution and $\mathrm{IG}(a/2, b/2)$ is the inverted gamma distribution with mean and variance respectively given by $b/(a-2)$ and $2b^2/\{(a-4)(a-2)^2\}$. The distribution of join point τ_1 is a discrete uniform distribution on $\{x_{l+1}, \ldots, x_{n-l-k+1}\}$, leaving out $l\,(\geqslant 0)$ values of x at both ends, and the conditional distribution of τ_u given $\{\tau_{u-1} = l', \tau_{u+1} = l''\}$ is a discrete uniform distribution on $\{l'+1, \ldots, l''-1\}$, $u = 2, \ldots, k$, with $\tau_{k+1} \equiv x_{n-1}$. The likelihood function is

$$\begin{aligned} L(\boldsymbol{\theta}_k|\mathbf{y}_n) &\propto \frac{1}{(\sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\{y_i - \beta_0 - \beta_1 x_i - \delta_1 s_1(x_i) - \ldots - \delta_k s_k(x_i)\}^2\right] \\ &\propto \frac{1}{(\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}\|\mathbf{y}_n - X(\tau_1, \ldots, \tau_k)\boldsymbol{\beta}_k\|^2\right\}, \end{aligned} \tag{8}$$

where, for any $(n \times 1)$-vector $\mathbf{a}_n = (a_1, \ldots, a_n)^{\mathrm{T}}$, $\|\mathbf{a}_n\| = \Sigma_{i=1}^n a_i^2$ and $X(\tau_1, \ldots, \tau_k)$ is the $n \times (k+2)$-design matrix with the ith row given by $\mathbf{x}(\tau_1, \ldots, \tau_k)(i)^{\mathrm{T}} = (1, x_i, s_1(x_i), \ldots, s_k(x_i))$.

The posterior distribution of $\boldsymbol{\theta}_k$ is given by

$$\pi(\boldsymbol{\theta}_k|M_k, \mathbf{y}_n) \propto \pi(\boldsymbol{\beta}_k|M_k)\,\pi(\tau_k|M_k)\,\pi(\sigma^2|M_k)\frac{1}{(\sigma^2)^{n/2}}\exp\left\{-\frac{1}{2\sigma^2}\|\mathbf{y}_n - X(\tau_1, \ldots, \tau_k)\boldsymbol{\beta}_k\|^2\right\},$$

Downloaded from https://academic.oup.com/jrsssc/article/54/5/919/7113011 by guest on 19 September 2024

from which the conditional posterior densities of β_k, τ_k and σ^2, under model M_k, are

$$\begin{aligned}\boldsymbol{\beta}_k|\mathbf{y}_n, \sigma^2, \boldsymbol{\tau}_k &\sim N_{k+2}(\hat{\boldsymbol{\beta}}_k, B_k),\\ \sigma^2|\mathbf{y}_n, \boldsymbol{\beta}_k, \boldsymbol{\tau}_k &\sim \mathrm{IG}\left(\frac{\nu}{2}, \frac{\delta}{2}\right),\end{aligned} \tag{9}$$

where

$$\begin{aligned}B_k &= \left(B_{0k}^{-1} + \frac{1}{\sigma^2} X(\tau_1, \ldots, \tau_k)^{\mathrm{T}} X(\tau_1, \ldots, \tau_k)\right)^{-1},\\ \hat{\boldsymbol{\beta}}_k &= B_k\left(B_{0k}^{-1}\boldsymbol{\beta}_{0k} + \frac{1}{\sigma^2} X(\tau_1, \ldots, \tau_k)\mathbf{y}_n\right),\\ \nu &= \nu_0 + n,\\ \delta &= \delta_0 + \|\mathbf{y}_n - X(\tau_1, \ldots, \tau_k)\boldsymbol{\beta}_k\|^2.\end{aligned}$$

The univariate conditional posterior distributions of $\tau_u|\boldsymbol{\tau}_k^{(-u)}$, $u = 1, \ldots, k$, are given by

$$P(\tau_u = r|\mathbf{y}_n, \beta_k, \sigma^2, \tau_{u-1} = l', \tau_{u+1} = l'') \propto \exp\left\{-\frac{1}{2\sigma^2}\|\mathbf{y}_n - X(\tau_1, \ldots, \tau_u = r, \ldots, \tau_k)\boldsymbol{\beta}_k\|^2\right\} \tag{10}$$

for $r = l' + 1, \ldots, l'' - 1$, where $\boldsymbol{\tau}_k^{(-u)} = \{\tau_1, \ldots, \tau_{u-1}, \tau_{u+1}, \ldots, \tau_k\}$ with $\tau_0 = x_{l+1}$ and $\tau_k = x_{n-l-k+1}$. The MCMC samples from the conditional distributions in expressions (9) and (10) are standard and can be drawn by using the method that was described in Chib (1995).

When the errors ε_i, $i = 1, \ldots, n$, are independent but not identically distributed, we assume that $\varepsilon_i \sim N(0, w_i\sigma^2)$, where the w_i are known and given in equation (1), and σ^2 has an inverted gamma prior as specified in expression (7). In this case, the posterior distribution of the parameters is

$$\begin{aligned}\pi(\boldsymbol{\theta}_k|M_k, \mathbf{y}_n) \propto{}& \pi(\boldsymbol{\beta}_k|M_k)\,\pi(\boldsymbol{\tau}_k|M_k)\,\pi(\sigma^2|M_k)\\ &\times \frac{1}{(\sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\frac{1}{w_i}\{y_i - \mathbf{x}(\tau_1, \ldots, \tau_k)(i)^{\mathrm{T}}\boldsymbol{\beta}_k\}^2\right].\end{aligned}$$

The conditional posterior distributions of the parameters can be derived as in expressions (9) and (10).

Suppose that in model (2) we allow the join points to occur not only at the observed values of the covariate x but also at $m - 1$ points $x_{i,u}$ between any two consecutive data points of x_i and x_{i+1}. The Bayesian model selection approach for this case can be implemented by assuming that, under model M_k, the vector of join points $\boldsymbol{\tau}_k^{\mathrm{T}}$ has support on the extended set $\{x_1, x_{1,1}, \ldots, x_{1,m-1}, x_2, x_{2,1}, \ldots, x_{2,m-1}, x_3, \ldots, x_{n-1}, x_{n-1,1}, \ldots, x_{n-1,m-1}, x_n\}$. Let $y_{i,u}$ be the unobserved values of the response variable y at $x_{i,u}$, $u = 1, \ldots, m - 1$. We rewrite the join point model (2) as

$$y_{i,u} = \beta_0 + \beta_1 x_{i,u} + \sum_{r=1}^{k}\delta_r s_r(x_{i,u}) + \varepsilon_{i,u}, \qquad i = 1, \ldots, n-1, \quad u = 0, \ldots, m-1,$$

$$y_n = \beta_0 + \beta_1 x_n + \sum_{r=1}^{k}\delta_r s_r(x_n) + \varepsilon_n,$$

where $y_{i,0}$ are observed values and $\varepsilon_{i,u}$ are IID normal random errors with mean 0 and variance σ^2. Under model M_k, the MCMC algorithm can now be implemented on the original set of parameters $\boldsymbol{\theta}_k = (\boldsymbol{\beta}_k^{\mathrm{T}}, \boldsymbol{\tau}_k^{\mathrm{T}}, \sigma^2)^{\mathrm{T}}$ together with the new set of parameters $\{y_{i,u} : u = 1, \ldots, m-1, i =$

Downloaded from https://academic.oup.com/jrsssc/article/54/5/919/7113011 by guest on 19 September 2024

$1, \ldots, n-1\}$. Note that the conditional distribution of $\{y_{i,u} : u = 1, \ldots, m-1, i = 1, \ldots, n-1\}$ given $\mathbf{y}_n$ and $\boldsymbol{\theta}_k$ is a multivariate normal distribution and hence is easy to simulate.

2.2. Computation of posterior probabilities $P(M_k|\mathbf{y}_n)$

Corresponding to any model M_r, let $\boldsymbol{\theta}_r = (\boldsymbol{\beta}_r, \boldsymbol{\tau}_r, \sigma_r^2)$ denote the model-specific parameters. The marginal likelihood for model M_r, in equation (6), can be rewritten as

$$m(\mathbf{y}_n|M_r) = \int f(\mathbf{y}_n|M_r, \boldsymbol{\beta}_r, \boldsymbol{\tau}_r, \sigma_r^2)\, \pi(\boldsymbol{\beta}_r, \boldsymbol{\tau}_r, \sigma_r^2|M_r)\, \mathrm{d}\boldsymbol{\beta}_r\, \mathrm{d}\boldsymbol{\tau}_r\, \mathrm{d}\sigma_r^2. \tag{11}$$

The marginal likelihood $m(\mathbf{y}_n|M_r)$ can be expressed (using the Bayes formula) as

$$m(\mathbf{y}_n|M_r) = \frac{f(\mathbf{y}_n|M_r, \boldsymbol{\beta}_r^*, \boldsymbol{\tau}_r^*, \sigma_r^{2*})\, \pi(\boldsymbol{\beta}_r^*, \boldsymbol{\tau}_r^*, \sigma_r^{2*}|M_r)}{\pi(\boldsymbol{\beta}_r^*, \boldsymbol{\tau}_r^*, \sigma_r^{2*}|M_r, \mathbf{y}_n)}, \tag{12}$$

where $(\boldsymbol{\beta}_r^*, \boldsymbol{\tau}_r^*, \sigma_r^{2*})$ is a high density point. In particular, $(\boldsymbol{\beta}_r^*, \boldsymbol{\tau}_r^*)$ are taken to be the mode and σ_r^{2*} is taken to be the mean. Note that the posterior mean and mode of σ^2 are approximately equal when n is sufficiently large because the mean-to-mode ratio is $(\nu_0 + n - 2)/(\nu_0 + n + 2)$. An estimate of the marginal likelihood in equation (12), on the log-scale, is given by

$$\log\{\hat{m}(\mathbf{y}_n|M_r)\} = \log\{f(\mathbf{y}_n|M_r, \boldsymbol{\beta}_r^*, \boldsymbol{\tau}_r^*, \sigma_r^{2*})\, \pi(\boldsymbol{\beta}_r^*, \boldsymbol{\tau}_r^*, \sigma_r^{2*}|M_r)\} - \log\{\hat{\pi}(\boldsymbol{\beta}_r^*, \boldsymbol{\tau}_r^*, \sigma_r^{2*}|M_r, \mathbf{y}_n)\}, \tag{13}$$

where the estimate of the posterior ordinate $\hat{\pi}(\boldsymbol{\beta}_r^*, \boldsymbol{\tau}_r^*, \sigma_r^{2*}|M_r, \mathbf{y}_n)$ is obtained by using the marginal–conditional decomposition formula

$$\begin{aligned}\pi(\boldsymbol{\beta}_r^*, \boldsymbol{\tau}_r^*, \sigma_r^{2*}|M_r, \mathbf{y}_n) = \pi(\tau_1^*|M_r, \mathbf{y}_n) \left\{ \prod_{u=2}^{r} \pi(\tau_u^*|M_r, \mathbf{y}_n, \tau_1^*, \ldots, \tau_{u-1}^*) \right\} \\ \times \pi(\sigma_r^{2*}|M_r, \mathbf{y}_n, \boldsymbol{\tau}_r^*)\, \pi(\boldsymbol{\beta}_r^*|M_r, \mathbf{y}_n, \sigma_r^{2*}, \boldsymbol{\tau}_r^*).\end{aligned} \tag{14}$$

In equation (14), the first mass function $\pi(\tau_1^*|M_r, \mathbf{y}_n)$ is estimated from output of the full MCMC run, for $u = 2, \ldots, r$, the mass functions $\pi(\tau_u^*|M_r, \mathbf{y}_n, \tau_1^*, \ldots, \tau_{u-1}^*)$ are estimated from the output of a sequence of reduced MCMC runs in which successive join points are fixed at their starred values $\tau_1^*, \ldots, \tau_{u-1}^*$ and the remaining join points $(\tau_v, v = u, u+1, \ldots, r)$ are sampled along with the other parameters $(\boldsymbol{\beta}_r, \sigma_r^2)$, where in each of these MCMC runs τ_u is set equal to τ_u^*. This process is simple and ensures that the τ_u^*, $u = 1, \ldots, r$, are high density points.

Similarly, we estimate the conditional density $\pi(\sigma_r^{2*}|M_r, \mathbf{y}_n, \boldsymbol{\tau}_r^*)$. Finally, we evaluate $\pi(\boldsymbol{\beta}_r^*|M_r, \mathbf{y}_n, \sigma_r^{2*}, \boldsymbol{\tau}_r^*)$. Thus,

$$\left.\begin{aligned}\pi(\tau_1^*|M_r, \mathbf{y}_n) &= \frac{1}{G}\sum_{g=1}^{G} \pi(\tau_1^*|M_r, \mathbf{y}_n, \tau_2^{(g)}, \ldots, \tau_r^{(g)}, \boldsymbol{\beta}_r^{(g)}, \sigma_r^{(g)2}), \\ \pi(\tau_u^*|M_r, \mathbf{y}_n, \tau_1^*, \ldots, \tau_{u-1}^*) &= \frac{1}{G}\sum_{g=1}^{G} \pi(\tau_u^*|M_r, \mathbf{y}_n, \tau_1^*, \ldots, \tau_{u-1}^*, \\ &\qquad \tau_{u+1}^{(g)}, \ldots, \tau_r^{(g)}, \boldsymbol{\beta}_r^{(g)}, \sigma_r^{(g)2}), \qquad u = 2, \ldots, r, \\ \pi(\sigma_r^{2*}|M_r, \mathbf{y}_n, \boldsymbol{\tau}_r^*) &= \frac{1}{G}\sum_{g=1}^{G} \pi(\sigma_r^{2*}|M_r, \mathbf{y}_n, \tau_1^*, \ldots, \tau_r^*, \boldsymbol{\beta}_r^{(g)}),\end{aligned}\right\} \tag{15}$$

where $\pi(\theta^*|z^*, w^{(g)})$ denotes the posterior conditional density of θ evaluated at θ^* when z is set

Downloaded from https://academic.oup.com/jrsssc/article/54/5/919/7113011 by guest on 19 September 2024

at z^* and the reduced MCMC run is $w^{(g)}$. The posterior probabilities can now be estimated by

$$\hat{P}(M_k|\mathbf{y}_n) = \hat{m}(\mathbf{y}_n|M_k) \Big/ \sum_{r=0}^{K} \hat{m}(\mathbf{y}_n|M_r), \qquad k = 0, \ldots, K, \tag{16}$$

and the BF B_{kl} can be estimated by

$$\hat{B}_{kl} = \hat{m}(\mathbf{y}_n|M_k)/\hat{m}(\mathbf{y}_n|M_l).$$

For other methods for estimating the marginal likelihood see Ritter and Tanner (1992), Newton and Raftery (1994), Kass and Raftery (1995), Zellner and Min (1995), Chib (1996) and Chib and Jeliazkov (2001), among others. Chib and Jeliazkov (2001) showed how to improve the Chib (1995) method when one or more full conditional is sampled by the Metropolis–Hastings algorithm. Recently, Mira and Nicholls (2004) have provided a modified version of the Chib and Jeliazkov estimator based on analogies to bridge sampling but usually the gains from this modification are typically small. Furthermore, this modification does not apply to our problem here because no conditional distribution is sampled by the Metropolis–Hastings algorithm.

An implementation via reversible jump MCMC sampling (Green, 1995; Denison *et al.*, 1998) can automatically provide posterior probabilities for the different models. However, the reversible jump method is much more delicate and difficult to set up, especially when there are a few candidate models to consider. In our case, obtaining the marginal likelihood of each model directly is simpler.

2.3. Computation of Bayes information criterion BIC (M_k)

For the linear model with independent normal errors ε_i, Schwarz's (1978) BIC for model M_k is equivalent to

$$\mathrm{BIC}(M_k) = \log\left\{\frac{\mathrm{RSS}(M_k)}{n}\right\} + \frac{p}{n}\log(n), \tag{17}$$

where $\mathrm{RSS}(M_k) = \Sigma_{i=1}^{n}(y_i - \hat{y}_i^{(k)})^2$ is the residual sum of squares and $\hat{y}_i^{(k)}$ is the predictor of y_i from model M_k based on the modal values of $(\beta_0, \beta_1, \delta_1, \ldots, \delta_k)$. Model M_k with the minimum value of BIC is selected as the best model.

It is common to evaluate BIC or the BF at the posterior mode in a Bayesian context since sometimes the MLE is difficult to find (Chib *et al.*, 2002). In our case, although the MLE can be estimated, it is convenient to use the posterior mode or mean in the Bayesian setting.

2.4. Bayesian model averaging

In standard practice, data analysts typically select a model from some class of models and then proceed as if the model selected had generated the data. This approach ignores model uncertainty, leading to overconfident inferences and decisions that are more risky than one thinks they are (Hoeting *et al.*, 1999). Bayesian model averaging (BMA) provides a coherent mechanism from accounting for model uncertainty. In particular, predictions based on model averaging are known to be better than those based on a fixed model (Clyde and George, 2004). Given the posterior model probabilities $P(M_k|\mathbf{y}_n)$ and the predicted value $\hat{y}_i^{(k)}$ of y_i under model M_k, the predicted value under BMA is

$$\hat{y}_i = \sum_{k=0}^{K} \hat{y}_i^{(k)} \hat{P}(M_k|\mathbf{y}_n).$$

Downloaded from https://academic.oup.com/jrsssc/article/54/5/919/7113011 by guest on 19 September 2024

The BMA predictors $\hat{y}_i$, $i = 1, \ldots, n$, are used to generate the plots of predicted values in the application to several major cancer sites.

3. Application

Prostate cancer and breast cancer are the most common cancers, other than skin cancer, for US men and women respectively; colorectal cancer is the fourth most common cancer for both men and women (American Cancer Society, 2004). Because of the progress in cancer detection methods, e.g. new imaging technologies, tumour markers and biopsy procedures, the incidences of these three cancer sites have experienced dramatic changes during the last three decades. It is of interest to examine the trend in incidence for these cancer sites.

The Bayesian model selection approach for the join point models that were developed above can now be applied to identify changes in the observed age-adjusted cancer incidence trend from 1973 to 1999 for colorectal cancer, prostate cancer, breast cancer in white women and breast cancer in black women. When fitting the join point model by using this software to SEER data in previous analyses, the maximum number of join points for any site is 4, so we set $K = 4$. The value of l is set to 2 because we do not expect join points to occur either at two consecutive years or at the first two or the last two years.

3.1. Specification of priors

Under model M_k $(k = 0, 1, \ldots, 4)$, the prior means and the prior variances for the regression parameters were specified as

$$\left.\begin{array}{c} \beta_0 \sim N(y_1, 100), \\ \beta_1, \delta_1, \ldots, \delta_k \overset{\text{IID}}{\sim} N(0, 10), \\ \sigma^2 \sim \text{IG}(\nu_0/2, \delta_0/2). \end{array}\right\} \tag{18}$$

The prior distributions of regression parameters β_0, β_1 and $\delta_1, \ldots, \delta_k$ are chosen to be quite flat. The choice of the prior mean of y_1 for β_0 does not affect the posterior analysis; thus it may be merely thought of as a starting value.

When the errors are IID, the prior mean and variance for σ^2 are taken as

$$\frac{\delta_0}{\nu_0 - 2} = \omega$$

and

$$\frac{2\delta_0^2}{(\nu_0 - 2)^2(\nu_0 - 4)} = 4\omega^2,$$

so that the prior standard error is within two units of the prior mean for σ^2. The value of ω is set to be 0.0001 in the application. The rationale for this choice of the prior mean of σ^2 is as follows. σ^2 is the common variance of the y_is, and under the Poisson assumptions for the number of incidence or mortality events, d_{ij}, it follows from equation (1) that, for equal values of the c_js, the variance of y_i is of order $(\Sigma_{j=1}^{J} d_{ij})^{-1}$. For most of the common cancer sites $\Sigma_{j=1}^{J} d_{ij} \approx 10000$ (American Cancer Society, 2004): hence the choice. Another choice of ω is $\bar{w}$, which is the average of w_i, the variance of ε_i. This choice is also motivated from equation (1). However, since $\bar{w}$ is data dependent, we do not use it in our analysis, but we would study the sensitivity of the prior distribution of σ^2 with respect to ω as the prior mean of σ^2. The prior for τ_k was taken to be as defined in expression (7) with the default value of $l = 2$.

Downloaded from https://academic.oup.com/jrsssc/article/54/5/919/7113011 by guest on 19 September 2024

When $\varepsilon_i \sim N(0, \omega_i \sigma^2)$, the weight ω_i was chosen to be the inverse of the variance of the cancer incidence rate obtained from the observed data, and the weights were standardized so that $\Sigma_{i=1}^{n} \omega_i = \bar{\omega}$.

3.2. Analysis of the incidence data

The Bayesian model selection procedure was applied to the observed age-adjusted incidence rates in the USA for colorectal cancer, prostate cancer, breast cancer in white women and breast cancer in black women for the period from 1973 to 1999. The data were obtained from the SEER programme with the SEER-STAT software (`http://www.seer.cancer.gov/seerstat`). The colorectal cancer data are for both males and females and the prostate cancer data relate to males only. The logarithm of the age-adjusted rate y_i at time x_i ($x_1 = 1, \ldots, x_{27} = 27$) were used in executing the Bayesian model selection computer program with the prior distributions as stated above.

The convergence of the MCMC samples of the parameters $\boldsymbol{\theta}_k = (\beta_0, \beta_1, \delta_1, \ldots, \delta_k, \tau_1, \ldots, \tau_k, \sigma^2)$ for all four cancer sites after 10000 MCMC simulations excluding 500 burn-in samples was examined by running the results of the MCMC simulations through the CODA (output analysis and diagnostics for MCMC simulations) package in R (`http://www.fis.iarc.fr/coda/`). As no single method is foolproof, we present the results of Geweke's (1992) and Heidelberger and Welch's (1983) convergence diagnostics. To see how stable the final estimates of the marginal likelihoods and posterior probabilities were, multiple independent runs were carried out. Gelman and Rubin's (1992) diagnostic is also presented by running more than two parallel chains with starting values that are overdispersed relative to the posterior distribution. Gelman and Rubin's (1992) diagnostic calculates the 'potential scale reduction factor' R for each parameter in $\boldsymbol{\theta}_k$, together with upper and lower confidence limits. Approximate convergence is diagnosed when the upper limits are close to 1.

The MCMC samples for the parameters for all four cancer sites converge after 10000 simulations. Using colorectal cancer as an example, the Geweke statistics are (−0.04703, 0.84650, −0.63558, 1.15226, −0.93957, 1.52299, 1.83233) for the parameters $(\beta_0, \beta_1, \delta_1, \delta_2, \tau_1, \tau_2, \sigma^2)$. Both stationarity and interval half-width tests (Heidelberger and Welch, 1983) passed for all the parameters. 10 independent runs were generated with random starting values for the prior means, where the starting values are equal to the posterior mean plus or minus twice the posterior standard deviations. For example, the starting values of the prior mean for σ^2 is $0.0002 \pm 2 \times 0.00006 = 0.00008$ or 0.00032. The posterior probabilities $P(M_k|\mathbf{y}_n)$ for the 10 runs are the same to the third digits and the best models picked are M_2. The posterior means of all the parameters are essentially the same, as shown in Table 1. The factors R for the parameters are (1.00, 1.00, 1.01, 1.00, 1.01, 1.01, 1.16). This strengthens the diagnosis of convergence of the MCMC algorithm.

Under model M_k, let $(\beta_0^{(g)}, \beta_1^{(g)}, \delta_1^{(g)}, \ldots, \delta_k^{(g)}, \tau_1^{(g)}, \ldots, \tau_k^{(g)})$, $g = 1, \ldots, G$, be the MCMC samples that are generated (not including the burn-in samples) by using the method that was described in Section 2.2. Note that $(\beta_0^{(g)}, \beta_1^{(g)}, \delta_1^{(g)}, \ldots, \delta_k^{(g)}, \tau_1^{(g)}, \ldots, \tau_k^{(g)})$ are the modal values for $(\beta_0, \beta_1, \delta_1, \ldots, \delta_k, \tau_1, \ldots, \tau_k)$ from each MCMC run. The predicted value of y_i was computed as

$$\hat{y}_i^{(k)} = \frac{1}{G} \sum_{g=1}^{G} \{\beta_0^{(g)} + \beta_1^{(g)} x_i + \delta_1^{(g)} (x_i - \tau_1^{(g)})^+ + \ldots + \delta_k^{(g)} (x_i - \tau_k^{(g)})^+\}.$$

The BIC value for model M_k was computed by using equation (17) where RSS is based on $\hat{y}_i^{(k)}$. The default value for the maximum number of join points is $K = 4$. The estimates of the

Downloaded from https://academic.oup.com/jrsssc/article/54/5/919/7113011 by guest on 19 September 2024

Table 1. Estimates of the regression parameters under the best model M_2 for colorectal cancer (1973–1999)

Parameter	*Results from the Bayesian method*				*Results from the PTB method*	
	Prior		*Posterior*			
	Mean	*Standard deviation*	*Mean*	*Standard deviation*	*Mean*	*Standard deviation*
Constant (β_0)	4.0860	10.0000	4.0962	0.0085	4.1080	0.0082
β_1	0.0000	3.1623	0.0108	0.0011	0.0105	0.0010
δ_1	0.0000	3.1623	−0.0303	0.0025	−0.0308	0.0024
δ_2	0.0000	3.1623	0.0233	0.0058	0.0244	0.0066
σ^2	0.0001	0.0002	0.0002	0.00006	0.0002	

posterior probabilities (16) of the models $M_0, M_1, \ldots, M_4$ were also computed from the posterior distribution of the parameters.

All the three approaches, the BF, BIC and the PTB method, chose M_2 as the optimal model. As one referee suggested, BIC can be computed by using the MLE. We compared the posterior estimates from the Bayesian method and the MLEs for $\boldsymbol{\theta}_k, k = 0, \ldots, 4,$ for all four cancer sites; they were very close (the data are not shown). Hence, the BIC values that were calculated on the basis of the posterior estimates and MLEs were similar. As an example, we present the posterior parameter estimates from the Bayesian method and the MLEs for colorectal cancer in Table 1. As expected, since the prior distributions of the regression parameters are quite flat, the Bayes estimates of the regression parameters, and hence the values of APCs, and the variance parameter σ^2 were close to their corresponding estimates based on the PTB method.

In Fig. 1, we show the posterior probabilities $p(M_k|\mathbf{y}_n)$ of $\{M_k, k = 0, 1, \ldots, 4\}$ in the left-hand panels and the marginal posterior densities conditional on the best model, $P(\tau_r = x|M_k, \mathbf{y}_n)$, for each of the join points in the right-hand panels. For colorectal cancer (Fig. 1(a)), model M_2 with probability 0.78 is the best model. The histograms underlined with —1— and —2— correspond to the marginal posterior densities $p(\tau_1|M_2, \mathbf{y}_n)$ and $p(\tau_2|M_2, \mathbf{y}_n)$ with each summing to 1. The most probable values (modes) for τ_1 and τ_2 are 1986 and 1995. The PTB method resulted in the same best model and join points as the Bayesian method. For comparison, we present the posterior probabilities $p(M_k|\mathbf{y}_n)$ and BIC(M_k) for all four cancer sites in Table 2. For colorectal cancer and breast cancer in white women, all three model selection methods picked the same optimal model.

The unconditional probability that there is a join point at x ($x = 1975, \ldots, 1997$) is

$$P(x \text{ is a join point}|\mathbf{y}_n) = \sum_{k=1}^{K} \left\{ \sum_{r=1}^{k} P(\tau_r = x|M_k, y_n) \right\} P(M_k|y_n).$$

This is the average of the posterior probabilities of join points over all models M_k (Fig. 2). The unconditional probabilities for the join points for colorectal cancer, prostate cancer and breast cancer in white women have shapes that are similar to their conditional probabilities for the best model. In Fig. 2, the unconditional probabilities for the join points for breast cancer in black women show two modes at 1978 and 1985 although the best model based on the BF is M_1 with join point mode at 1977. This is because model M_2 also has substantial posterior probability

Downloaded from https://academic.oup.com/jrsssc/article/54/5/919/7113011 by guest on 19 September 2024

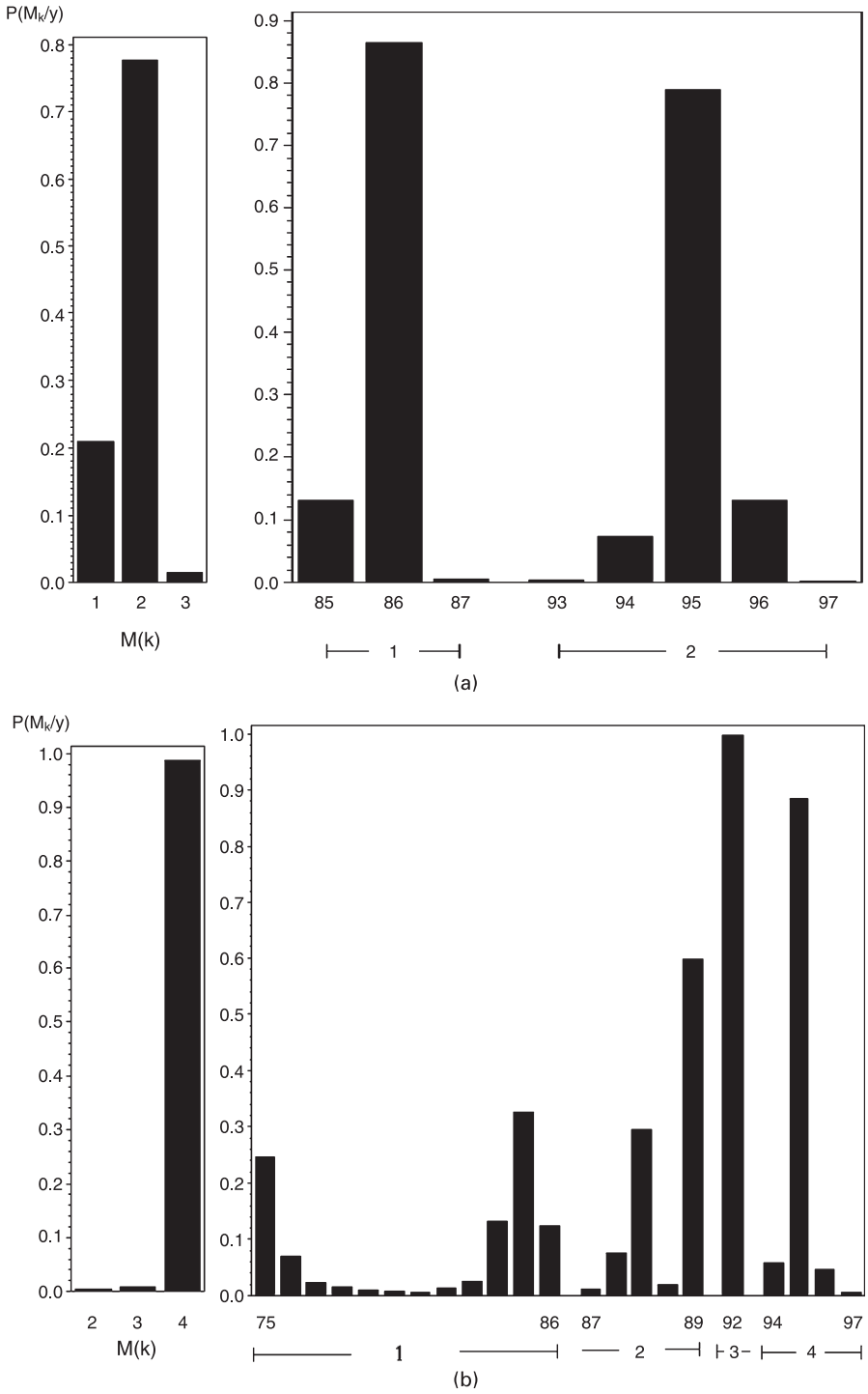

Fig. 1. Posterior probabilities $P(M_k|\mathbf{y}_n)$ and $P(\tau_r|M_k, \mathbf{y}_n)$ under the best model based on the BF for four cancer sites: (a) colorectal cancer; (b) prostate cancer; (c) breast cancer in white women; (d) breast cancer in black women

Downloaded from https://academic.oup.com/jrsssc/article/54/5/919/7113011 by guest on 19 September 2024

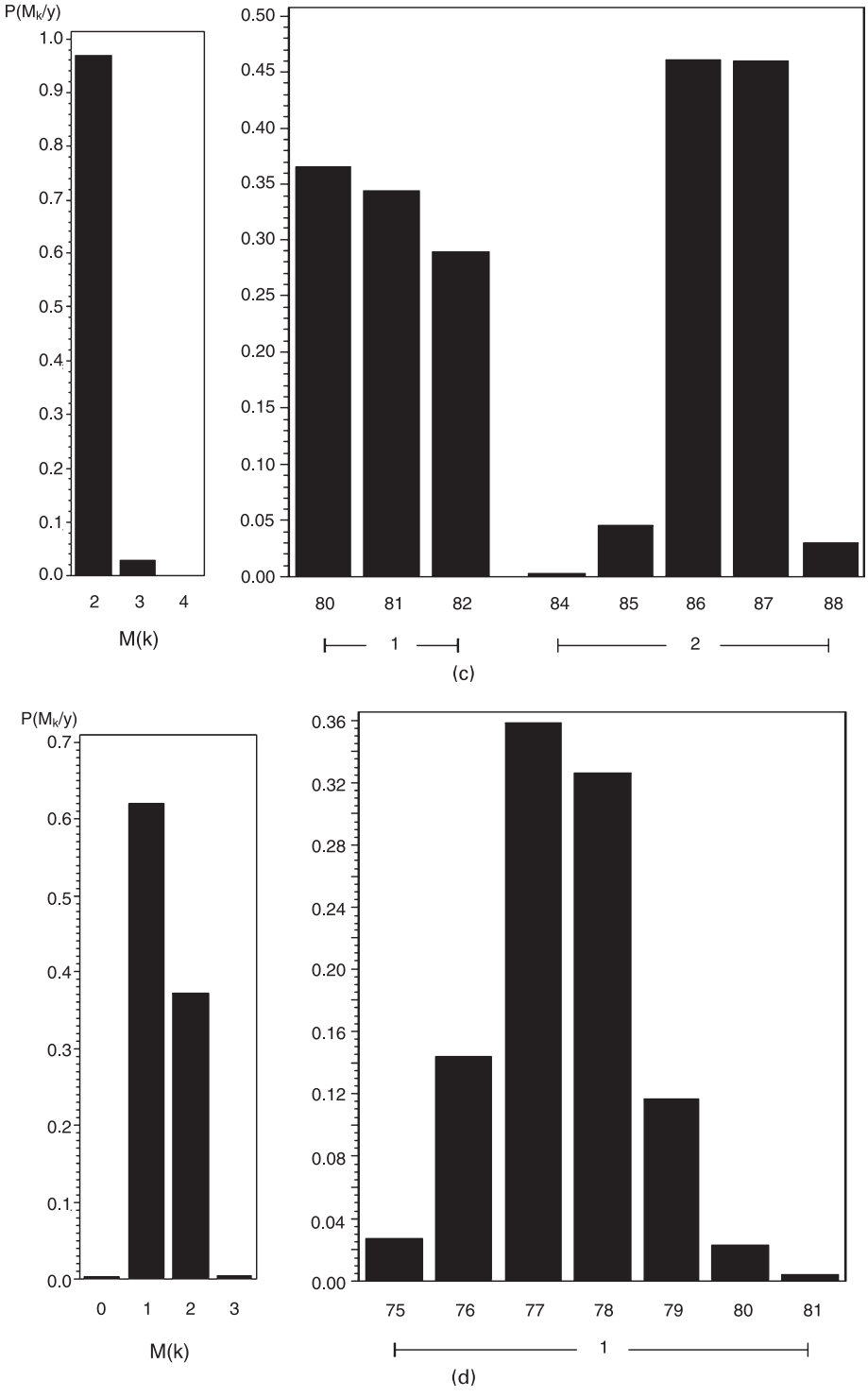

Fig. 1 (*continued*)

Downloaded from https://academic.oup.com/jrsssc/article/54/5/919/7113011 by guest on 19 September 2024

Table 2. Posterior probabilities $P(M_k|\mathbf{y}_n)$ and Bayes information criterion $\mathrm{BIC}(M_k)$ for the four cancer sites†

Cancer site	*Best model (PTB)*	*Bayesian model selection criterion*	*Probabilities for the following values of k:*				
			0	*1*	*2*	*3*	*4*
Colorectal cancer	M_2	BF, $p(M_k\|\mathbf{y}_n)$	0.00	0.21	*0.78*	0.00	0.00
		$\mathrm{BIC}(M_k)$	−5.90	−7.90	*−8.59*	−6.16	−5.32
Prostate cancer	M_4	BF, $p(M_k\|\mathbf{y}_n)$	0.00	0.00	0.00	0.01	*0.99*
		$\mathrm{BIC}(M_k)$	−3.51	−3.92	−5.65	*−5.82*	−4.53
Breast cancer (white women)	M_2	BF, $p(M_k\|\mathbf{y}_n)$	0.00	0.00	*0.95*	0.04	0.00
		$\mathrm{BIC}(M_k)$	−5.64	−5.87	*−6.98*	−6.36	−5.90
Breast cancer (black women)	M_2	BF, $p(M_k\|\mathbf{y}_n)$	0.00	*0.63*	0.36	0.00	0.00
		$\mathrm{BIC}(M_k)$	−5.81	−6.03	*−6.12*	−5.07	−4.55

†Numbers in italics indicate the optimal model M_k from the BF or BIC method.

0.35 of being the best model. Note that the sum of the unconditional probabilities is not 1 over all x. The probabilities in Fig. 2 stand for a series of success probabilities for a Bernoulli trial at x from 1975 to 1999.

For all four sites, the graphs of the join point models by using the fitted values $\hat{y}_i^{(k)}$ under the best model M_k selected by the three model selection methods are given in Fig. 3, which also gives the plots of predicted value $\hat{y}_i$ by using BMA. The predicted values of y_i from the PTB method are very close to those from BIC and the BF and to those from BMA. When the PTB method and both Bayesian methods pick the same best model M_k, the regression coefficients $(\beta_0, \beta_1, \delta_1, \ldots, \delta_k)$, and hence the APC values, are very close (the results are not shown).

The priors that were employed for the regression coefficients in the Bayesian analysis were quite flat. To study the sensitivity of the prior for σ^2 with respect to the choice of prior mean ω for σ^2, we took the prior variance of σ^2 to be $4\omega^2$. For the application, the value of ω was chosen to be 0.0001. As the value of $\bar{w}$ for most of the common cancer sites, e.g. prostate, lung, breast and colorectal, is of the order 0.0001 the Bayesian analysis using $\bar{w}$ as the prior mean for σ^2 led to the same number of join points as with $\omega = 0.0001$. It is interesting that we did not observe any additional join points for any of the four cancer sites that were analysed above when we lowered ω from 10^{-4} to say 10^{-6}. When we increased ω from 10^{-4} to 10^{-2}, the number of join points that were selected by BIC remained the same as reported in Table 2 for all four cancer sites; the number of join points that were selected by the BF, however, decreased to 1, 2, 0 and 0 for colorectal cancer, prostate cancer, breast cancer in white women and breast cancer in black women respectively. This showed that the BF method was sensitive to the specification of the prior value ω. We also examined a rare cancer site, i.e. brain cancer. The PTB method selected one join point and the BIC method selected one join point for both $\omega = 10^{-4}$ and $\omega = 10^{-2}$.

When the errors were not identically distributed, the standardized inverses of the variances for cancer mortality rates were used as weights ω_i in the analysis. Because the weights were close throughout 1973–1999, the results remained similar, i.e. the same numbers of join points were picked and the parameter estimates were close. When the join points were allowed to be between two data points in the augmented data $x_i, x_{i,1}, \ldots, x_{i,m-1}, x_{i+1}$, $i = 1, \ldots, n-1$, with $m = 2$, the same numbers of join points were picked and the parameter estimates were close. However, compared with Fig. 2, the posterior distributions of the join points were more spread, but the shape and range of the histograms remained the same.

Downloaded from https://academic.oup.com/jrsssc/article/54/5/919/7113011 by guest on 19 September 2024

Downloaded from https://academic.oup.com/jrsssc/article/54/5/919/7113011 by guest on 19 September 2024

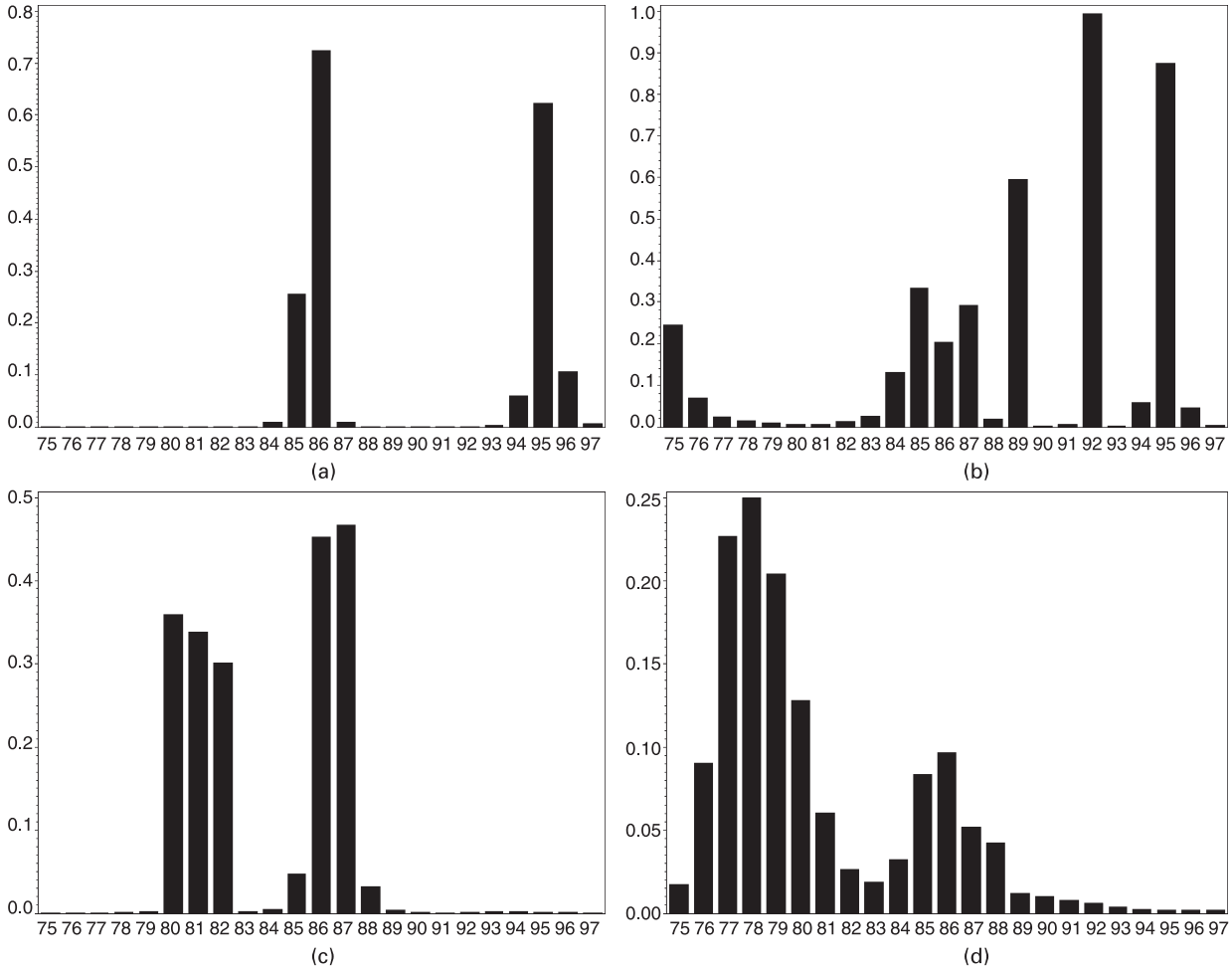

Fig. 2. Unconditional probabilities that there is a join point at a specified year x for (a) colorectal cancer, (b) prostate cancer, (c) breast cancer in white women and (d) breast cancer in black women

Downloaded from https://academic.oup.com/jrsssc/article/54/5/919/7113011 by guest on 19 September 2024

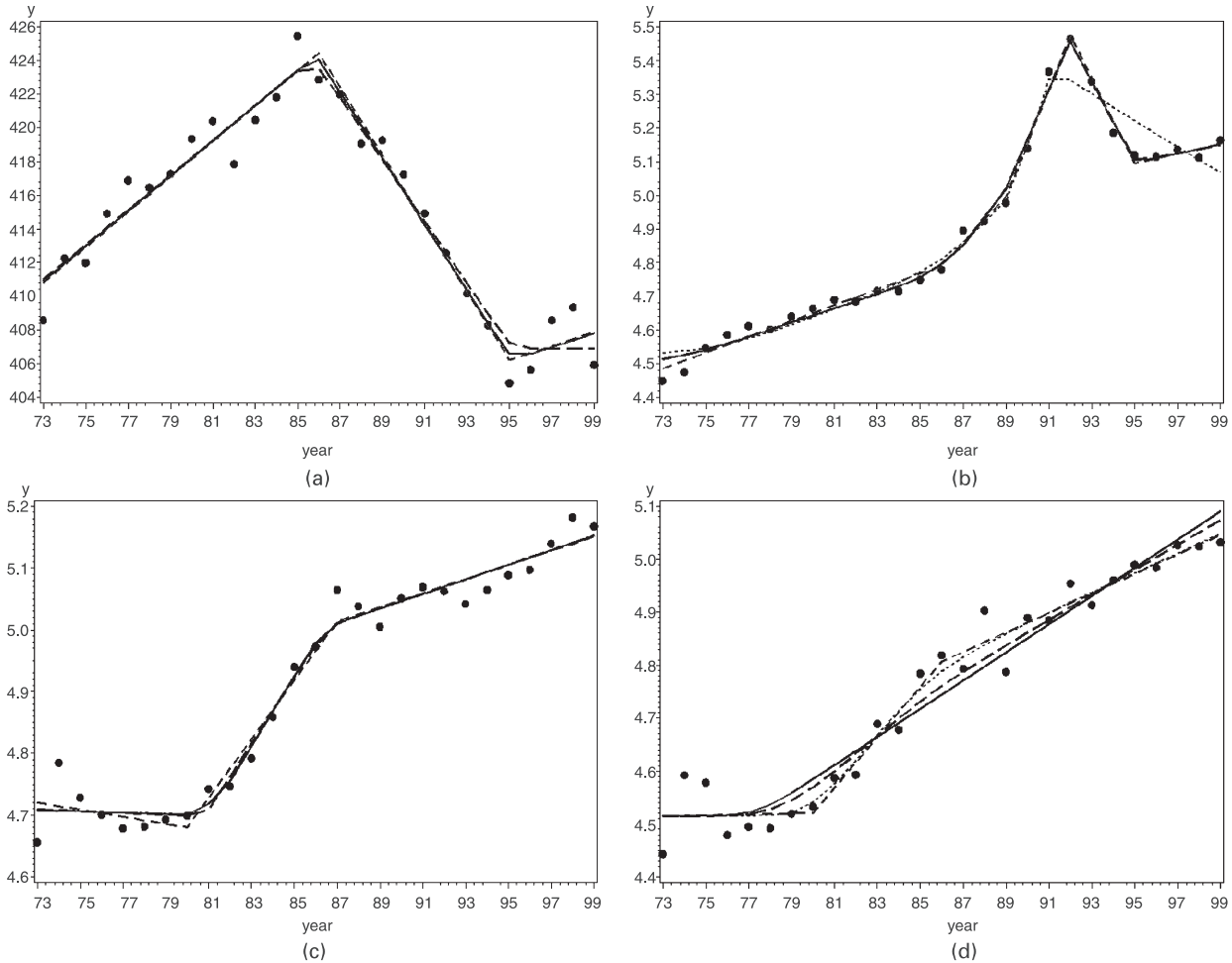

Fig. 3. Predicted values $\hat{y}$ under the best join point models and BMA (•, data point; ———, BF; · – · – ·, BIC; · · · · · · · , PTB method; – – –, BMA): (a) colorectal cancer; (b) prostate cancer; (c) breast cancer in white women; (d) breast cancer in black women

4. Simulation studies

Let $n = 27$ and $x \in \{1, 2, \ldots, 27\}$. Assume that the join point model is log-linear with normal error ε,

$$\log(y) = \beta_0 + \beta_1 x + \sum_{r=1}^{k_0} \delta_r (x - \tau_r)^+ + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2),$$

with the true number of join points $k_0 = 1$ or $k_0 = 2$, $\beta_0 = 5$, $\beta_1 = \log(1 + 0.01\ \mathrm{APC}_1)$ and $\delta_r = \log(1 + 0.01\ \mathrm{APC}_{r+1}) - \log(1 + 0.01\ \mathrm{APC}_r)$. The parameters $(\mathrm{APC}_r, \sigma^2)$ are specified in Tables 3 and 4. The signal-to-noise ratio $|\delta_r|/\sigma$ indicates the magnitude of the change. The simulations were carried out as follows.

(a) Generate the data from the above models.
(b) Fit the Bayesian join point models and PTB join point model with 5% significance level to the simulated data. To reduce the computation time, the maximum number of join points K is set to be 3. The prior specifications for the parameters $\boldsymbol{\theta}_k = (\beta_0, \beta_1, \delta_1, \ldots, \delta_k, \tau_1, \ldots, \tau_k, \sigma^2)$ are given in expression (18). The prior mean and variance for σ^2 are ω and $4\omega^2$. To assess the sensitivity to the prior of σ^2, the models were fitted for $\omega = 10^{-6}, 10^{-4}, 10^{-2}$.
(c) Find the best join point model M_k with k join points and the corresponding coefficients $(\beta_0, \beta_1, \delta_1, \ldots, \delta_k)$ and the Bayesian estimates of the join point locations $(\tau_1, \ldots, \tau_k)$.
(d) Repeat steps (a)–(c) 500 times. Find the frequencies of the best model M_k that has the correct number of join points, i.e. $k = k_0$. The posterior means $\hat{\tau}_j$ are rounded to the closest integer. The root-mean-square error (RMSE) of the estimates of the join points, $\hat{\tau}_j$, was calculated as

Table 3. Optimal model M_k selected from three model selection methods and the RMSE of join point estimates $\hat{\tau}_k$ for the one-join-point model $\log(y) = \beta_0 + \beta_1 x + \delta_1 (x - \tau_1)^+ + \varepsilon$, where $\tau_1 = 13$

σ^2	*APC*	$\lvert\delta_1\rvert/\sigma$	*Method*	ω	$\%(M_0)$	$\%(M_1)$	$\%(M_2)$	$\%(M_3)$	$\Delta(\hat{\tau}_1 \mid M_1)$
0.0002	(3,2)	0.7	PTB		1.8	97.4	0.8	—	1.633
			BIC	10^{-4}	—	85.6	9.4	5.0	1.610
			BIC	10^{-2}	—	97.2	2.8	—	1.464
			BF	10^{-4}	2.4	97.4	0.2	—	1.572
			BF	10^{-2}	100.0	—	—	—	†
	(3,1)	1.4	PTB		—	99.0	1.0	—	0.623
			BIC	10^{-4}	—	84.4	10.2	5.4	0.608
			BIC	10^{-2}	—	97.4	2.6	—	0.658
			BF	10^{-4}	—	100.0	—	—	0.620
			BF	10^{-2}	—	100.0	—	—	0.659
0.0010	(3,2)	0.3	PTB		68.0	31.4	0.6	—	3.485
			BIC	10^{-4}	20.4	64.2	10.4	5.0	3.224
			BIC	10^{-2}	26.6	68.8	4.4	0.2	2.574
			BF	10^{-4}	80.4	19.6	—	—	3.148
			BF	10^{-2}	99.6	0.4	—	—	3.536
	(3,1)	0.6	PTB		4.4	94.4	1.2	—	1.604
			BIC	10^{-4}	—	85.0	8.2	6.8	1.561
			BIC	10^{-2}	—	94.8	4.8	0.4	1.579
			BF	10^{-4}	3.4	96.4	0.2	—	1.532
			BF	10^{-2}	36.6	63.4	—	—	1.531

†Not applicable.

Downloaded from https://academic.oup.com/jrsssc/article/54/5/919/7113011 by guest on 19 September 2024

Table 4. Optimal model M_k selected from three model selection methods and the RMSE of join point estimates $\hat{\tau}_k$ for a two-join-point model $\log(y) = \beta_0 + \beta_1 x + \delta_1(x - \tau_1)^+ + \delta_2(x - \tau_2)^+ + \varepsilon$, where $\tau_1 = 8$ and $\tau_2 = 18$

σ^2	*APC*	$(\lvert\delta_1\rvert/\sigma, \lvert\delta_2\rvert/\sigma)$	*Method*	ω	$\%(M_0)$	$\%(M_1)$	$\%(M_2)$	$\%(M_3)$	$\Delta(\hat{\tau}_1 \mid M_2)$	$\Delta(\hat{\tau}_2 \mid M_2)$
0.0002	(2,3,1)	(0.7,1.4)	PTB		0.2	48.4	50.8	0.6	1.904	0.874
			BIC	10^{-4}	—	9.6	79.0	11.4	2.251	0.836
			BIC	10^{-2}	—	23.2	69.6	7.2	4.311	1.247
			BF	10^{-4}	—	66.6	33.4	—	2.102	0.800
			BF	10^{-2}	95.8	4.2	—	—	†	†
	(1,3,1)	(1.4,1.4)	PTB		—	0.4	99.6	—	0.895	0.870
			BIC	10^{-4}	—	—	86.2	13.8	0.866	0.851
			BIC	10^{-2}	—	—	95.6	4.4	1.248	0.769
			BF	10^{-4}	3.4	0.2	96.4	—	0.865	0.827
			BF	10^{-2}	100.0	—	—	—	†	†
0.0010	(2,3,1)	(0.3,0.6)	PTB		36.4	56.4	7.2	—	3.362	2.108
			BIC	10^{-4}	5.0	51.8	36.0	7.2	4.068	2.003
			BIC	10^{-2}	8.0	63.4	27.8	0.8	3.701	1.365
			BF	10^{-4}	57.4	40.8	1.8	—	3.266	2.309
			BF	10^{-2}	97.4	2.6	—	—	†	†
	(1,3,1)	(0.6,0.6)	PTB		49.6	14.4	36.0	—	2.133	2.022
			BIC	10^{-4}	11.0	4.8	74.8	9.4	2.269	2.088
			BIC	10^{-2}	17.8	6.4	72.2	3.6	2.235	1.636
			BF	10^{-4}	92.0	2.8	5.2	—	2.410	1.871
			BF	10^{-2}	100.0	—	—	—	†	†

†Not applicable.

$$\Delta(\hat{\tau}_j \mid M_k; k = k_0) = \sqrt{[\text{average of } \{(\hat{\tau}_j - \tau_j)^2 : M_k, k = k_0\}]}, \qquad j = 1, \ldots, k.$$

Tables 3 and 4 give the percentages of selecting the best model M_k by the Bayesian model selection methods (the BF and BIC) and the PTB method and the RMSE of the join point estimates $\hat{\tau}_j$, $j = 1, \ldots, k$, conditioning on the best model M_k. The results from $\omega = 10^{-6}$ were very close to those from $\omega = 10^{-4}$ and are not shown here. The PTB method is used as the bench-mark for the comparison with the Bayesian methods (the BF and BIC).

The results for the one-join-point model (Table 3) are summarized below.

(a) The PTB methods selected the correct model M_1 with high percentages except for $\sigma^2 = 0.0010$ and $|\delta_1|/\sigma = 0.3$.
(b) When $\omega = 10^{-2}$, the BIC method picked numbers of M_1 which were similar to those of the PTB method or even outperformed the PTB method ($\sigma^2 = 0.0010$ and $|\delta_1|/\sigma = 0.3$). When $\omega = 10^{-4}$, the BIC method selected M_1 less often.
(c) When $\sigma^2 = 0.0002$ and $|\delta_1|/\sigma = 1.4$, the BF method was perfect. However, the BF method was sensitive to the value ω for the prior. When ω increases from 10^{-4} to 10^{-2}, the BF method picked the correct model M_1 less often.
(d) The RMSEs for $\hat{\tau}_1$ from the three methods were very close.

Thus, the prior mean of the error variance, ω, does affect the Bayesian model selection methods. The BF method is more sensitive. When $\omega = 10^{-2}$, the BIC method outperformed the PTB method.

Downloaded from https://academic.oup.com/jrssc/article/54/5/919/7113011 by guest on 19 September 2024

The results for the two-join-point model (Table 4) are summarized below.

(a) The PTB method picked the correct model M_2 fewer times when σ^2 increased or $|\delta_r|/\sigma$ decreased. For example, when $\sigma^2 = 0.0010$ and $(|\delta_1|/\sigma, |\delta_2|/\sigma) = (0.3, 0.6)$, the PTB method picked the correct model M_2 only 7.2% of times.
(b) The BIC method performed well, especially for $\omega = 10^{-2}$.
(c) The BF method was sensitive to the prior specification ω. The BF method performed well when the change in APC is big and prior $\omega = 10^{-4}$. When $\omega = 10^{-2}$, the BF method essentially picked the model without any join point.
(d) The RMSEs conditioned on the correct model M_2 were similar for all the three models.

Overall, the BIC and the PTB methods selected similar numbers of join points. When the variance $\sigma^2 = 0.001$ or $|\delta_r|/\sigma$ is small, the BIC worked better compared with the PTB. Generally, the BIC method performed better than the PTB and BF methods under the scenarios that were considered in the simulation study.

5. Discussion

The Bayesian model selection method based on the BF and BIC, as demonstrated through the analysis that is carried out in Section 3.2, is a good competitor to the existing PTB approach and has an advantage over the latter as it yields a probability density of the models $M_0, \ldots, M_K$ and the distribution of the join points conditional on the best model M_k. In cases where the posterior probabilities are not concentrated at one mode, as in the case of breast cancer in black women (see Fig. 1(d)), we need to examine the results further. The Bayesian approach also produces a probability distribution on the locations of the join points corresponding to any of the models $M_0, \ldots, M_K$. However, for pure model selection BIC based on the MLE suffices.

On the basis of the results of the application and simulation, we believe that, for the Bayesian analysis to be competitive with the frequentist PTB method, the BIC method with a prior for σ^2 with a mean such as $\omega = 0.01$ is appropriate. There was no effect on the results when we lowered l, the number of observations that are left out at the two ends, from 2 to 0 or to 1. Also, increasing the number of runs from 10000 to 20000 did not make any significant difference in the outcomes of the Bayesian analysis.

As a demonstration, we ran an augmented model M_2 for colorectal cancer. The posterior distributions of join points τ_1 and τ_2 were more spread than their distributions in Fig. 1 (not shown here). Thus, introducing augmented data points x_{iu} without having their observed y_{iu}-values did not help in selecting the best model as it did not result in higher posterior probabilities for the join points. For non-IID errors discussed above, the augmented data case is more complicated as it requires the specification of weights $\{w_i\}$ at the augmented values of y.

Finally, the Bayesian model selection methodology that is developed here can be extended to incorporate the case when join points are continuous. For example, we may assume that the normalized spacings (gaps) defined by the join points have a Dirichlet distribution over the time interval (x_1, x_n). In this case, the posterior density of a join point conditional on the rest and other parameters is a mixture of normal densities and is easy to simulate from. The results will be addressed in a separate paper.

6. Conclusion

Bayesian model selection based on the BF and BIC for comparing a given number of join point regression models was developed and applied to study the trend in the US age-adjusted cancer

Downloaded from https://academic.oup.com/jrsssc/article/54/5/919/7113011 by guest on 19 September 2024

incidence rates for the prostate and colorectal cancer sites. The MCMC method of Chib (1995) was used to generate the samples from the posterior, and to estimate the marginal likelihood of the models under comparison. For the cancer data that were examined here, we found that the BIC method compared quite well with the PTB method. The robustness of the Bayesian model selection methods with respect to the choice of the prior for the error variance in the model was studied, and some extensions were stated.

We recommend the Bayesian methods as a useful supplement to the PTB method, since the posterior distribution of the number and location of the join points gives additional insight into the plausibility of other join point models which could have been selected.

Downloaded from https://academic.oup.com/jrssc/article/54/5/919/7113011 by guest on 19 September 2024

Acknowledgements

The authors thank the Joint Editor and two referees for their helpful comments and suggestions that greatly improved this paper.

References

American Cancer Society (2004) *Cancer Facts and Figures 2004*. Atlanta: American Cancer Society.
Berger, J. O. (1985) *Statistical Decision Theory and Bayesian Analysis*, 2nd edn. New York: Springer.
Carlin, B. P., Gelfand, A. E. and Smith, A. F. M. (1992) Hierarchical Bayesian analysis of changepoint problems. *Appl. Statist.*, **41**, 389–405.
Chib, S. (1995) Marginal likelihood from the Gibbs output. *J. Am. Statist. Ass.*, **90**, 1313–1321.
Chib, S. (1996) Calculating posterior distributions and model estimates in Markov mixture models. *J. Econometr.*, **75**, 79–98.
Chib, S. and Jeliazkov, I. (2001) Marginal likelihood from the Metropolis-Hastings output. *J. Am. Statist. Ass.*, **96**, 270–281.
Chib, S., Nardari, F. and Shephard, N. (2002) Markov chain Monte Carlo methods for stochastic volatility models. *J. Econometr.*, **108**, 281–316.
Clyde, M. and George, E. (2004) Model uncertainty. *Statist. Sci.*, **19**, 81–94.
Denison, D. G. T., Mallick, B. K. and Smith, A. F. M. (1998) Automatic Bayesian curve fitting. *J. R. Statist. Soc.* B, **60**, 333–350.
Edwards, B. K., Howe, H. L., Ries, L. A. G., Thun, M. J., Rosenberg, H. M., Yancik, R., Wingo, P. A., Jemal, A. and Feigal, E. G. (2002) Annual report to the nation on the status of cancer, 1973-1999, featuring implications of age and aging on the US cancer burden. *Cancer*, **94**, 2766–2792.
Gelman, A. and Rubin, D. B. (1992) Inference from iterative simulation using multiple sequences. *Statist. Sci.*, **7**, 457–511.
Geweke, J., (1992) Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In *Bayesian Statistics 4* (eds J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith). Oxford: Clarendon.
Green, P. J. (1995) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82**, 711–732.
Heidelberger, P. and Welch, P. D. (1983) Simulation run length control in the presence of an initial transient. *Ops Res.*, **31**, 1109–1144.
Hinkley, D. V. (1970) Inference about the changepoint in a sequence of random variables. *Biometrika*, **57**, 1–17.
Hinkley, D. V. (1971) Inference about the change-point from cumulative sum tests. *Biometrika*, **58**, 509–523.
Hoeting, J. A., Madigan, D., Raftery, A. E. and Volinsky, C. T. (1999) Bayesian model averaging: a tutorial. *Statist. Sci.*, **14**, 382–417.
Kass, R. and Raftery, A. (1995) Bayes factors. *J. Am. Statist. Ass.*, **90**, 773–795.
Kass, R. E. and Wasserman, L. (1995) A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *J. Am. Statist. Ass.*, **90**, 928–934.
Kim, H. J., Fay, M., Feuer, E. J. and Midthune, D. N. (2000) Permutation tests for joinpoint regression with applications to cancer rates. *Statist. Med.*, **19**, 335–351.
Lerman, P. M. (1980) Fitting segmented regression models by grid search. *Appl. Statist.*, **17**, 77–84.
Mira, A. and Nicholls, G. (2004) Bridge estimation of the probability density at a point. *Statist. Sin.*, **14**, 603–612.
Newton, M. A. and Raftery, A. E. (1994) Approximate Bayesian inference with the weighted likelihood bootstrap (with discussion). *J. R. Statist. Soc.* B, **56**, 3–48.
Pauler, D. K. (1998) The Schwarz criterion and related methods for normal linear models. *Biometrika*, **85**, 13–27.
Pettitt, A. N. (1980) A simple cumulative sum type test statistic for the change point problem with zero-one observations. *Biometrika*, **67**, 79–84.

Ries, L. A. G., Eisner, M. P., Kosary, C. L., Hankey, B. F., Miller, B. A., Clegg, L. and Edwards, B. K. (2002) *SEER Cancer Statistics Review*. Bethesda: National Cancer Institute. (Available from `http://seer.cancer.gov/csr/1973_1999/`.)

Ritter, C. and Tanner, M. A. (1992) Facilitating the Gibbs sampler. *J. Am. Statist. Ass.*, **87**, 861–868.

Schwarz, G. (1978) Estimating the dimension of a model. *Ann. Statist.*, **6**, 461–464.

Smith, A. F. M. (1975) A Bayesian approach to inference about a changepoint in a sequence of random variables. *Biometrika*, **62**, 407–416.

Smith, A. F. M. (1980) Change-point problems: approaches and applications. In *Bayesian Statistics* (eds J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith), pp. 83–89. Valencia: Valencia University Press.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and van der Linde, A. (2002) Bayesian measures of model complexity and fit (with discussion). *J. R. Statist. Soc.* B, **64**, 583–639.

Stephens, D. A. (1994) Bayesian retrospective multiple-changepoint identification. *Appl. Statist.*, **43**, 159–178.

Surveillance, Epidemiology and End Results Program (1999) *The Portable Survival System/Mainframe Survival System*. Bethesda: National Cancer Institute.

Tan, M., Tian, G.-L. and Ng, K. W. (2003) A noniterative sampling method for computing posteriors in the structure of EM-type algorithms. *Statist. Sin.*, **13**, 625–639.

Zellner, A. and Min, C. (1995) Gibbs sampler convergence criteria (GSC^2). *J. Am. Statist. Ass.*, **90**, 921–927.

Downloaded from https://academic.oup.com/jrsssc/article/54/5/919/7113011 by guest on 19 September 2024