Communications for Statistical Applications and Methods
2022, Vol. 29, No. 3, 287–299

https://doi.org/10.29220/CSAM.2022.29.3.287
Print ISSN 2287-7843 / Online ISSN 2383-4757

Estimating dose-response curves using splines: a nonparametric Bayesian knot selection method

Jiwon Lee[a], Yongku Kim[a], Young Min Kim[1,a]

[a]Department of Statistics, Kyungpook National University, Korea

Abstract

In radiation epidemiology, the excess relative risk (ERR) model is used to determine the dose–response relationship. In general, the dose-response relationship for the ERR model is assumed to be linear, linear-quadratic, linear-threshold, quadratic, and so on. However, since none of these functions dominate other functions for expressing the dose-response relationship, a Bayesian semiparametric method using splines has recently been proposed. Thus, we improve the Bayesian semiparametric method for the selection of the tuning parameters for splines as the number and location of knots using a Bayesian knot selection method. Equally spaced knots cannot capture the characteristic of radiation exposed dose distribution which is highly skewed in general. Therefore, we propose a nonparametric Bayesian knot selection method based on a Dirichlet process mixture model. Inference of the spline coefficients after obtaining the number and location of knots is performed in the Bayesian framework. We apply this approach to the life span study cohort data from the radiation effects research foundation in Japan, and the results illustrate that the proposed method provides competitive curve estimates for the dose-response curve and relatively stable credible intervals for the curve.

Keywords: Bayesian analysis, Dirichlet process mixture, dose-response estimation, excess relative risk, splines

1. Introduction

In identifying the potential risk of exposure to ionizing radiation, the excess relative risk (ERR) is a common measure to quantify the relationship between radiation exposure and the risk of cancer incidence or mortality. The crucial research on radiation-associated cancer risk assessment using the ERR have been conducted using the life span study (LSS) cohort data of Japanese atomic bomb survivors in Hiroshima and Nagasaki (Grant *et al.*, 2017). The LSS cohort is a long-term follow-up data and has been used to investigate health-related effects, such as mortality and the incidence of radiation-associated cancers and cardiovascular diseases.

Understanding the dose–response relationship is a main goal for research on radiation-associated health effects. The dose–response function represents this relationship; thus, estimating this function has been a primary concern. In general, parsimonious models are preferred as simple parametric forms, such as linear non-threshold, linear-quadratic, and quadratic functions in radiation epidemiological studies. However, the estimates from these models are sometimes unable to reflect the uncertainty at low doses. In addition, the risk at higher doses is more influential than that at lower doses

This work was supported by the National Research Foundation of Korea (NRF-2019R1F1A1061691) and research Grants of Korea Forest Service project (No.2019149A00-2123-0301).

[1] Corresponding author: Department of Statistics, Kyungpook National University, Daegu 41566, Korea.
E-mail: kymmyself@knu.ac.kr

Published 31 May 2022 / journal homepage: http://csam.or.kr

(Furukawa *et al.*, 2016), emphasizing the need for more sophisticated, flexible, and nonparametric models to examine the low-dose effect.

One alternative for understanding the dose–response relationship is to use piecewise functions which divides the domain into several equally spaced intervals and applies a different subfunction to each interval. Furukawa *et al.* (2016) proposed a piecewise linear dose–response model, setting the similarity among adjacent dose intervals as the prior distribution of coefficients. Park and Kim (2020) employed the same model, but applied a Gaussian process prior to adjust rapid changes in slope at each knot. Similar Bayesian approaches have been studied by Buscot *et al.* (2017), Kern *et al.* (2005), and Holmes and Mallick (2001).

Splines defined by a sum of piecewise polynomials are more flexible methods. When using splines, the number and location of knots should be specified in advance. The important tuning parameters for splines are the number and location of knots. Kauermann and Opsomer (2011) proposed data-driven selection of the number of spline basis functions in the penalized spline regression based on a likelihood criterion. Dung and Tjahjowidodo (2017) investigated the identification of the number and location of knots in non-uniform space using B-spline functions. In addition, Dimatteo *et al.* (2001) used the reversible-jump Markov chain Monte Carlo to capture the number or location of knots for an exponential family. However, the radiation dose distribution is highly skewed (Grant *et al.*, 2017) and is not a member of an exponential family. Thus the choice of equally spaced knots may be inappropriate. Therefore, we propose a knot selection method based on a Dirichlet process mixture model (DPMM), which is one of the most widely used models in nonparametric Bayesian statistics (MacEachern and Müller, 1998). McAuliffe *et al.* (2006) applied the DPMM to nonparametric empirical Bayes problems, and Da Silva (2007) analyzed brain magnetic resonance imaging (MRI) tissue using the DPMM. Straub *et al.* (2015) modeled directional data on a spherical domain utilizing the DPMM, and Ngan *et al.* (2015) used the DPMM for the detection of outliers.

We assume that the dose distribution follows an infinite mixture of normal distributions; thus, we can select knots as the overlapping points of two normal distributions. The DPMM is an infinite mixture model with a countably infinite number of clusters inferred from data (Teh *et al.*, 2006; Teh, 2011). This flexibility makes the DPMM more promising than finite mixture models. After selecting the knots, the coefficients of the spline are estimated using the Bayesian inference. The Bayesian method can compute credible intervals more easily than the frequentist approach, although the computational time may be higher due to the large size of LSS data.

The remainder of this paper is organized as follows. Section 2 describes the ERR model and properties of the radiation dose distribution. In Section 3, we explains the knot selection method based on the DPMM, and then an estimation method of the spline function in a Bayesian framework. Section 4 presents an example using real data to illustrate the proposed knot selection method based on the Bayesian approach, and Section 5 provides concluding remarks.

2. Excess relative risk model

The ERR describes the proportional risk increase above the baseline rate $\lambda(t, d = 0, z)$ defined as

$$\mathrm{ERR}(t, d, z) = \frac{\lambda(t, d, z) - \lambda(t, d = 0, z)}{\lambda(t, d = 0, z)} = \mathrm{RR}(t, d, z) - 1, \tag{2.1}$$

where RR is the relative risk, d is the exposed dose, t is the event time, and z is a vector of covariates, such as age at exposure, birth year, attained age, city etc. In particular, the ERR (2.1) is relevant for modeling the dose–response relationship with effect modification. Thus, the ERR (2.1) can be

fitted to individual failure time data using partial likelihood or to person-year data using Poisson rate regression (Furukawa *et al.*, 2016). The person-year data consists of strata of subjects on radiation exposure dose groups, sex, age at exposure etc. We count the event occurrences for exposed and non-exposed subjects, respectively and compute the sum of each person-year on each stratum (NRC, 2006). In general, ERR models (2.2) tend to be nonlinear in the parameter coefficients using a person-year data. Grant *et al.* (2017) stated that the ERR model for the LSS cohort data (person-year data) at the radiation effects research foundation (RERF), which has conducted health-associated studies for atomic bomb survivors in Hiroshima and Nagasaki, Japan for more than 70 years, is defined as

$$\lambda(t, d, z) = \lambda(t, d = 0, z)\,[1 + \mathrm{ERR}(d, z)]\,. \tag{2.2}$$

The ERR model consists of the dose-response function $\rho(d)$ and the effect modification term $\varepsilon(z)$ as $\mathrm{ERR}(d, z) = \rho(d)\varepsilon(z)$. For example, the effect modification $\varepsilon(z)$ for the LSS data is defined in (3.7). The baseline incidence rate $\lambda(t, d = 0, z)$ is an exponential function of sex, age at exposure, birth year, and other factors. RERF estimated the ERR model using maximum likelihood estimation in general estimating the baseline risk and Furukawa *et al.* (2016) proposed the Bayesian approach using piecewise linear dose–response model. We apply the maximum likelihood estimation for the baseline risk, find the number and locations of knots in the nonparametric Bayesian knot selection method, and then estimate the ERR in the Bayesian inference. Hence,

$$Y_i \sim \mathrm{Pois}\left(\mathrm{PY}_i e^{\eta_i}(1 + \mathrm{ERR}\,(d_i, z_i))\right) \quad \text{for } i = 1, \ldots, n,$$

where Y_i is the number of event occurrences at each stratum, PY_i is the sum of person-years in the i^{th} row, and e^{η_i} is the baseline incidence rate, $\lambda(t_i, d_i = 0, z_i)$. η_i for the LSS data will be described in Section 3.2. Common parametric dose–response functions for $\rho(d)$ are as follows,

$$\begin{aligned} \text{Linear} \quad & \rho(d) = \beta_1 d, \\ \text{Linear-quadratic} \quad & \rho(d) = \beta_1 d + \beta_2 d^2, \\ \text{Quadratic} \quad & \rho(d) = \beta_1 d^2. \end{aligned} \tag{2.3}$$

To determine a more stable dose–response relationship than the above parametric dose-response functions, we consider a spline function as smooth piecewise polynomials. This allows us to apply a polynomial function to each subinterval, and connect the pieces to construct a smooth function. The piecewise quadratic spline function is defined as,

$$\rho(d) = \sum_{k=1}^{2} \beta_k d^k + \sum_{j=1}^{C} \beta_{j+2} \left(d - \delta_j\right)_+^2, \quad \text{for} \quad \left(d - \delta_j\right)_+^2 = \begin{cases} \left(d - \delta_j\right)^2, & d > \delta_j, \\ 0, & \text{otherwise.} \end{cases} \tag{2.4}$$

However, it is difficult to define the correct number and location of knots, $\boldsymbol{\delta} = \{\delta_1, \ldots, \delta_C\}$. The number of knots is usually determined by cross-validation. One of clear ways to determine the location of knots is to divide the domain into subintervals with an equal length. However, as depicted in Figure 1, the dose distribution is highly skewed, and has a long right tail. Thus, equally spaced subintervals may not be suitable for this ERR model.

3. Bayesian inference

3.1. Knot selection based on Dirichlet process mixture model

The goal of this subsection is to divide a spline domain by clustering dose observations. The DPMM and infinite mixture of Gaussian models are widely used to estimate a density function (Escobar and

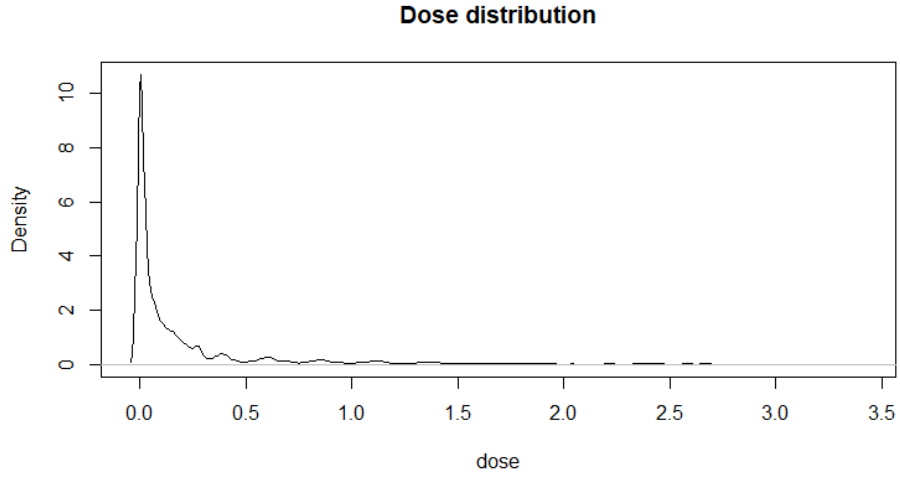

Figure 1: *Distribution of radiation dose in LSS cohort data.*

West, 1995; Müller *et al.*, 1996; Rasmussen, 1999). Although it can be used for density estimation, the DPMM has also been applied to a clustering algorithm (Dahl, 2006; Yu *et al.*, 2010; Reich and Bondell, 2011). The most significant difference between the DPMM and common clustering methods (e.g., K-means or Gaussian finite mixture model) is that the DPMM can estimate the appropriate number of clusters from data. Thus, the DPMM has an advantage because the shape of the spline curve is greatly affected by the number of knots.

Let the density function be $f(x) = \int f(x|\theta)G(\theta), d\theta$ where $G = \sum_{k=1}^{N} \pi_k \delta_{\theta_k}$ is a mixing distribution (Orbanz and Teh, 2010). This simply means that $f(x)$ has a parameter θ, and the assigned probability for each θ_k is π_k. Then, it is written as a finite mixture model. We can extend it to an infinite mixture model with $G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}$. Bayesian assigns the Dirichlet process (DP) as a prior for G, and $f(x)$ is then the DPMM. The DP is a stochastic process with its marginal distribution as a Dirichlet distribution, denoted as $G \sim \text{DP}(\alpha, G_0)$, where α is a concentration parameter and G_0 is the base distribution. α controls π_k and $\theta_k \overset{\text{iid}}{\sim} G_0$. Since the DP is on the infinite dimension, instead of writing it as an explicit formula, many studies have focused on developing construction schemes. Three famous schemes are the Chinese restaurant process, the Polya urn scheme, and the stick-breaking construction. Because the computation time is shorter in the stick-breaking construction for large data, we select this scheme to generate samples from the DP. When the mixing distribution G is a normal distribution, $f(x)$ is a DP Gaussian mixture model (DPGMM). For simplicity, $f(x) = \sum_{k=1}^{\infty} \pi_k N(\mu_k, \tau_k^{-1})$ for $\theta_k = (\mu_k, \tau_k^{-1})$. Using the stick-breaking construction, it can be summarized as follows,

$$f(x) = \sum_{k=1}^{\infty} \pi_k N\left(\mu_k, \tau_k^{-1}\right), \quad \left(\mu_k, \tau_k^{-1}\right) \sim G_0,$$

$$\pi_1 = U_1, \quad U_j \sim \text{Beta}(1, \alpha), \quad \pi_j = U_j \prod_{i=1}^{j-1} (1 - U_i) \quad \text{for } j \geq 2,$$

By properties of the beta distribution, as α is close to zero, only the first several clusters are important. In contrast, as α increases, the assigned probabilities for each cluster become equal. We assume that the density function of the dose distribution is $f_d(\cdot) = \sum_{k=1}^{\infty} \pi_k N(\mu_k, \tau_k^{-1})$. After estimating (μ_k, τ_k^{-1})

and the number of clusters, we select a knot as an overlapping point of two normal distributions that has a maximum likelihood value. We repeat it to obtain reasonable knots. This is demonstrated in Figure 2.

Let d_i be the i^{th} dose observation. Then, the model structure to cluster the dose observations can be written as a stick-breaking construction,

$$d_i \sim f(\theta_i), \quad i = 1, \ldots, n, \tag{3.1}$$

$$\theta_i \sim G, \tag{3.2}$$

$$G = \sum_{k=1}^{N} \pi_k \delta_{\phi_k}, \tag{3.3}$$

$$\phi_k \sim G_0, \tag{3.4}$$

$$\pi_1 = U_1, \quad U_j \sim \text{Beta}(1, \alpha), \quad \pi_j = U_j \prod_{\ell=1}^{j-1} (1 - U_\ell) \quad \text{for } j \geq 2. \tag{3.5}$$

Note that the finite number N is introduced as the number of cluster components. This is because the modeling phase cannot handle infinity. Thus, by using a sufficiently large number N instead of infinity, the probabilities can be calculated. Choosing the exact value of N depends on the concentration parameter, α. Further details about this problem can be found in Ishwaran and James (2002). For this model (3.1) to be the DPGMM, f must be a normal density function with $\theta_j = (\mu_j, \tau_j^{-1})$. In clustering analysis, the cluster indicator variables $\mathbf{K} = (K_1, \ldots, K_n)$ and $\mathbf{Z} = (Z_1, \ldots, Z_N)$ indicate that $K_i = j$ signifies $Z_{K_i} = \theta_j = (\mu_j, \tau_j^{-1})$. With these variables, the model (3.1) can be rewritten as

$$\begin{aligned}
(d_i|\mathbf{Z}, \mathbf{K}) &\overset{\text{ind}}{\sim} N\left(d_i|Z_{K_i}\right), \quad i = 1, \ldots, n, \\
(K_i|\boldsymbol{\pi}) &\overset{\text{iid}}{\sim} \sum_{k=1}^{N} \pi_k \delta_k(\cdot), \\
(\boldsymbol{\pi}, \mathbf{Z}) &\sim \pi(\boldsymbol{\pi}) \times G_0^N(\mathbf{Z}).
\end{aligned}$$

To estimate the parameters, we must obtain values from the posterior distribution of $(\mathbf{Z}, \mathbf{K}, \boldsymbol{\pi}|\mathbf{D})$. This can be achieved by the block Gibbs algorithm (Ishwaran and James, 2001), which sequentially updates each parameter in the following order,

$$\begin{aligned}
&(\mathbf{Z}|\mathbf{K}, \mathbf{D}), \\
&(\mathbf{K}|\mathbf{Z}, \boldsymbol{\pi}, \mathbf{D}), \\
&(\boldsymbol{\pi}|\mathbf{K}).
\end{aligned}$$

In each iteration, $\mathbf{K}^* = \{K_1^*, \ldots, K_m^*\}$ denotes the set of m unique values of $\mathbf{K}$. The procedure to estimate the parameters is described by the following steps,

1. Sample $\mathbf{Z} = (Z_1, \ldots, Z_N)$,

   (a) Sample $\mathbf{Z}^* = (Z_{K_1}^*, \ldots, Z_{K_m}^*)$ from the full conditional distribution $f(Z_{K_j}^*|\mathbf{K}, \mathbf{D})$,

$$f\left(Z_{K_j}^*|\mathbf{K}, \mathbf{D}\right) \propto G_0\left(Z_{K_j}^*\right) \prod_{\{i:K_i=K_j^*\}} f\left(d_i|Z_{K_j}^*\right), \quad j = 1, \ldots, m$$

(b) For each $k \in \mathbf{K} - \{K_1^*, \ldots, K_m^*\}$, draw Z_k from the base distribution G_0.

More specifically, we can rewrite this as an explicit formula. Let $h^*_{K_j} = \#\{i : K_i = K_j^*\}$ and G_0 be a normal-gamma distribution. The normal-gamma distribution can be used as a conjugate prior when we do not know either the mean or variance of a normal distribution (Görür and Rasmussen, 2010). There are four hyperparameters, $\mu_0, \lambda_0, \alpha_0$, and β_0, and we can write it as $\mathrm{NG}(\mu_0, \lambda_0, \alpha_0, \beta_0)$. Then, (a) and (b) can simply be achieved by the following procedure,

1) Draw $\tau_{Z^*_{K_j}} \sim \mathrm{Gamma}(\alpha_0 + (h^*_{K_j})/2,\ \beta_0 + 1/2(h^*_{K_j} s + (\lambda_0 h^*_{K_j}(\bar{d} - \mu_0)^2)/(\lambda_0 + h^*_{K_j})))$
   where $\bar{d} = \sum_{\{i:K_i=K_j^*\}} d_i/h^*_{K_j}$ and $s = \sum_{\{i:K_i=K_j^*\}}(d_i - \bar{d})^2/h^*_{K_j}$
2) Draw $\mu_{Z^*_{K_j}} \sim N((\lambda_0\mu_0 + h^*_{K_j}\bar{d})/(\lambda_0 + h^*_{K_j}),\ 1/\{(\lambda_0 + h^*_{K_j})\tau_{Z^*_{K_j}}\})$
3) For each $k \in \mathbf{K} - \{K_1^*, \ldots, K_m^*\}$,
   i. Draw $\tau_{Z_k} \sim \mathrm{Gamma}(\alpha_0, \beta_0)$
   ii. Draw $\mu_{Z_k} \sim N(\mu_0, 1/(\lambda_0\tau_{Z_k}))$

2. Sample $\mathbf{K} = (K_1, \ldots, K_n)$,

$$(K_i|\mathbf{Z}, \boldsymbol{\pi}, \alpha, \mathbf{D}) \overset{\text{ind}}{\sim} \sum_{1}^{N} \pi^*_{k,i}\delta_k(\cdot), \quad i = 1, \ldots, n,$$

where $\pi^*_{k,i} \propto \pi_k f(d_i|Z_k)$. This is equivalent to sampling K_i from Multinomial$(N, \boldsymbol{\pi}_i^*)$, where $\boldsymbol{\pi}_i^* = (\pi^*_{1,i}, \ldots, \pi^*_{N,i})$

3. Sample $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_N)$,

$$U_j \sim \mathrm{Beta}\left(1 + M_j, \alpha + \sum_{j+1}^{N} M_k\right),$$

$$\pi_1 = U_1, \quad \pi_j = U_j \prod_{\ell=1}^{j-1}(1 - U_\ell), \quad \text{for } j = 1, \ldots, N,$$

where $M_j = \#\{d_i : K_i = j\}$ and $\pi_N = 1 - \sum_{j=1}^{N-1}\pi_j$. To effectively define a measure of α, $U_N = 1$ (Ishwaran and James, 2001). In this paper, we use a fixed α of 0.5. We iterate the above steps until the chain of $(\mathbf{Z}, \mathbf{K}, \boldsymbol{\pi}|\mathbf{D})$ converges. The number of clusters n is determined by choosing the mode of the posterior samples.

3.2. Bayesian inference of spline coefficients

After choosing the knots, we estimate the coefficients $\boldsymbol{\beta}$ of $\rho(d)$ (2.3) using the Bayesian paradigm. We must estimate not only $\boldsymbol{\beta}$, but also the coefficients of $\lambda(t, d = 0, z)$ (2.2) and the effect modification term $\varepsilon(z)$. For simplicity, however, we focus only on estimating $\boldsymbol{\beta}$, and fix other parameter estimates from maximum likelihood estimation. We do this because our priority is to make an inference about the shape of the dose-response curves, and inferences about other parameters remain secondary. Bayesian updates the parameters via MCMC. The Metropolis algorithm is used to draw samples from the multivariate distribution of $\boldsymbol{\beta} = \{\beta_1, \ldots, \beta_{k+c}\}$, $f(\boldsymbol{\beta}|\mathbf{Y}, \boldsymbol{\alpha}, \gamma, \phi)$. The prior distribution

Table 1: Number (n) and percent (%) of cluster components for males and females

		# of clusters				
		6	7	8	9	10
Males	n		3464	1290	225	21
	%		69.28	25.80	4.50	0.42
Females	n	3437	1310	223	227	3
	%	68.74	26.20	4.46	0.54	0.06

of $\boldsymbol{\beta}$ is a multivariate normal distribution with mean 0 and variance 100 considering a noninformative prior. The posterior distribution of $\boldsymbol{\beta}$ is defined as,

$$\begin{aligned} f(\boldsymbol{\beta}|\mathbf{Y}, \boldsymbol{\alpha}, \gamma, \phi) &\propto L(\boldsymbol{\beta}|\mathbf{Y}, \boldsymbol{\alpha}, \gamma, \phi) \times \pi(\boldsymbol{\beta}) \\ &\propto \prod_{i=1}^{n} \frac{\lambda_i^{y_i} e^{-\lambda_i}}{y_i!} \times \exp\left(-\frac{1}{2}\boldsymbol{\beta}^T \Sigma \boldsymbol{\beta}\right), \end{aligned}$$

where $\lambda_i = PY_i e^{\eta_i}(1 + \rho_s(d_i)\varepsilon(z_i))$. Here, the subscript s represents a sex-specific parameter, and $\rho_s(d)$ is the spline defined above. For clarity, explicit formulas for e^{η_i} and $\varepsilon(z_i)$ are defined as follows,

$$\begin{aligned} \exp(\eta_i) = \exp\Bigg[&\alpha_{1s} + \alpha_{2s}\log\left(\frac{a_i}{70}\right) + \alpha_{3s}\log\left(\frac{a_i}{70}\right)^2 + \alpha_{4s}\log\left(\frac{a_i}{70}\right)^2 I(a_i > 70) + \alpha_{5s} b_i \\ &+ \alpha_{11}\text{Hiroshima} \times \text{NIC}_i + \alpha_{12}\text{Nagasaki} \times \text{NIC}_i\Bigg] \end{aligned} \tag{3.6}$$

$$\varepsilon(z_i) = \exp\left(\gamma e_i + \phi_s \log\left(\frac{a_i}{70}\right)\right), \tag{3.7}$$

where a is the attained age, e is the age at exposure, yr is the calendar year, b is the birth year defined as $b = (\text{floor}(yr - a) - 1915)/10$, and NIC is the "not in city" indicator. We ran 30,000 MCMC and treated 5,000 samples for a burn-in period. Due to the high correlation between iterations, every 100th sample remained, and four chains were used. Thus, a total of $250 \times 4 = 1{,}000$ samples approximated the posterior distribution. In addition, we used Gelman-Rubin statistics to diagnose the convergence of the chains. As mentioned, the coefficients $\boldsymbol{\alpha} = (\alpha_{1s}, \ldots, \alpha_{12})$ of $\lambda_0(d)$ and (γ, ϕ_s) of $\varepsilon(z)$ were fixed as the maximum likelihood estimates for simplicity.

4. Application

4.1. Life span study

The LSS cohort pertains to atomic bomb survivors in Hiroshima and Nagasaki, Japan, and provides the risk estimates of cancers related to radiation exposure. The LSS cohort is a main source of radiation-associated risk assessment for humans. The report in Grant *et al.* (2017) contained information for a 52-year follow-up period, and a total of 3,079,484 person-years. The total number of subjects was 105,444, and several covariates, including city, gender, age at exposure, and attained age, were considered. The target outcome was all types of solid cancers, and more details are provided by Grant *et al.* (2017).

4.2. Knot selection results

Table 1 presents the estimated number of cluster components. The mode was used as the appropriate number of clusters for males and females: 7 for males and 6 for females. The estimates of the means

Table 2: Posterior means and variances of each cluster for male and females

		Cluster						
		1	2	3	4	5	6	7
Males	Mean	0.006	0.072	0.193	0.430	0.784	1.314	2.394
	Variance	0.0002	0.0017	0.0066	0.0261	0.0654	0.1402	0.1135
Females	Mean	0.008	0.096	0.238	0.593	1.194	2.314	
	Variance	0.0002	0.0017	0.0261	0.0654	0.0006	0.1402	

Table 3: Knot selection for males and females

	Knot					
	1	2	3	4	5	6
Males	0.03	0.13	0.31	0.60	1.04	1.85
Females	0.05	0.16	0.41	0.89	1.75	

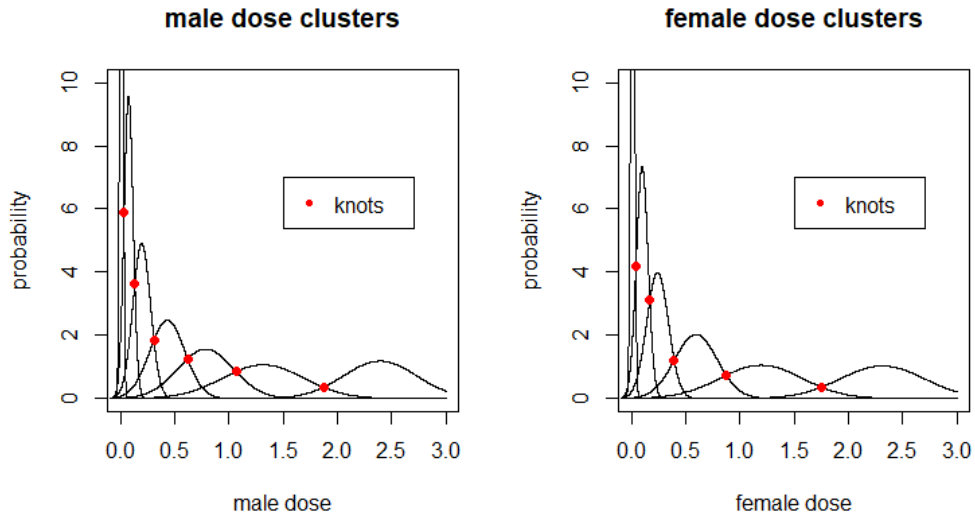

Figure 2: *Knot selection for males and females. The red points denote the overlapping points with the maximum likelihood. The estimated total number of knots is 6 for males and 5 for females.*

and variances of each cluster are presented in Table 2. As illustrated in Figure 2, we selected each red point as a knot that had the maximum likelihood in the overlapping region of two normal distributions, and the results are presented in Table 3. It should be noted that the selected knots in Figure 2 appear to reflect the highly skewed dose distribution in Figure 1. Thus, as expected, the knot selection using the DPMM can be an effective method to draw data-driven knots.

4.3. Sensitivity to the choice of the concentration parameter

Since the concentration parameter α of the DP plays a vital role in determining the number of cluster components, it was important to appropriately determine this parameter. Although it can also be estimated by its posterior distribution, we fixed it to 0.5 for simplicity. However, it was necessary to verify its effect on the final clustering results described in Figure 2. We thus varied its value from 0.1 to 10. Whereas the estimated number of cluster components did not change when α had a value

Table 4: Parameter estimates and Gelman-Rubin statistics for males and females

	Posterior mean								
Sex	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	GR
Males	0.053	0.028	0.030	0.025	0.021	0.016	0.023	0.107	1.05
Females	0.362	0.027	0.028	0.023	0.019	0.022	0.071		1.02

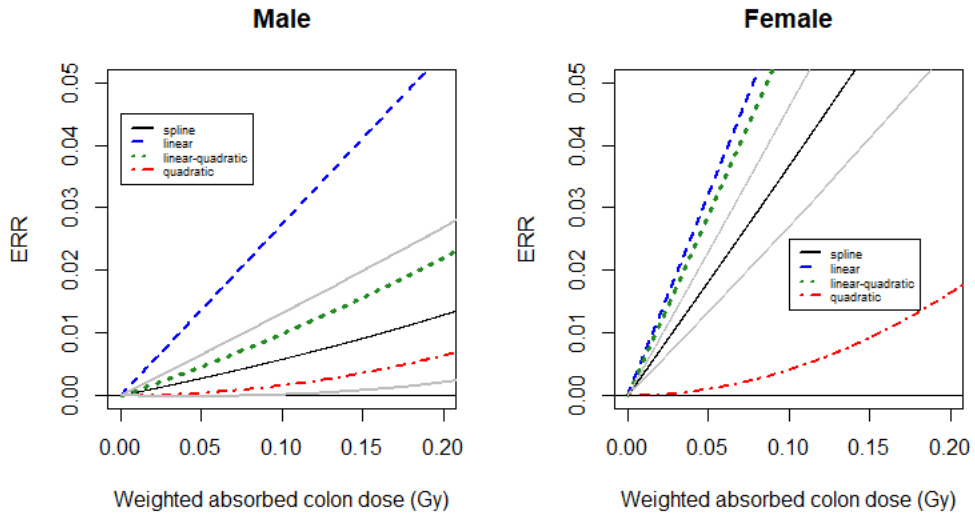

Figure 3: *Dose-response curves for different dose: spline (black, solid), linear (blue, dashed), linear-quadratic (green, dotted), and quadratic (red, dash-dotted). Solid gray lines depict the 95% credible interval of the spline.*

close to 0.5, it increased to 10 and 11 as α became 5 and 10, respectively. The variances of the added clusters, however, were much higher than those of the initially estimated clusters with $\alpha = 0.5$. Therefore, the number of cluster components may not be sensitive to the choice of α.

4.4. Excess relative risk model estimates

The chains converged because the Gelman-Rubin statistics were close to 1 in Table 4. There was also the estimated $\boldsymbol{\beta}$, which was represented as the posterior mean. Based on this result, we compared the ERR calculated by several models for age 70 after exposure at age 30, as illustrated in Figure 3. Overall, the estimated ERR curves for females increased more sharply than those for males. In comparing the individual curves, the spline curve (black, solid) for males was fairly similar to the linear-quadratic (green, dotted) and quadratic curves (red, dash-dotted) but different from the linear curve (blue, dashed). However, it estimated the ERR to be smaller than the other two curves. This was also a prominent feature in the curve estimation for females. In particular, the estimated ERR at doses of < 0.2 Gy is presented in Figure 4. The spline curve traced the linear-quadratic curve rather than the quadratic curve. An interesting result is that the interval estimates of the spline included zero at doses of < 0.08 Gy, indicating that there was no difference between the male exposed and unexposed group when the doses were < 0.08 Gy, whereas there was no such interval for females. The quadratic curves estimated the ERR to be very small; however, the ERR increased rapidly as the doses increased. This was due to characteristics of a quadratic function. Except for the quadratic curves the spline curves

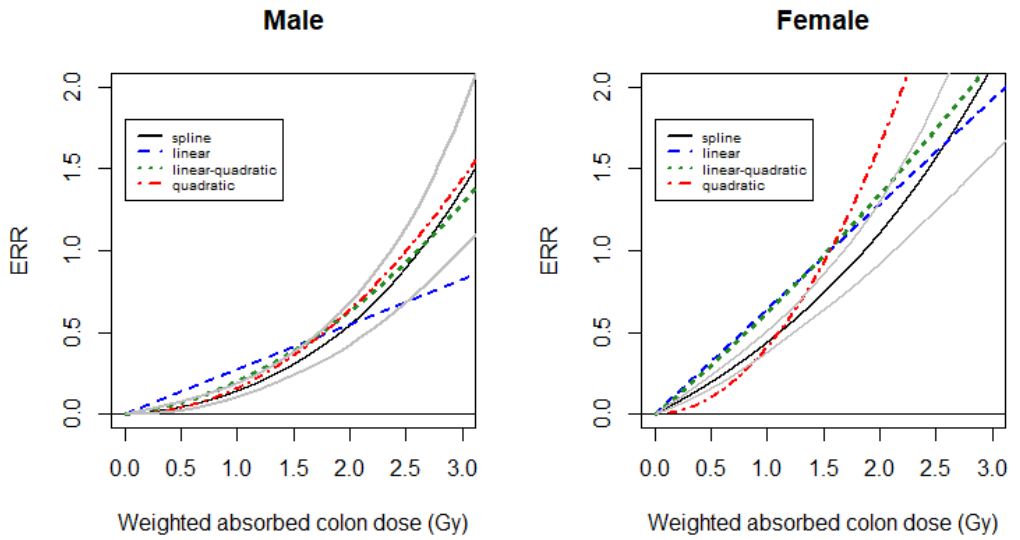

Figure 4: *Dose-response curves at low doses: spline (black, solid), linear (blue, dashed), linear-quadratic (green, dotted), and quadratic (red, dash-dotted). Solid gray lines depict the 95% credible interval of the spline.*

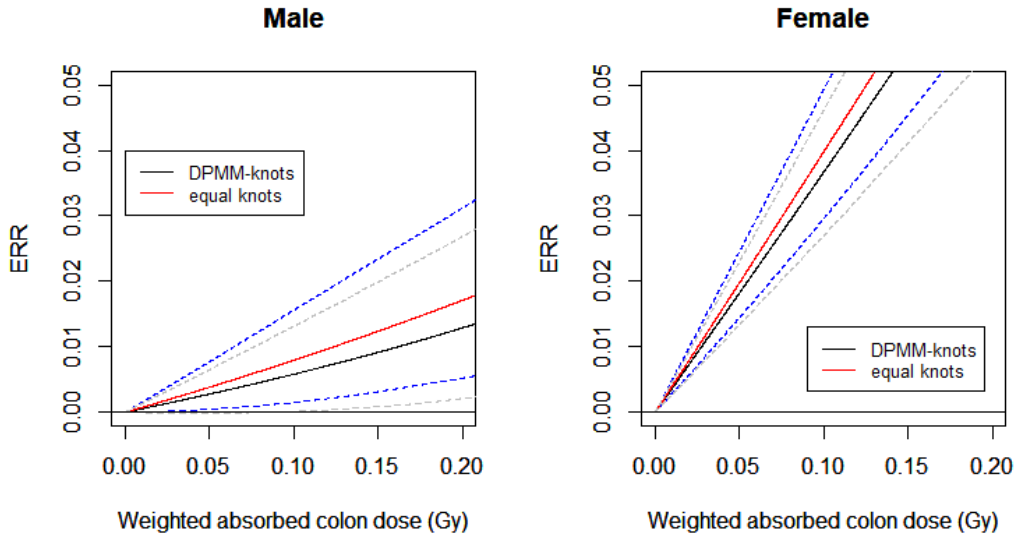

Figure 5: *Dose-response curves for different dose: DPMM-based spline (black, solid) and equally spaced knots-based spline (red, solid). The gray dotted lines represent the 95% credible interval of the solid black curve, and the blue dotted lines represent the 95% credible interval of the red curve.*

also estimated the ERR to be smaller than the other curves at lower doses.

To demonstrate the importance of the knot selection procedure, we also compared two spline curves using DPMM-based knots (black) and equally spaced knots (red) in Figures 5 and 6, respec-

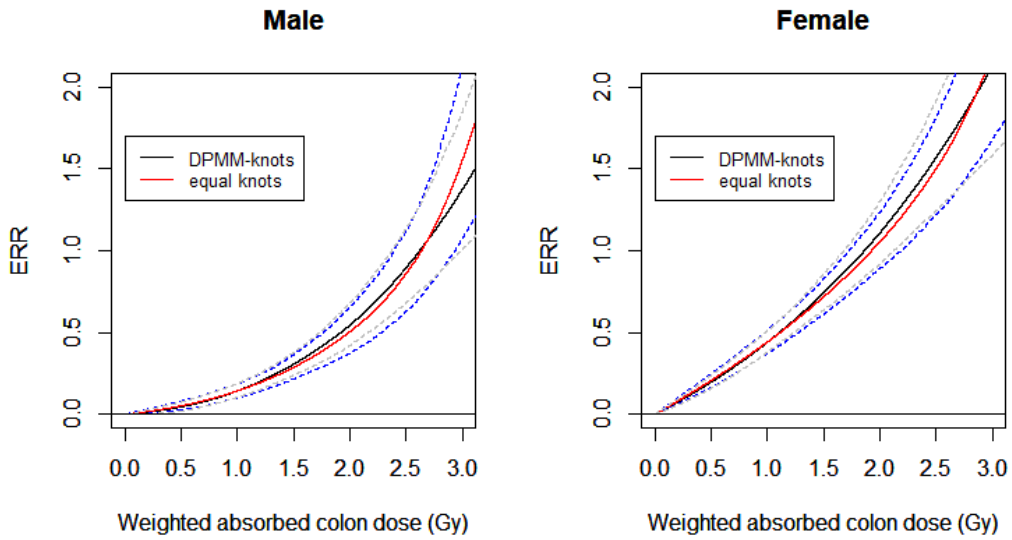

Figure 6: *Dose-response curves at low doses: DPMM-based spline (black, solid) and equally spaced knots-based spline (red, solid). The gray dotted lines represent the 95% credible interval of the solid black curve, and the blue dotted lines represent the 95% credible interval of the red curve.*

tively. In this case, the number of knots remained unchanged, and only the location of the knots was altered. Over the entire dose range, there appeared to be little difference between the two curves. At low doses, however, the red curve estimated the ERR to be smaller than the black curve. The clearest distinction was that the interval estimates of the DPMM-based spline contained zero at doses of < 0.08 Gy, whereas the interval estimates of the spline with equally spaces knots did not include zero over the entire dose range for both males and females.

5. Concluding remarks

In this paper, we proposed spline curve fitting of the dose–response function for the excess relative risk model commonly used in radiation epidemiology. Since the choice of the number and location of knots is crucial in splines, we proposed the Dirichlet process mixture model in selecting data-driven knots, treating knot selection as a clustering problem. When the dose distribution is highly skewed, it is particularly effective due to its flexibility in the number of clusters. The chosen knots appeared to successfully reflect the dose distribution. Then, based on these knots, we estimated the Excess Relative Risk model in the Bayesian framework using the Lifetime Span Study cohort, and compared the spline curve to other parametric dose-response functions. For both males and females, the spline curves estimated the Excess Relative Risk model to be smaller than other curves; however, the Excess Relative Risk model became similar to that of the other curves as the dose increased. The estimation of the dose-response curves was greatly affected by observations at higher doses, but the use of the spline can alleviate this problem. Although we assumed a noninformative prior for the coefficients of the spline, a Gaussian process prior with a covariance matrix depending on the knots can be used as an alternative. The limitation of this research only provided that a Bayesian knot selection is conduced separately from fitting the Bayesian excess relative model.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF-2019R1F1A1061691) and research Grants of Korea Forest Service project (No.2019149A00-2123-0301)

References

Buscot MJ, Wotherspoon SS, Magnussen CG, *et al.* (2017). Bayesian hierarchical piecewise regression models: A tool to detect trajectory divergence between groups in long-term observational studies, *BMC Medical Research Methodology*, **17**, 1–15.

Dahl DB (2006). Model-based clustering for expression data via a Dirichlet process mixture model, *Bayesian Inference for Gene Expression and Proteomics*, **4**, 201–218.

Da Silva ARF (2007). A Dirichlet process mixture model for brain MRI tissue classification, *Medical Image Analysis*, **11**, 169–182.

Dimatteo L, Genovese CR, and Kass RE (2001). Bayesian curve-fitting with free-knot splines, *Biometrika*, **88**, 1055–1071.

Dung VT and Tjahjowidodo T (2017). A direct method to solve optimal knots of B-spline curves: An application for non-uniform B-spline curves fitting, *PLOS one*, **12**.

Escobar MD and West M (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, **90**, 577–588.

Furukawa K, Misumi M, Cologne JB, and Cullings HM (2016). A Bayesian semiparametric model for radiation dose-response estimation, *Risk Analysis*, **36**, 1211–1223.

Görür D and Rasmussen CE (2010). Dirichlet process gaussian mixture models: Choice of the base distribution, *Journal of Computer Science and Technology*, **25**, 653–664.

Grant EJ, Brenner A, Sugiyama H, *et al.* (2017). Solid cancer incidence among the life span study of atomic bomb survivors, *Radiation Research*, **187**, 513–537.

Holmes CC and Mallick BK (2001). Bayesian regression with multivariate linear splines, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **63**, 3–17.

Ishwaran H and James LF (2001). Gibbs sampling methods for stick-breaking priors, *Journal of the American Statistical Association*, **96**, 161–173.

Ishwaran H and James LF (2002). Approximate Dirichlet process computing in finite normal mixtures: smoothing and prior information, *Journal of Computational and Graphical Statistics*, **11**, 508–532.

Kauermann G and Opsomer JD (2011). Data-driven selection of the spline dimension in penalized spline regression, *Biometrika*, **98**, 225–230.

Kern I, John C, and Cohen SM (2005). Menopausal symptom relief with acupuncture: Bayesian analysis using piecewise regression, *Communications in Statistics-Simulation and Computation*, **34**, 783–798.

MacEachern SN and Müller P (1998). Estimating mixture of Dirichlet process models, *Journal of Computational and Graphical Statistics*, **7**, 223–238.

McAuliffe JD, Blei DM, and Jordan MI (2006). Nonparametric empirical Bayes for the Dirichlet process mixture model, *Statistics and Computing*, **16**, 5–14.

Müller P, Erkanli A, and West M (1996). Bayesian curve fitting using multivariate normal mixtures, *Biometrika*, **83**, 67–79.

Ngan HY, Yung NH, and Yeh AG (2015). Outlier detection in traffic data based on the Dirichlet process mixture model, *IET Intelligent Transport Systems*, **9**, 773–781.

NRC (2006). Health risks from exposure to low levels of ionizing radiation: BEIR VII phase 2

(National Research Council and others), *National Academies Press*, **7**.

Orbanz P and Teh YW (2010). Bayesian nonparametric models, *Encyclopedia of Machine Learning*, **1**.

Park Y and Kim Y (2020). Estimation of the excess relative risk using the piecewise linear model with Gaussian process, *Journal of the Korean Data and Information Science Society*, **31**, 1145–1153.

Rasmussen CE (1999). The infinite Gaussian mixture model. In *Proceedings of the 12th International Conference on Neural Information Processing Systems*, **12**, 554–560.

Reich BJ and Bondell HD (2011). A spatial Dirichlet process mixture model for clustering population genetics data, *Biometrics*, **67**, 381–390.

Straub J, Chang J, Freifeld O, Fisher I, and John W (2015). A Dirichlet process mixture model for spherical data, *Artificial Intelligence and Statistics*, 930–938.

Teh YW, Jordan MI, Beal MJ, and Blei DM (2006). Hierarchical Dirichlet processes, *Journal of the American Statistical Association*, **101**, 1566–1581.

Teh YW (2011). Dirichlet process, *In Encyclopedia of Marchine Learning(pp280-897)*, New York, Springer.

Yu G, Huang R, and Wang Z (2010). Document clustering via Dirichlet process mixture model with feature selection. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 763–772.

Received June 21, 2021; Revised August 17, 2021; Accepted October 27, 2021