Journal of the Korean Statistical Society (2006), 35: 3, pp 261–280

PARTIAL INTRINSIC BAYES FACTOR

Y. JOO[1] AND G. CASELLA[2]

ABSTRACT

We have developed a new model selection criteria, the partial intrinsic Bayes factor, which is designed for cases when we select a model among a small number of candidate models. For example, we can choose only a few candidate models after exploring scatter plots. By simulation study, we have showed that PIBF performs better than AIC, BIC and GCV.

AMS 2000 subject classifications. Primary 62F15; Secondary 62A15.
Keywords. Bayes factor, intrinsic Bayes factor, Bayesian model selection.

1. INTRODUCTION

We often select a model from a finite number of candidate models, all of which encompass a small model. For example, suppose that cubic regression spline model (Ruppert *et al.*, 2003) is applied to capture a curvy trend in Figure 1.1. From this scatter plot, analyst knows that the trend may not be explained with linear, quadratic or cubic regressions. A cubic regression spline with 5 equally spaced knots and a natural cubic regression spline (Hastie *et al.*, 2001) with 4 equally spaced knots are applied. In this case, all cubic regression spline models have four common regression parameters corresponding to intercept, linear, quadratic and cubic terms. As an example of regression analysis, some of explanatory variables may be obviously important and need be included in the finally-chosen model, while importance of others should be verified. All candidate models have these important variables commonly. The idea of the Partial Intrinsic Bayes Factor ($\mathrm{BF}^{p\text{-}int}$) is to incorporate this information into a model selection criterion. Most of classical model selection criteria, such as AIC (Akaike, 1973) and BIC (Schwarz, 1978), do not incorporate such knowledge into the selection procedure. Although the prior distribution in a Bayesian framework might be used to employ this information, there is not any clearly suggested objective rule to convert this information into a prior density.

Received June 2006; accepted August 2006.
[1]Corresponding author. Division of Biostatistics, University of Florida, FL 32610, U.S.A (e-mail: yjoo@phhp.ufl.edu)
[2]Department of Statistics, University of Florida, FL 32610, U.S.A

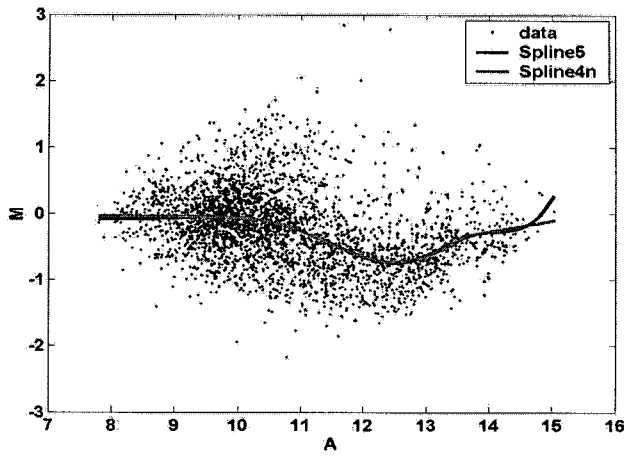

FIGURE 1.1. *M vs. A plot from apo AI microarray experiment, (Dudoit et al., 2-002). The cubic regression spline with 5 equally spaced knots is drawn in a dark line and a natural cubic regression spline with 4 equally spaced knots is in a light line.*

The $\text{BF}^{p\text{-}int}$ divides the parameters into two groups. While the regular intrinsic Bayes factor ($\text{BF}^{intrinsic}$; Berger and Pericchi, 1996) starts with improper noninformative priors for every parameter, the $\text{BF}^{p\text{-}int}$ uses proper informative priors for the parameters that the analyst believes are important and starts with improper noninformative priors, as in $\text{BF}^{intrinsic}$, for the parameters of which importance he/she wants to verify. For example, suppose that parameters θ_A and θ_B are known or assumed to be important and should be included in the finally-chosen model. Also suppose that importance of parameters θ_C and θ_D should be verified. In this case, we may consider four possible candidate models: a model with only θ_A and θ_B, a model with θ_A, θ_B and θ_C, a model with θ_A, θ_B and θ_D, a model with θ_A, θ_B, θ_C and θ_D. Note that all candidate models has θ_A and θ_B commonly because they should be in the finally-chosen model among candidates. To use the $\text{BF}^{p\text{-}int}$, candidate models do not have to have nested structure. However, candidate models should share common explanatory variables (parameters).

Various types of Bayes factors in other researches are summarized in Section 2. General definition of partial intrinsic Bayes factor ($\text{BF}^{p\text{-}int}$) is described in Section 3. Also, $\text{BF}^{p\text{-}int}$ for regression spline model is calculated. Simulation study in Section 4 shows that the $\text{BF}^{p\text{-}int}$ performs better than AIC, BIC, $\text{BF}^{intrinsic}$ and

the generalized cross validation (GCV_ω) for penalized least square estimation (Brumback *et al.*, 1999). Finally, in Section 5, these model selection criteria are applied to the dye bias correction of cDNA microarray data.

2. VARIOUS BAYES FACTORS

In many literatures (Gelfand and Dey, 1994), the Bayes factor is described as

$$\mathrm{BF}_{12}^{std} = \frac{\pi(\theta_1|M_1,y)/\pi(\theta_2|M_2,y)}{\pi(\theta_1|M_1)/\pi(\theta_2|M_2)} = \frac{\int f(y|\theta_1,M_1)\pi(\theta_1|M_1)d\theta_1}{\int f(y|\theta_2,M_2)\pi(\theta_2|M_2)d\theta_2} = \frac{m(y|M_1)}{m(y|M_2)}, \tag{2.1}$$

where $f(y|\theta_i,M_i)$ is the likelihood function of the model M_i that has the parameter, θ_i and $\pi(\theta_i|M_i)$ is the prior of θ_i in the model M_i. Here $f(\cdot)$ and $\pi(\cdot)$ are used as general notations for the probability density function of observation and parameters. When $\pi(M_1) = \pi(M_2) = 1/2$, the posterior odds and the Bayes factor are one-to-one functions. The Bayes factor is closely related to the ratio of maximum likelihoods in a sense that the Bayes factor integrates out θ_i in the likelihood instead of maximizing it. We will call this the standard Bayes factor, BF^{std}. Also, various Bayes factors have been developed using different types of predictive distributions:

$$\mathrm{BF}_{12} = \frac{f(y_{s_{v1}}|y_{s_{c1}},M_1)}{f(y_{s_{v2}}|y_{s_{c2}},M_2)},$$

where

$$\begin{aligned} f(y_{s_{vi}}|y_{s_{ci}},M_i) &= \int f(y_{s_{vi}}|\theta_i,M_i)\pi(\theta_i|y_{s_{ci}},M_i)d\theta_i \\ &= \frac{\int f(y_{s_{vi}}|\theta_i,M_i)f(y_{s_{ci}}|\theta_i,M_i)\pi(\theta_i|M_i)d\theta_i}{\int f(y_{s_{ci}}|\theta_i,M_i)\pi(\theta_i|M_i)d\theta_i} \\ &= \frac{\int f(y_S|\theta_i,M_i)\pi(\theta_i|M_i)d\theta_i}{\int f(y_{s_{ci}}|\theta_i,M_i)\pi(\theta_i|M_i)d\theta_i} \end{aligned}$$

with $S = \{1,2,\ldots,n\} = s_{vi} \cup s_{ci}$, $s_{vi} \subset S$ and $s_{ci} \subset S$. Data label sets S, s_{vi} and s_{ci} are used to indicate that $y_{s_{ci}}$ and $y_{s_{vi}}$ are subsets of the whole data set y_S . We will call s_{vi} the validation set (for checking goodness of fit) and s_{ci} the construction set (for constructing $\pi(\theta_i|y_{s_{ci}},M_i)$). Two construction sets, s_{v1} and s_{v2}, are not necessarily the same. Neither are the two validation sets, s_{c1} and s_{c2}. However, for simplicity of notation, we will use a common notation s_v for s_{v1} and s_{v2}, and s_c for s_{c1} and s_{c2} when the discrimination between them is not

necessary. Gelfand and Dey (1994) summarized various Bayes factors in terms of s_c and s_v as follows.

(i) BF^{std}: $s_v = S$ and $s_c = \{\cdot\}$ yield the standard marginal density and the standard BF in (2.1).

(ii) $\text{BF}^{intrinsic}$: $s_v = S \setminus s_c$ and s_c is a minimal set to make the posterior density proper. This is so called the intrinsic BF and is often used with improper noninformative priors (Berger and Pericchi, 1996).

(iii) $\text{BF}^{O'Hagan}$: $s_v = S \setminus s_c$ and $s_c = \{1, 2, \ldots, [\rho n]\}$, where $0 < \rho < 1$ and $[\,]$ denotes the greatest integer function, are used in Atkinson (1978) and O'Hagan (1991).

(iv) $\text{BF}^{P\&T}$: Peña and Tiao (1992) proposed to use s_v = small set and $s_c = S \setminus s_v$.

(v) BF^{CV}: $s_v = \{r\}$ and $s_c = S \setminus \{r\}$ results in the so-called cross validation Bayes factor, which appeared in Geisser and Eddy (1979). To remove the effect by an arbitrary choice of r, they suggested the pseudo Bayes Factor:

$$\text{BF}^{pseudo} = \frac{m_1}{m_2},$$

where $m_i = \prod_{r=1}^{N} [f(y_{s_v}|y_{s_c}, M_i)]_{s_v=\{r\}}$.

(vi) $\text{BF}^{posterior}$: $s_v = S$ and $s_c = S$ yields Aitkin's (1991) posterior BF. This is similar to the idea of a frequentist likelihood calculation in the sense that it uses all observations to estimate parameters (or distribution of parameters) and evaluate the likelihood based on it.

3. Partial Intrinsic Bayes Factor for Regression Spline Models

3.1. Partial noninformative Bayes factor and partial intrinsic Bayes factor

The partial noninformative Bayes factors are developed for the case when a small model is commonly nested in all candidates. The $\text{BF}^{p\text{-}non}$ divides parameters into two groups, one group $(\theta_i^{(i)})$ for parameters that are included in all

candidate models in common and another ($\theta_i^{(n)}$) for the other parameters. Interest of model selection is finding whether $\theta_i^{(n)}$ should be included in the finally-chosen model or not. Then, the BF$^{p\text{-}non}$ uses a proper informative proper priors $\pi(\theta_i^{(i)}|M_i,\xi)$ for the first group of parameters and estimates the hyper parameters ξ using all observations, and uses a noninformative prior $\pi(\theta_i^{(n)})$ for the parameters in the second group. This is summarized as follows.

DEFINITION 3.1. (Partial noninformative Bayes factor (BF$^{p\text{-}non}$)). *For model M_i, let*

$$\hat{\xi}_i = \arg\max_{\xi_i} f(y_S|M_i,\xi),$$

where

$$f(y_S|M_i,\xi_i) = \int\int f(y_S|\theta_i^{(i)},\theta_i^{(n)},M_i)\pi(\theta_i^{(n)}|M_i)\pi(\theta_i^{(i)}|M_i,\xi_i)d\theta_i^{(n)}d\theta_i^{(i)},$$

$\pi(\theta_i^{(n)}|M_i)$ is a noninformative prior, and $\pi(\theta_i^{(i)}|M_i,\xi)$ is an informative proper distribution with hyper parameter ξ. The partial noninformative Bayes factor is defined as

$$BF_{12}^{p\text{-}non} = \frac{f(y_S|M_1,\hat{\xi}_1)}{f(y_S|M_2,\hat{\xi}_2)}.$$

However, because noninformative improper priors are sensitive to constants that can be multiplied to improper prior, we propose to use intrinsic Bayes factor (Berger and Pericchi, 1996) set-up for the noninformative prior as follows.

DEFINITION 3.2. (Partial intrinsic Bayes factor (BF$^{p\text{-}int}$)). *Let $s_v = S \setminus s_c$ and s_c is a minimal set to make the posterior density of $\theta_i^{(n)}$ proper. The partial intrinsic Bayes factor is defined as*

$$BF_{12}^{p\text{-}int} = \frac{f(y_{s_{v1}}|y_{s_{c1}},M_1)}{f(y_{s_{v2}}|y_{s_{c2}},M_2)}, \tag{3.1}$$

where

$$\begin{aligned} f(y_{s_{vi}}|y_{s_{ci}},M_i) &= \int\int f(y_{s_{vi}}|\theta_i^{(i)},\theta_i^{(n)},M_i)\pi(\theta_i^{(n)}|y_{s_{ci}},M_i)\pi(\theta_i^{(i)}|y_{s_{ci}},M_i,\hat{\xi})d\theta_i \\ &= \frac{\int\int f(y_S|\theta_i^{(i)},\theta_i^{(n)},M_i)\pi(\theta_i^{(n)}|M_i)\pi(\theta_i^{(i)}|M_i,\hat{\xi})d\theta_i^{(n)}d\theta_i^{(i)}}{\int\int f(y_{s_{ci}}|\theta_i^{(i)},\theta_i^{(n)},M_i)\pi(\theta_i^{(n)}|M_i)\pi(\theta_i^{(i)}|M_i,\hat{\xi})d\theta_i^{(n)}d\theta_i^{(i)}} \end{aligned} \tag{3.2}$$

In general, the prior setup of $\text{BF}^{p\text{-}int}$ has the following motivation. By estimating ξ, major portion of probability density in the informative prior of $\theta_i^{(i)}$ tends to be allocated closed to the posterior mode. In posterior probability calculation, information in the likelihood updates both informative prior of $\theta_i^{(i)}$ and noninformative prior of $\theta_i^{(n)}$. Because informative prior of $\theta_i^{(i)}$ is already estimated using the likelihood information, the likelihood makes less contribution in updating the posterior distribution of $\theta_i^{(i)}$ than that of $\theta_i^{(n)}$. In other words, posterior of $\theta_i^{(n)}$ can react more sensitively to the likelihood information in $\text{BF}^{p\text{-}int}$ than in $\text{BF}^{intrinsic}$. Although this paper focuses on comparing the regression models or linear smoothers as candidates, the idea of the $\text{BF}^{p\text{-}int}$ can be generalized to any situation when a group of the same parameters are commonly used in all candidate models.

3.2. Partial intrinsic Bayes factor for the linear regression models

In this section, we calculate partial intrinsic Bayes factor for the linear regression models. Also, the relationship between the partial intrinsic Bayes factor and the intrinsic Bayes factor, which uses improper noninformative priors for all parameters, is shown. Consider regression models

$$y = X_\psi \psi + \epsilon = X_\alpha \alpha + X_\beta \beta + \epsilon, \tag{3.3}$$

where $\alpha = (\alpha_1, \ldots, \alpha_{p_\alpha})^T$, $\beta = (\beta_1, \ldots, \beta_{p_\beta})^T$, X_α and X_β are corresponding design matrix of α and β, $X_\psi = (X_\alpha, X_\beta)$ and $\epsilon \sim N(0, \sigma^2 I)$. For these models, α is the parameter vector, of which all elements are known or assumed to have none zero values and are included in all candidate models. Parameter vector β may have non-zero elements.

3.2.1. Partial intrinsic Bayes factor. The parameters β and σ have improper noninformative priors and $\alpha \sim N(\overline{\alpha}, \Sigma_\alpha)$. For convenience, the notation that indicates the model will be omitted from now on unless it is necessary. For example, $\theta = \theta_i$ and $f(y|\theta) = f(y|\theta_i, M_i)$. Also, suppose $\overline{\alpha}$ and Σ_α are given for now. If $\int f(y_{s_v}|\theta)\pi(\theta)d\theta$ and $\int f(y_S|\theta)\pi(\theta)d\theta$ are calculated, we can get the intrinsic Bayes factor from equations (3.1) and (3.2). Mathematically, logics in the calculations of $\int f(y_{s_v}|\theta)\pi(\theta)d\theta$ and $\int f(y_S|\theta)\pi(\theta)d\theta$ are equivalent. Therefore, we will focus on the general calculation of $\int f(y|\theta)\pi(\theta)d\theta$ for now. For the calculation of $\int f(y|\theta)\pi(\theta)d\theta$, we will use Bayes theorem,

$$\int f(y|\theta)\pi(\theta)d\theta = \frac{f(y|\theta)\pi(\theta)}{\pi(\theta|y)}. \tag{3.4}$$

This model has $\theta = \{\alpha, \beta, \sigma\}$. Assume $\pi(\beta|\sigma) = 1$, and $\alpha|\sigma \sim N(\overline{\alpha}, \Sigma_\alpha)$. The overline of hyper parameter, *i.e.* $\overline{\alpha}$, is used to indicate a parameter in the prior distribution. We will use the double overlines for the parameters in the posterior distribution, *i.e.* $\overline{\overline{\alpha}}$. For convenience of calculation, set $A = \Sigma_\alpha^{-1}\sigma^2$. Then

$$\begin{aligned}\pi(\alpha|\sigma) &= (2\pi)^{-p_\alpha/2}|\Sigma_\alpha|^{-1/2}\exp\{-(\alpha-\overline{\alpha})^T\Sigma_\alpha^{-1}(\alpha-\overline{\alpha})/2\}\\ &\propto \sigma^{-p_\alpha}\exp\{-(\alpha-\overline{\alpha})^T A(\alpha-\overline{\alpha})/2\sigma^2\}\\ &= \sigma^{-p_\alpha}\exp\{-(\psi-\overline{\psi})^T B(\psi-\overline{\psi})/2\sigma^2\},\end{aligned}$$

where

$$B = \begin{pmatrix} A & 0_{p_\alpha p_\beta} \\ 0_{p_\beta p_\alpha} & 0_{p_\beta p_\beta}\end{pmatrix}$$

and 0_{ij} is the $i \times j$ matrix of which all elements are zeros. The hyper parameter, $\overline{\beta}$ in $\overline{\psi} = (\overline{\alpha}^T, \overline{\beta}^T)^T$, can be any vector of finite constants because it will disappear when it is multiplied with $0_{p_\beta p_\beta}$. Finally, assume $\pi(\sigma) = \sigma^{-1}$.

Hence, with the assumption of the independence between $\alpha|\sigma$ and $\beta|\sigma$,

$$\pi(\psi, \sigma) = \pi(\alpha|\sigma)\pi(\beta|\sigma)\pi(\sigma) \propto \sigma^{-p_\alpha - 1}\exp\{-(\psi-\overline{\psi})^T B(\psi-\overline{\psi})/2\sigma^2\}.$$

Also,

$$\begin{aligned}f(y|\psi,\sigma) &= (2\pi)^{-n/2}\sigma^{-n}\exp\{-(y-X\psi)^T(y-X\psi)/2\sigma^2\}\\ &\propto \sigma^{-n}\exp\{-(y-X\psi)^T(y-X\psi)/2\sigma^2\}.\end{aligned}$$

Hence,

$$\begin{aligned}&\pi(\psi,\sigma|y) \propto f(y|\psi,\sigma)\pi(\psi,\sigma)\\ &= (2\pi)^{-n/2}\sigma^{-n}\exp\{-(y-X\psi)^T(y-X\psi)/2\sigma^2\}\\ &\quad\times(2\pi)^{-p_\alpha/2}\sigma^{-p_\alpha-1}|A|^{1/2}\exp\{-(\alpha-\overline{\alpha})^T A(\alpha-\overline{\alpha})/2\sigma^2\}. \qquad (3.5)\end{aligned}$$

Equation (3.5) is the numerator in equation (3.4). Also,

$$\begin{aligned}\pi(\psi,\sigma|y) \propto \sigma^{-n-p_\alpha-1}\exp\{&-((y-X\psi)^T(y-X\psi)\\ &+(\psi-\overline{\psi})^T B(\psi-\overline{\psi}))/2\sigma^2\}. \qquad (3.6)\end{aligned}$$

For calculation of the denominator in equation (3.4), the posterior distribution, we will use equation (3.6) as follows. The posterior calculation becomes similar to the case when a multivariate normal distribution is assigned for all elements

in ψ and $\pi(\sigma) = \sigma^{-1}$. Also, $\pi(\sigma) = \sigma^{-1}$ can be considered as a special case of the kernel in the inverted gamma-2 ($IG2$) probability dentisy:

$$\begin{aligned}
\pi\left(\sigma^{-2}\right) &= \frac{\bar{b}^{\bar{a}}}{\Gamma(\bar{a})}(\sigma^{-2})^{\bar{a}-1}e^{-\bar{b}(\sigma^{-2})} : \text{Gamma}(\bar{a},\bar{b}) \\
\Longleftrightarrow \pi(\sigma) &= \frac{2\bar{b}^{\bar{a}}}{\Gamma(\bar{a})}\sigma^{-2\bar{a}-1}e^{-\bar{b}\sigma^{-2}} : IG2(\bar{a},\bar{b}) \\
&\propto \sigma^{-2\bar{a}-1}e^{-\bar{b}\sigma^{-2}} : \text{the kernel of } IG2(\bar{a},\bar{b}). \qquad (3.7)
\end{aligned}$$

If $\bar{a} = \bar{b} = 0$, then the kernel becomes σ^{-1}. Though the parameters in the gamma or the $IG2$ distributions should be greater than 0, we can still treat this as a special case of the kernel of the $IG2$ distribution (3.7) because kernels do not have the same restrictions as probability density functions.

Then, we can use the calculations in other studies (Judge *et al.*, 1988, pp. 306–311) for the proper informative prior setup, in which ψ has a normal distribution and σ has a inverted gamma-2 distribution for this problem. Finally, the BF^{p_int} is calculated as (see Appendix for detailed calculation):

$$\begin{aligned}
\text{BF}_{12}^{p_int} &= \frac{f(y_{s_{v1}}|y_{s_{c1}}, M_1)}{f(y_{s_{v2}}|y_{s_{c2}}, M_2)} \\
&= (2\pi)^{\{C(s_{c1})-C(s_{c2})\}/2}\left(\frac{|B_1 + X_{s_{c1}}^{[1]T}X_{s_{c1}}^{[1]}|/|B_2 + X_{s_{c2}}^{[2]T}X_{s_{c2}}^{[2]}|}{|B_1 + X_S^{[1]T}X_S^{[1]}|/|B_2 + X_S^{[2]T}X_S^{[2]}|}\right)^{1/2} \\
&\quad \times \frac{\Gamma\left((n-p_{\beta\,1})/2\right)}{\Gamma\left((n-p_{\beta\,2})/2\right)}\frac{\Gamma\left(\{C(s_{c2})-p_{\beta\,2}\}/2\right)}{\Gamma\left(\{C(s_{c1})-p_{\beta\,1}\}/2\right)}\frac{\bar{\bar{b}}_{s_{c1}}^{[1]\{C(s_{c1})-p_{\beta\,1}\}/2}}{\bar{\bar{b}}_{s_{c2}}^{[2]\{C(s_{c2})-p_{\beta\,2}\}/2}}\frac{\bar{\bar{b}}_S^{[2](n-p_{\beta\,2})/2}}{\bar{\bar{b}}_S^{[1](n-p_{\beta\,1})/2}}, \qquad (3.8)
\end{aligned}$$

where superscript $^{[j]}$ indicated model M_j and

$$\begin{aligned}
\bar{\bar{b}} &= \frac{(y - X\bar{\bar{\psi}})^T(y - X\bar{\bar{\psi}}) + (\bar{\psi} - \bar{\bar{\psi}})^T B(\bar{\psi} - \bar{\bar{\psi}})}{2} \\
&= \frac{y^Ty + \bar{\psi}^TB\bar{\psi} - (B\bar{\psi} + X^TX\hat{\psi})(B + X^TX)^{-1}(B\bar{\psi} + X^TX\hat{\psi})}{2}.
\end{aligned}$$

In this paper, we adopted the idea of arithmetic mean IBF (Berger and Pericchi, 1996) to remove dependence of s_c. The construction set, s_c, is randomly and repeatedly selected from S. Then, BF^{p_int} is calculated for each s_c. The average of BF^{p_int} is used as an model selection criterion in actual applications.

3.2.2. Intrinsic Bayes factor. All parameters (α, β and σ) have improper noninformative priors. Let the prior for the parameters for model j be

$$\pi\left(\psi^{[j]}, \sigma^{[j]} \middle| M_j\right) = 1/\sigma^{[j]}.$$

Then, assuming that $s_{v1} = s_{v2} = s_v$ and $s_{c1} = s_{c2} = s_c$, the intrinsic Bayes factor is (Berger and Pericchi, 1996)

$$\begin{aligned} \mathrm{BF}_{12}^{intrinsic} = & \frac{|X_S^{[2]T} X_S^{[2]}|^{1/2}}{|X_S^{[1]T} X_S^{[1]}|^{1/2}} \frac{|y_S - X_S^{[2]}\hat{\psi}_S^{[2]}|^{n-p_{\alpha 2}-p_{\beta 2}}}{|y_S - X_S^{[1]}\hat{\psi}_S^{[1]}|^{n-p_{\alpha 1}-p_{\beta 1}}} \frac{\Gamma\big((n-p_{\alpha 1}-p_{\beta 1})/2\big)}{\Gamma\big((n-p_{\alpha 2}-p_{\beta 2})/2\big)} \\ & \times \frac{\Gamma\big((C(s_c)-p_{\alpha 2}-p_{\beta 2})/2\big)}{\Gamma\big((C(s_c)-p_{\alpha 1}-p_{\beta 1})/2\big)} \frac{|X_{s_c}^{[1]T} X_{s_c}^{[1]}|^{1/2}}{|X_{s_c}^{[2]T} X_{s_c}^{[2]}|^{1/2}} \\ & \times \frac{|y_{s_c} - X_{s_c}^{[1]}\hat{\psi}_{s_c}^{[1]}|^{C(s_c)-p_{\alpha 1}-p_{\beta 1}}}{|y_{s_c} - X_{s_c}^{[2]}\hat{\psi}_{s_c}^{[2]}|^{C(s_c)-p_{\alpha 2}-p_{\beta 2}}}. \end{aligned} \tag{3.9}$$

The intrinsic Bayes factor in equation (3.9), $\mathrm{BF}_{12}^{intrinsic}$, can be obtained using equation (3.8) by setting $s_{v1} = s_{v2} = s_v$, $s_{c1} = s_{c2} = s_c$, $B_j = 0_{p_{\beta j} p_{\beta j}}$ and $p_{\alpha j} = 0$. The $\mathrm{BF}^{intrinsic}$ is a special case of the $\mathrm{BF}^{p\text{-}int}$.

3.3. Estimation of informative prior using empirical Bayes Gibbs sampling

In this subsection, we demonstrate the estimation of Σ_α using empirical Bayes Gibbs sampling. To make the model selection criterion conservative, we set $\overline{\alpha}$=0. Suppose $A = \lambda^{-2} I_{p_\alpha p_\alpha}$, where $I_{p_\alpha p_\alpha}$ is the $p_\alpha \times p_\alpha$ identity matrix. Equivalently, $\Sigma_\alpha = \lambda^2 \sigma^2 I_{p_\alpha p_\alpha}$. Remind that σ^2 is the variance of ϵ and is included in definition of Σ_α for calculational convenience of $\mathrm{BF}^{p\text{-}int}$. To estimate λ, we propose to use empirical Bayes Gibbs sampling (Casella, 2001), which is basically an EM algorithm in Bayesian framework. Detailed calculation for this algorithm is in Appendix.

4. SIMULATION STUDY

We conducted simulation studies to evaluate the performance of the $\mathrm{BF}^{p\text{-}int}$ with sample size 50 and 100, considering a cubic regression model and cubic regression spline models with 1, 2 and 3 (equally spaced) knots as candidate models. We used the setup of priors described in Section 3.2. In this case, four regression parameters (intercept and three parameters corresponding to the first, second and third order terms) in the cubic regression are the commonly

used parameters, α in (3.3), in all candidate models. These parameters will be included in the finally chosen model. Parameters corresponding to each knot are the parameters, β in (3.3), to be verified.

Using simulated data sets from each of the candidates, we estimated and recorded the probability of choosing each candidate model in Tables 4.1, 4.2, 4.3 and 4.4. The specific model that is used as the true model in each simulation study is described in the title of each table. For example, when 50 (=n) observations are simulated from a cubic regression model, $y = 1 + x + x^2 + x^3 + N(0, 2^2)$ and $X \in [0, 2]$, the $\text{BF}^{p\text{-}int}$ chooses a cubic regression model without any knot with probability 0.864 in 5,000 simulation runs (See Table 4.1). The probability of choosing the correct model is typed in bold in each table. The $\text{BF}^{intrinsic}$, AIC, BIC, GCV_0 and GCV_ω (Brumback *et al.*, 1999) are considered as competitors of the $\text{BF}^{p\text{-}int}$. Model selection criteria are called conservative when they tends to choose a smaller candidate model. AIC and BIC are well-known anti-conservative model and conservative selectors, respectively. Most of other popular model selectors are somewhere between AIC and BIC in terms of conservativeness. GCV_ω is the generalized cross validation with a smoothing parameter ω. Brumback *et al.* (1999) suggested to use GCV_ω as a model selection criterion for regression spline models. Even though our interests is knot selection rather than smoothing parameter estimation, we considered GCV_ω in simulation study because it is widely used. GCV_0 is GCV_ω with $\omega = 0$. In other words, GCV_0 is the GCV that we use for regular regression models without any smoothing parameter.

When data come from a cubic regression model (Table 4.1), the $\text{BF}^{p\text{-}int}$ does not perform best. But it performs almost as good as the best performing model selection criteria, $\text{BF}^{intrinsic}$ and BIC, and does much better than AIC, GCV_0 and GCV_ω. Because BIC has a big penalty term for overfitting, it has a strong tendency to choose a smaller model. $\text{BF}^{intrinsic}$ is asymptotically equivalent to BIC (Joo, 2003). Because the smallest candidate (cubic regression model) is the true model, BIC and other asymptotically equivalent model selectors are expected to perform well. When the data come from a cubic regression spline model with 1 knot (Table 4.2), the $\text{BF}^{p\text{-}int}$ performs best. Particularly, when n=50 (n is small), the $\text{BF}^{p\text{-}int}$ performed much better than others. The $\text{BF}^{p\text{-}int}$ chooses the correct model 85.5% of cases, but the others choose $40 - 65\%$. When data come from a cubic regression spline model with 2 knots (Table 4.3), the $\text{BF}^{p\text{-}int}$ dominates all other criteria. When the data come from a cubic regression spline model with 3 knots (Table 4.4), the $\text{BF}^{p\text{-}int}$ performs best along with AIC and GCV_ω. Because AIC has a small penalty term, it has strong tendency to choose a bigger

model. GCV_0 is asymptotically equivalent to AIC. In this case, the biggest candidate (cubic regression model) is the true model. Therefore, AIC and other asymptotically equivalent model selectors are expected to perform well.

TABLE 4.1. *Probability of selecting the model when the data is simulated from a cubic regression model:* $y = 1 + x + x^2 + x^3 + N(0, 2^2)$ *and* $X \in [0, 2]$

n	$BF^{p\text{-}int}$				$BF^{intrinsic}$			
	0	1	2	3	0	1	2	3
50	**0.864**	0.084	0.032	0.020	**0.910**	0.064	0.022	0.004
100	**0.879**	0.095	0.017	0.009	**0.938**	0.051	0.011	0.000
n	BIC				AIC			
	0	1	2	3	0	1	2	3
50	**0.918**	0.051	0.023	0.008	**0.712**	0.135	0.077	0.076
100	**0.957**	0.035	0.007	0.001	**0.725**	0.127	0.088	0.060
n	GCV_0				GCV_ω			
	0	1	2	3	0	1	2	3
50	**0.750**	0.128	0.066	0.056	**0.604**	0.134	0.105	0.157
100	**0.747**	0.117	0.082	0.054	**0.625**	0.133	0.121	0.121

TABLE 4.2. *Probability of selecting the model when the data is simulated from a cubic regression spline model with* **1** *knot:* $y = 1 + x + x^2 + x^3 - 15(x-1)_+^3 + N(0, 2^2)$ *and* $X \in [0, 2]$

n	$BF^{p\text{-}int}$				$BF^{intrinsic}$			
# of knots	0	1	2	3	0	1	2	3
50	0.058	**0.855**	0.063	0.024	0.321	**0.614**	0.049	0.016
100	0.009	**0.907**	0.069	0.015	0.072	**0.879**	0.044	0.005
n	BIC				AIC			
# of knots	0	1	2	3	0	1	2	3
50	0.331	**0.598**	0.046	0.025	0.133	**0.615**	0.139	0.113
100	0.096	**0.854**	0.042	0.008	0.009	**0.732**	0.158	0.101
n	GCV_0				GCV_ω			
# of knots	0	1	2	3	0	1	2	3
50	0.155	**0.632**	0.125	0.088	0.069	**0.415**	0.243	0.273
100	0.010	**0.747**	0.156	0.087	0.004	**0.424**	0.273	0.299

In summary, the $BF^{p\text{-}int}$ performed as well as BIC when the smallest model is true, did as well as AIC when the biggest model is true, and did better than all other model selectors in other cases.

TABLE 4.3. *Probability of selecting the model when the data is simulated from a cubic regression spline model with* **2** *knots:* $y = 1 + x + x^2 + x^3 - 10(x - 1/3)_+^3 + 5(x - 2/3)_+^3 + N(0, 0.5^2)$ *and* $X \in [0, 2]$

n	$BF^{p\text{-}int}$				$BF^{intrinsic}$			
# of knots	0	1	2	3	0	1	2	3
50	0.307	0.150	**0.498**	0.045	0.717	0.162	**0.110**	0.011
100	0.186	0.102	**0.664**	0.048	0.572	0.208	**0.210**	0.010
n	BIC				AIC			
# of knots	0	1	2	3	0	1	2	3
50	0.742	0.139	**0.102**	0.017	0.427	0.205	**0.250**	0.118
100	0.632	0.163	**0.189**	0.016	0.240	0.194	**0.430**	0.136
n	GCV_0				GCV_ω			
# of knots	0	1	2	3	0	1	2	3
50	0.479	0.203	**0.233**	0.085	0.403	0.074	**0.174**	0.349
100	0.253	0.206	**0.419**	0.122	0.236	0.058	**0.259**	0.447

TABLE 4.4. *Probability of selecting the model when the data is simulated from a cubic regression spline model with* **3** *knots:* $y = 1 + x + x^2 + x^3 - 15(x - 0.5)_+^3 + 50(x - 1.0)_+^3 - 50(x - 1.5)_+^3 + N(0, 0.5^2)$ *and* $X \in [0, 2]$

n	$BF^{p\text{-}int}$				$BF^{intrinsic}$			
# of knots	0	1	2	3	0	1	2	3
50	0.000	0.157	0.004	**0.839**	0.000	0.455	0.016	**0.529**
100	0.000	0.014	0.001	**0.985**	0.000	0.119	0.002	**0.879**
n	BIC				AIC			
# of knots	0	1	2	3	0	1	2	3
50	0.000	0.400	0.010	**0.590**	0.000	0.147	0.012	**0.841**
100	0.000	0.108	0.002	**0.890**	0.000	0.010	0.001	**0.989**
n	GCV_0				GCV_ω			
# of knots	0	1	2	3	0	1	2	3
50	0.000	0.182	0.012	**0.806**	0.000	0.141	0.020	**0.839**
100	0.000	0.014	0.001	**0.985**	0.000	0.058	0.005	**0.937**

5. EXAMPLE: CORRECTION OF DYE BIAS IN MICROARRAY DATA

In cDNA Microarray analyses, the experimenter prepares two tissue samples of interest and applies fluorescent green (Cy3) and red (Cy5) dyes, which get activated only when mRNA is bound with complimentary DNA. After hybridizing mRNA of two tissue samples, the print tip (or printing machine) delivers

hybridized mRNA into each well. If the gene in a well has a complimentary DNA for mRNA in a sample, the corresponding dye will get activated. The intensity of fluorescence indicates the abundance of DNA's that the tissue sample has. Intensity of both dye responses increase monotonically with the abundance of DNA. However, these responses are not identical. This is called dye bias. Dudoit *et al.* (2002) suggested removing the dye bias by capturing the trend on the plot of M $(= \log(R) - \log(G))$ *vs.* A $(= \{\log(R) + \log(G)\}/2)$, where R is the red fluorescence intensity from pooled tissue samples of control group individuals and G is the green fluorescence intensity from a control group individual. If the responses of two dyes are identical, there should not be any trend on M *vs.* A plot. Dudoit *et al.* (2002) proposes to estimate the trend in the mean function using nonparametric regression.

The apo AI experiment used eight normal C57B1/6 mice for the control group (mouse id=1,2,...,8) and another eight mice with the apo AI knocked-out for the treatment group (mouse id=1,2,...,8). Model selection criteria are applied to choose a proper model that captures the trend of M *vs.* A plot of these data (Figure 1.1). As candidate models, we considered regular cubic regression spline models with equally-spaced 1-6 knots, and natural cubic regression spline models with equally-spaced 2-6 knots, and a cubic regression model. Among 12 candidate models, BIC, $\text{BF}^{intrinsic}$ and BF^{p_int} support the natural cubic regression spline model with 4 knots and AIC, GCV_0 and GCV_ω support the cubic regression spline model with 5 knots. It seems that the natural spline model performed a little better in the right edge of the data (see where $A \in (14, 15)$ in Figure 1.1). However, there is not a big difference among these models in a practical sense. Model selection criteria were sensitive for this case, because the data has a large number (3192) of observations.

6. CONCLUDING REMARK

We developed the partial intrinsic Bayes factor to select the best model when all candidate models have common parameters. Knot selection in cubic regression splines often belongs to this case because four regression parameters in the cubic regression part are usually included in all candidate models. Using simulation study, it is demonstrated that the partial intrinsic Bayes factor performs better than others. Although main ideas in this method can be widely applied in many types of regression analyses, computational difficulty can be an disadvantage to apply it to any model beyond the normal linear regression.

ACKNOWLEDGEMENTS

The authors thanks to the referee's useful comments that made the paper more readable.

APPENDIX

Calculation of partial intrinsic Bayes factor

By Judge *et al.* (1988),

$$\begin{aligned}&(y - X\psi)^T(y - X\psi) + (\psi - \overline{\psi})^T B(\psi - \overline{\psi}) \\ &= \left\{(\psi - \overline{\overline{\psi}})^T(B + X^TX)(\psi - \overline{\overline{\psi}})\right\} \\ &\quad + \left\{(y - X\overline{\overline{\psi}})^T(y - X\overline{\overline{\psi}}) + (\overline{\psi} - \overline{\overline{\psi}})^T B(\overline{\psi} - \overline{\overline{\psi}})\right\},\end{aligned}$$

where $\overline{\overline{\psi}} = (B + X^TX)^{-1}(B\overline{\psi} + X^TX\hat{\psi})$ and $\hat{\psi} = (X^TX)^{-1}X^Ty$. Though B does not have full rank, $B + X^TX$ is positive definite because B is non-negative definite and X^TX is positive definite. Hence, the inverse of $(B + X^TX)$ exists. Therefore, equation (3.5) can be written

$$\begin{aligned}&\pi(\psi, \sigma|y) \\ &\propto \sigma^{-n-p_\alpha-1} \exp\left[-\left\{(y - X\psi)^T(y - X\psi) + (\psi - \overline{\psi})^T B(\psi - \overline{\psi})\right\}/2\sigma^2\right] \\ &\propto \sigma^{-p_\beta-p_\alpha} \exp\left\{-\frac{(\psi - \overline{\overline{\psi}})^T(B + X^TX)(\psi - \overline{\overline{\psi}})}{2\sigma^2}\right\} \\ &\quad \times\sigma^{-n+p_\beta-1} \exp\left\{-\frac{(y - X\overline{\overline{\psi}})^T(y - X\overline{\overline{\psi}}) + (\overline{\psi} - \overline{\overline{\psi}})^T B(\overline{\psi} - \overline{\overline{\psi}})}{2\sigma^2}\right\} \\ &\propto \pi(\psi|\sigma, y) \cdot \pi(\sigma|y),\end{aligned}$$

where

$$\psi|\sigma, y \sim N\left(\overline{\overline{\psi}}, \sigma^2(B + X^TX)^{-1}\right),$$

$$\sigma|y \sim IG2\left(\overline{\overline{a}}, \overline{\overline{b}}\right),$$

$$\overline{\overline{a}} = \frac{n - p_\beta}{2}$$

and

$$\begin{aligned}\bar{\bar{b}} &= \frac{(y - X\bar{\bar{\psi}})^T(y - X\bar{\bar{\psi}}) + (\bar{\psi} - \bar{\bar{\psi}})^T B(\bar{\psi} - \bar{\bar{\psi}})}{2} \\ &= \frac{y^T y + \bar{\psi}^T B\bar{\psi} - \bar{\bar{\psi}}^T(B + X^T X)\bar{\bar{\psi}}}{2} \\ &= \frac{y^T y + \bar{\psi}^T B\bar{\psi} - (B\bar{\psi} + X^T X\hat{\psi})(B + X^T X)^{-1}(B\bar{\psi} + X^T X\hat{\psi})}{2}.\end{aligned}$$

In other words,

$$\begin{aligned}\pi(\psi, \sigma|y) &= \pi(\psi|\sigma, y)\pi(\sigma|y) \\ &= (2\pi)^{-(p_\beta + p_\alpha)/2}\sigma^{-(p_\beta + p_\alpha)}\sqrt{|(B + X^T X)|} \\ &\quad \times \exp\{-(\psi - \bar{\bar{\psi}})^T(B + X^T X)(\psi - \bar{\bar{\psi}})/2\sigma^2\} \\ &\quad \times \frac{2\bar{\bar{b}}^{(n-p_\beta)/2}}{\Gamma\big((n - p_\beta)/2\big)}\sigma^{-n+p_\beta-1}e^{-\bar{\bar{b}}\sigma^{-2}},\end{aligned}$$

The marginal distribution becomes

$$\begin{aligned}&\int f(y|\alpha, \beta, \sigma)\pi(\alpha, \beta, \sigma)\, d(\alpha, \beta, \sigma) \\ &= \frac{f(y|\alpha, \beta, \sigma)\pi(\alpha, \beta, \sigma)}{\pi(\alpha, \beta, \sigma|y)} \\ &= \Big[(2\pi)^{-n/2}\sigma^{-n}\exp\{-(y - X\psi)^T(y - X\psi)/2\sigma^2\} \\ &\quad \times (2\pi)^{-p_\alpha/2}\sigma^{-p_\alpha}|A|^{1/2}\exp\{-(\psi - \bar{\psi})^T B(\psi - \bar{\psi})/2\sigma^2\}\sigma^{-1}\Big] \\ &\quad \div \Bigg[(2\pi)^{-(p_\beta + p_\alpha)/2}\sigma^{-(p_\beta + p_\alpha)}|B + X^T X|^{1/2} \\ &\quad \times \exp\{-(\psi - \bar{\bar{\psi}})^T(B + X^T X)(\psi - \bar{\bar{\psi}})/2\sigma^2\} \\ &\quad \times \frac{2\bar{\bar{b}}^{(n-p_\beta)/2}}{\Gamma\big((n - p_\beta)/2\big)}\sigma^{-n+p_\beta-1}e^{-\bar{\bar{b}}\sigma^{-2}}\Bigg] \\ &= \Big[(2\pi)^{-n/2} \times (2\pi)^{-p_\alpha/2}|A|^{1/2}\Big] \\ &\quad \div \left[(2\pi)^{-(p_\beta + p_\alpha)/2}|B + X^T X|^{1/2}\frac{2\bar{\bar{b}}^{(n-p_\beta)/2}}{\Gamma\big((n - p_\beta)/2\big)}\right] \\ &= (2\pi)^{-(n-p_\beta)/2}\left(|A|/|B + X^T X|\right)^{1/2}\frac{\Gamma\big((n - p_\beta)/2\big)}{2\bar{\bar{b}}^{(n-p_\beta)/2}}.\end{aligned}$$

Naturally, define the expressions, B_j, r_j, $p_{\alpha j}$, $X_s^{[j]}$ and $\bar{\bar{b}}_s^{[j]}$, to show the depen dence on the model j and the set s. Equation (3.2) becomes

$$
\begin{aligned}
f(y_{s_{vj}}|y_{s_{cj}}, M_j) &= \frac{\int f(y_{s_{vj}}, y_{s_{cj}}|\alpha,\beta,\sigma, M_j)\pi(\alpha,\beta,\sigma|M_j)\, d(\alpha,\beta,\sigma|M_j)}{\int f(y_{s_{cj}}|\alpha,\beta,\sigma, M_j)\pi(\alpha,\beta,\sigma|M_j)\, d(\alpha,\beta,\sigma|M_j)} \\
&= \frac{(2\pi)^{-(n-p_{\beta j})/2}\left(|A|/|B_j + X_S^{[j]T}X_{jS}^{[j]}|\right)^{1/2}\frac{\Gamma(\{n-p_{\beta j}\}/2)}{2\bar{\bar{b}}_S^{[j](n-p_{\beta j})/2}}}{(2\pi)^{-\{C(s_{cj})-p_{\beta j}\}/2}\left(|A|/|B_j + X_{s_{cj}}^{[j]T}X_{s_{cj}}^{[j]}|\right)^{1/2}\frac{\Gamma(\{C(s_{cj})-p_{\beta j}\}/2)}{2\bar{\bar{b}}_{s_{cj}}^{[j]\{C(s_{cj})-p_{\beta j}\}/2}}} \\
&= (2\pi)^{-\{n-C(s_{cj})\}/2}\left(|B_j + X_{s_{cj}}^{[j]T}X_{s_{cj}}^{[j]}|/|B_j + X_S^{[j]T}X_S^{[j]}|\right)^{1/2} \\
&\quad \times \frac{\Gamma\big((n-p_{\beta j})/2\big)}{\Gamma\big(\{C(s_{cj})-p_{\beta j}\}/2\big)}\frac{\bar{\bar{b}}_{s_{cj}}^{[j]\{C(s_{cj})-p_{\beta j}\}/2}}{\bar{\bar{b}}_S^{[j](n-p_{\beta j})/2}},
\end{aligned}
$$

where $C(s)$ is the size of set s. Hence, the $\mathrm{BF}^{p\text{-}int}$ is

$$
\begin{aligned}
\mathrm{BF}_{12}^{p_{intrinsic}} &= \frac{f(y_{s_{v1}}|y_{s_{c1}}, M_1)}{f(y_{s_{v2}}|y_{s_{c2}}, M_2)} \\
&= \left\{(2\pi)^{-\{n-C(s_{c1})\}/2}\left(|B_1 + X_{s_{c1}}^{[1]T}X_{s_{c1}}^{[1]}|/|B_1 + X_S^{[1]T}X_S^{[1]}|\right)^{1/2}\right. \\
&\quad \left.\times \frac{\Gamma\big((n-p_{\beta 1})/2\big)}{\Gamma\big(\{C(s_{c1})-p_{\beta 1}\}/2\big)}\frac{\bar{\bar{b}}_{s_{c1}}^{[1]\{C(s_{c1})-p_{\beta 1}\}/2}}{\bar{\bar{b}}_S^{[1](n-p_{\beta 1})/2}}\right\} \\
&\quad \div\left\{(2\pi)^{-\{n-C(s_{c2})\}/2}\left(|B_2 + X_{s_{c2}}^{[2]T}X_{s_{c2}}^{[2]}|/|B_2 + X_S^{[2]T}X_S^{[2]}|\right)^{1/2}\right. \\
&\quad \left.\times \frac{\Gamma\big((n-p_{\beta 2})/2\big)}{\Gamma\big((C(s_{c2})-p_{\beta 2})/2\big)}\frac{\bar{\bar{b}}_{s_{c2}}^{[2]\{C(s_{c2})-p_{\beta 2}\}/2}}{\bar{\bar{b}}_S^{[2](n-p_{\beta 2})/2}}\right\} \\
&= (2\pi)^{\{C(s_{c1})-C(s_{c2})\}/2}\left(\frac{|B_1 + X_{s_{c1}}^{[1]T}X_{s_{c1}}^{[1]}|/|B_2 + X_{s_{c2}}^{[2]T}X_{s_{c2}}^{[2]}|}{|B_1 + X_S^{[1]T}X_S^{[1]}|/|B_2 + X_S^{[2]T}X_S^{[2]}|}\right)^{1/2} \\
&\quad \times \frac{\Gamma\big((n-p_{\beta 1})/2\big)}{\Gamma\big((n-p_{\beta 2})/2\big)}\frac{\Gamma\big(\{C(s_{c2})-p_{\beta 2}\}/2\big)}{\Gamma\big(\{C(s_{c1})-p_{\beta 1}\}/2\big)}\frac{\bar{\bar{b}}_{s_{c1}}^{[1]\{C(s_{c1})-p_{\beta 1}\}/2}}{\bar{\bar{b}}_{s_{c2}}^{[2]\{C(s_{c2})-p_{\beta 2}\}/2}}\frac{\bar{\bar{b}}_S^{[2](n-p_{\beta 2})/2}}{\bar{\bar{b}}_S^{[1](n-p_{\beta 1})/2}}.
\end{aligned}
$$

Empirical Bayes Gibbs Sampling

Notice that the marginal likelihood of λ is

$$L(\lambda, y) = \frac{L(\lambda, \psi, \sigma, y)}{\pi(\psi, \sigma|\lambda, y)},$$

where $L(\lambda, \psi, \sigma, y)$ is the complete likelihood that will be used in the EM algorithm. The expectation of the log complete data likelihood can be approximated as follows. Recall that B is a function of λ ($B = B(\lambda)$). Let an initial value of λ in the EM algorithm λ_o. When a variable or parameter depends on λ_o, we use the subscript o, for example, $B_o = B(\lambda_o)$. We have

$$\begin{aligned}
&Q(\lambda|\lambda_o) = E_{\psi,\sigma|\lambda_o,Y}\left[\log L(\lambda, \psi, \sigma, y)\right] \\
&= E_{\psi,\sigma|\lambda_o,Y}\left[\log L(\lambda|y, \alpha, \beta, \sigma)\right] \\
&= E_{\psi,\sigma|\lambda_o,Y}\left[\log f(y|\alpha, \beta, \sigma) + \log \pi(\alpha|\lambda, \sigma) + \log \pi(\beta, \sigma)\right] \\
&= E_{\psi,\sigma|\lambda_o,Y}\left[\log \pi(\alpha|\lambda, \sigma)\right] \\
&= E_{\psi,\sigma|\lambda_o,Y}\left[\frac{1}{2}\log\left|\frac{1}{\sigma^2}A\right| - \frac{1}{2\sigma^2}(\alpha - \bar{\alpha})^T A(\alpha - \bar{\alpha})\right] \\
&= E_{\psi,\sigma|\lambda_o,Y}\left[-p_\alpha \log(\lambda) - \frac{1}{2\lambda^2\sigma^2}(\alpha - \bar{\alpha})^T(\alpha - \bar{\alpha})\right] \\
&= -p_\alpha \log(\lambda) - \frac{1}{2\lambda^2}E_{\psi,\sigma|\lambda_o,Y}\left[\frac{(\alpha - \bar{\alpha})^T(\alpha - \bar{\alpha})}{\sigma^2}\right],
\end{aligned}$$

and

$$(\alpha - \bar{\alpha})^T(\alpha - \bar{\alpha}) = {\alpha_1}^2 + {\alpha_2}^2 + \cdots + {\alpha_{p_\alpha}}^2$$

because $\bar{\alpha}_1 \equiv \bar{\alpha}_2 \equiv \cdots \equiv \bar{\alpha}_{p_\alpha} \equiv 0$. Therefore,

$$\begin{aligned}
&E_{\psi,\sigma|\lambda_o,Y}\left[\frac{(\alpha - \bar{\alpha})^T(\alpha - \bar{\alpha})}{\sigma^2}\right] \\
&= E_{\psi,\sigma|\lambda_o,Y}\left[\frac{{\alpha_1}^2 + {\alpha_2}^2 + \cdots + {\alpha_{p_\alpha}}^2}{\sigma^2}\right] \\
&= E_{\psi,\sigma|\lambda_o,Y}\left[\frac{{\alpha_1}^2}{\sigma^2}\right] + E_{\psi,\sigma|\lambda_o,Y}\left[\frac{{\alpha_2}^2}{\sigma^2}\right] + \cdots + E_{\psi,\sigma|\lambda_o,Y}\left[\frac{{\alpha_{p_\alpha}}^2}{\sigma^2}\right].
\end{aligned}$$

Let $\left({\tau_1}^2, {\tau_2}^2, \ldots, {\tau_{p_\alpha+p_\beta}}^2\right)$ be the diagonal elements of $(B + X^T X)^{-1}$, that is,

$$\left({\tau_1}^2, {\tau_2}^2, \ldots, {\tau_{p_\alpha+p_\beta}}^2\right)^T = \text{diag}\{(B_o + X^T X)^{-1}\}.$$

Then,

$$\begin{aligned} E_{\psi|\sigma,\lambda_o,Y}\left[\frac{{\alpha_1}^2}{\sigma^2}\right] &= \frac{1}{\sigma^2}E_{\psi|\sigma,\lambda_o,Y}\left[{\alpha_1}^2\right] \\ &= \frac{1}{\sigma^2}\int {\alpha_1}^2\phi\left(\psi|\bar{\bar{\psi}},\sigma^2(B_o+X^TX)^{-1}\right)d\psi \\ &= \frac{1}{\sigma^2}\int \alpha_1^2\phi(\alpha_1|\bar{\bar{\alpha}}_1,\tau_1^2\sigma^2)d\alpha_1 \\ &= \frac{1}{\sigma^2}\left(\mathrm{Var}_{\alpha_1|\sigma,\lambda_o}(\alpha_1)+E^2_{\alpha_1|\sigma,\lambda_o}[\alpha_1]\right) \\ &= \tau_1^2+{\bar{\bar{\alpha}}_1}^2\frac{1}{\sigma^2}. \end{aligned}$$

Because $\bar{\bar{\alpha}}_1$ is independent of σ,

$$\begin{aligned} E_{\psi,\sigma|\lambda_o,Y}\left[\frac{{\alpha_1}^2}{\sigma^2}\right] &= E_{\sigma|\lambda_o}\left[E_{\psi|\sigma,\lambda_o}\left[\frac{{\alpha_1}^2}{\sigma^2}\right]\right] \\ &= \tau_1^2+{\bar{\bar{\alpha}}_1}^2E_{\sigma|\lambda_o,Y}\left[\sigma^{-2}\right], \end{aligned}$$

where

$$\begin{aligned} E_{\sigma|\lambda_o,Y}\left[\sigma^{-2}\right] &= \int \sigma^{-2}\frac{2\bar{\bar{b}}^{\bar{\bar{a}}}}{\Gamma(\bar{\bar{a}})}\sigma^{-2\bar{\bar{a}}-1}e^{-\bar{\bar{b}}\sigma^{-2}}d\sigma \\ &= \int \frac{2\bar{\bar{b}}^{\bar{\bar{a}}}}{\Gamma(\bar{\bar{a}})}\sigma^{-2(\bar{\bar{a}}+1)-1}e^{-\bar{\bar{b}}\sigma^{-2}}d\sigma \\ &= \int \frac{\Gamma(\bar{\bar{a}}+1)}{\Gamma(\bar{\bar{a}})\cdot\bar{\bar{b}}}\frac{2\bar{\bar{b}}^{\bar{\bar{a}}+1}}{\Gamma(\bar{\bar{a}}+1)}\sigma^{-2(\bar{\bar{a}}+1)-1}e^{-\bar{\bar{b}}\sigma^{-2}}d\sigma \\ &= \frac{\Gamma(\bar{\bar{a}}+1)}{\Gamma(\bar{\bar{a}})\cdot\bar{\bar{b}}}. \end{aligned}$$

Therefore,

$$E_{\psi,\sigma|\lambda_o,Y}\left[\frac{{\alpha_1}^2}{\sigma^2}\right] = \tau_1^2+{\bar{\bar{\alpha}}_1}^2\left(\frac{\Gamma(\bar{\bar{a}}+1)}{\Gamma(\bar{\bar{a}})\cdot\bar{\bar{b}}}\right),$$

and

$$E_{\psi,\sigma|\lambda_o,Y}\left[\frac{(\alpha-\bar{\alpha})^T(\alpha-\bar{\alpha})}{\sigma^2}\right] = \sum_{i=1}^{p_\alpha}\left(\tau_i^2+{\bar{\bar{\alpha}}_i}^2\frac{\Gamma(\bar{\bar{a}}+1)}{\Gamma(\bar{\bar{a}})\cdot\bar{\bar{b}}}\right).$$

Hence,

$$\begin{aligned}\frac{\partial}{\partial\lambda}Q(\lambda|\lambda_o) &= \frac{\partial}{\partial\lambda}\left\{-p_\alpha\log(\lambda) - \frac{1}{2\lambda^2}E_{\psi,\sigma|\lambda_o,Y}\left[\frac{(\alpha-\bar{\alpha})^T(\alpha-\bar{\alpha})}{\sigma^2}\right]\right\} \\ &= \frac{\partial}{\partial\lambda}\left\{-p_\alpha\log(\lambda) - \frac{1}{2\lambda^2}\sum_{i=1}^{p_\alpha}\left(\tau_i^2 + \overline{\overline{\alpha}}_i{}^2\frac{\Gamma(\overline{\overline{a}}+1)}{\Gamma(\overline{\overline{a}})\cdot\overline{\overline{b}}}\right)\right\} \\ &= -p_\alpha\lambda^{-1} + \lambda^{-3}\sum_{i=1}^{p_\alpha}\left(\tau_i^2 + \overline{\overline{\alpha}}_i{}^2\frac{\overline{\overline{a}}}{\overline{\overline{b}}}\right)\end{aligned}$$

and

$$\begin{aligned}&\frac{\partial}{\partial\lambda}Q(\lambda|\lambda_o) = 0 \\ &\Leftrightarrow \lambda = \sqrt{\frac{\sum_{i=1}^{p_\alpha}\left(\tau_i^2 + \overline{\overline{\alpha}}_i{}^2\frac{\overline{\overline{a}}}{\overline{\overline{b}}}\right)}{p_\alpha}}.\end{aligned}$$

Now, we are ready to set up an EM algorithm as

Step 1. Let $l = 1$ and $\overline{\overline{a}} = (n - p_\beta)/2$. Assign an initial value, $\hat{\lambda} = \hat{\lambda}^{(0)}$.

Step 2. Calculate

$$\begin{aligned}&B^{(l-1)} = \begin{pmatrix} \frac{1}{(\hat{\lambda}^{(l-1)})^2}\cdot I_{p_\alpha p_\alpha} & 0_{p_\alpha p_\beta} \\ 0_{p_\beta p_\alpha} & 0_{p_\beta p_\beta}\end{pmatrix}, \\ &\overline{\overline{\psi}}^{(l-1)} = (B^{(l-1)} + X^TX)^{-1}(B^{(l-1)}\overline{\psi} + X^TX\hat{\psi}), \\ &\overline{\overline{\alpha}}_k^{(l-1)} = \overline{\overline{\psi}}_k^{(l-1)}, \text{ where } k = 1, 2, \ldots, p_\alpha, \\ &\left(\tau_1^{(l-1)^2}, \tau_2^{(l-1)^2}, \ldots, \tau_{p_\alpha+p_\beta}^{(l-1)\ 2}\right)^T = \text{diag}\{(B^{(l-1)} + X^TX)^{-1}\}, \\ &\overline{\overline{b}}^{(l-1)} = (y - X\overline{\overline{\psi}}^{(l-1)})^T(y - X\overline{\overline{\psi}}^{(l-1)}) + (\overline{\psi} - \overline{\overline{\psi}}^{(l-1)})^TB^{(l-1)}(\overline{\psi} - \overline{\overline{\psi}}^{(l-1)}).\end{aligned}$$

Step 3. Calculate

$$\hat{\lambda}^{(l)} = \sqrt{\frac{\sum_{j=1}^{p_\alpha}\left(\tau_j^{(l-1)^2} + \overline{\overline{\alpha}}_j{}^{(l-1)^2}\frac{\overline{\overline{a}}}{\overline{\overline{b}}^{(l-1)}}\right)}{p_\alpha}}.$$

Step 4. Let $l = l + 1$ and go to Step 2 until $\hat{\lambda}^{(l)}$ converges.

Once the algorithm converges, we use $\hat{\lambda}^{(l)}$ as estimates for λ.

REFERENCES

AITKIN, M. (1991). "Posterior Bayes factors", *Journal of the Royal Statistical Society*, Ser. B, **53**, 111–142.

AKAIKE, H. (1973). "Information theory and an extension of the maximum likelihood principle", In *Second International Symposium on Information Theory* (Petrov, B. N. and Csáki, F., eds.), 267–281.

ATKINSON, A. C. (1978). "Posterior probabilities for choosing a regression model", *Biometrika*, **65**, 39–48.

BERGER, J. O. AND PERICCHI, L. R. (1996). "The intrinsic Bayes factor for linear models", In *Bayesian Statistics 5* (Bernardo, J. M. *et al.*, eds.), 23–42, Oxford University Press, New York.

BRUMBACK, B. A., RUPPERT, D. AND WAND, M. P. (1999). Comment on "Variable selection and function estimation in additive nonparametric regression using a data-based prior" (by Shively, T. S. *et al.*), *Journal of the American Statistical Association*, **94**, 794–797.

CASELLA, G. (2001). "Empirical Bayes Gibbs sampling", *Biostatistics*, **2**, 485–500.

DUDOIT, S., YANG, Y. H., CALLOW, M. J. AND SPEED, T. P. (2002). "Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments", *Statistica Sinica*, **12**, 111–139.

GEISSER, S. AND EDDY, W. F. (1979). "A predictive approach to model selection", *Journal of the American Statistical Association*, **74**, 153–160.

GELFAND, A. E. AND DEY, D. K. (1994). "Bayesian model choice: asymptotics and exact calculations", *Journal of the Royal Statistical Society*, Ser. B, **56**, 501–514.

HASTIE, T., TIBSHIRANI, R. AND FRIEDMAN, J. (2001). *The Elements of Statistical Learning*, Springer-Verlag, New York.

JOO, Y. (2003). *Evaluation of Model Selection Criteria in Log Spline Models*, Ph. D. Dissertation, Cornell University, New York.

JUDGE, G. G., HILL, R. C., GRIFFITHS, W. E., LÜTKEPOHL, H. AND LEE, T. C. (1988). *Introduction to the Theory and Practice of Econometrics*, 2nd ed., John Wiley & Sons, New York.

O'HAGAN, A. (1991). Discussion on "Posterior Bayes factors" (by Aitkin, M.), *Journal of Royal Statistic Society*, Ser. B, **53**, 136.

PEÑA, D. AND TIAO, G. C. (1992). "Bayesian robustness functions for linear models", In *Bayesian Statistics 4* (Bernardo, J. M. *et al.* eds.), 365–389, Oxford University Press, New York.

RUPPERT, D., WAND, M. P. AND CAROLL, R. J. (2003). *Semiparametric Regression*, Cambridge University Press, Cambridge.

SCHWARZ, G. (1978). "Estimating the dimension of a model", *The Annals of Statistics*, **6**, 461–464.