Bayesian Analysis (2006) 1, Number 2, pp. 375–383

Misinformation in the conjugate prior for the linear model with implications for free-knot spline modelling

Christopher J. Paciorek*

Abstract. In the conjugate prior for the normal linear model, the prior variance for the coefficients is a multiple of the error variance parameter. However, if the prior mean for the coefficients is poorly chosen, the posterior distribution of the model can be seriously distorted because of prior dependence between the coefficients and error variance. In particular, the error variance will be overestimated, as will the posterior variance of the coefficients. This occurs because the prior mean, which can be thought of as a weighted pseudo-observation, is an outlier with respect to the real observations. While this situation will be easily noticed and avoided in simple models, in more complicated models, the effect can be easily overlooked. The issue arises in the unit information (UI) prior, a conjugate prior in which the prior contributes information equal to that in one observation. In particular, a successful Bayesian nonparametric regression model — Bayesian Adaptive Regression Splines (BARS) — that relies on the UI prior for its model selection step suffers from this problem, and addressing the problem within the Bayesian paradigm alters the penalty on model dimensionality.

Keywords: Bayes factor, BIC, model selection, nonparametric regression, unit information prior

1 Introduction

Consider the simple normal mean problem with observations, $Y_i,\ i = 1, \ldots, n$, independent and identically distributed, $Y_i \sim \mathrm{N}(\mu, \sigma^2)$. The conjugate normal inverse-gamma prior is

$$\begin{aligned} \mu|\sigma^2 &\sim \mathrm{N}(\mu_0, \sigma^2/n_0) \\ \sigma^2 &\sim \mathrm{IG}(a, b), \end{aligned} \tag{1}$$

where n_0 is a parameter that can be interpreted as the number of prior observations. Gelman et al. (2003, p. 71) discuss this prior and note the prior dependence between μ and σ^2, saying that it provides a way to calibrate the prior for μ based on the scale of measurement of the observations. They also mention that additional uncertainty is introduced into the model based on the difference between the prior mean and the sample mean. Here I expand on that to note that when the prior mean is far from the sample mean and n_0 is not close to zero, this additional uncertainty can cause the

*Department of Biostatistics, Harvard School of Public Health, Boston, MA, http://www.biostat.harvard.edu/~paciorek

model to overestimate both the error variance and the posterior variance of μ, as the deviation of the prior mean from the posterior estimate cascades through the model. The prior dependence causes the data to inform σ^2 not just through their deviation from the estimate of μ, but also through the deviation of the estimated μ from μ_0. In essence a poorly chosen prior mean is a pseudo-observation that is an outlier with respect to the true observations. The situation is similar for the general linear model, for which the conjugate prior takes the form

$$\boldsymbol{\beta} \sim \mathrm{N}(\boldsymbol{\beta_0}, c\sigma^2(B^T B)^{-1}), \tag{2}$$

where B is the design matrix and c scales the prior. George and Foster (2000) discuss this prior in detail for the variable selection problem in the normal linear model. The conjugate prior with $c = 1/g$ is the g-prior of Zellner (1986).

One important form of the conjugate prior is the unit information (UI) prior. The UI prior for a parameter, $\boldsymbol{\psi}$, is defined to have variance equal to the inverse Fisher information arising from one observation (Kass and Wasserman 1995). A normal UI prior would be

$$\boldsymbol{\psi} \sim \mathrm{N}\left(\boldsymbol{\psi_0}, I^{-1}_{\boldsymbol{\psi}\boldsymbol{\psi}}(\boldsymbol{\psi_0}, \boldsymbol{\theta})\right),$$

where $I^{-1}_{\boldsymbol{\psi}\boldsymbol{\psi}}$ is the inverse of the block of the Fisher information matrix corresponding to $\boldsymbol{\psi}$. This inverse is the asymptotic variance of the MLE, $\hat{\boldsymbol{\psi}}$, assuming that the information matrix is block diagonal with respect to a nuisance parameter, $\boldsymbol{\theta}$. Kass and Wasserman (1995) and Pauler (1998) consider the UI prior as a default prior and show that with this prior, the Bayes factor for model selection can be approximated by the Schwarz criterion, also known as the Bayes information criterion (BIC). Earlier suggestions of such a prior include Jeffreys (1967) and Zellner and Siow (1980). Returning to the normal mean problem, the UI prior for μ is

$$\mu \sim \mathrm{N}(\mu_0, \sigma^2),$$

where σ^2 is the error variance. Since the prior variance for μ is the same as the variance of an observation, the prior contributes one unit of information to the posterior. One can think of the prior mean, μ_0, as one pseudo-observation, or prior observation. This is a special case of the general conjugate prior (1), in which the prior contributes as much information as n_0 observations. Similarly, the UI prior for the linear model has $c = n$ in (2), and the resulting conditional posterior precision matrix for $\boldsymbol{\beta}|\sigma^2$ is $\frac{1}{\sigma^2}\left(\frac{1}{n} + 1\right) B^T B$, with the prior contributing the term $\frac{1}{n}$. The prior contributes one unit of information, or $\frac{1}{n}$ as much information as the likelihood. DiMatteo et al. (2001) use this prior for a free-knot spline nonparametric regression model, with B a cubic B-spline basis matrix that varies with the number and location of knots. The knots are parameters in the model and are estimated via Markov chain Monte Carlo (MCMC).

The conjugate prior for the linear model provides a convenient closed form posterior, and its UI form has intuitive appeal and is related to using BIC for model selection. However, if the prior mean is poorly chosen, the resulting posterior can give badly distorted inference for both the posterior variance of the parameter of interest and

for the posterior distribution of the error variance, as I show for a simple example in section 2. In section 3, I discuss what happens in a more complicated model, the nonparametric regression model of DiMatteo et al. (2001), why one might overlook the problem in practice, and why it is difficult to resolve the problem when the UI prior is used for model selection purposes. I close with some suggestions for detecting and avoiding the problem.

2 Prior misinformation

To illustrate the problem more concretely, I take the simple normal mean problem as an illustrative example and use the unit information form of the conjugate prior. Let the observations, Y_i, $i = 1, \ldots, n$, be independent and identically distributed, $Y_i \sim \mathrm{N}(\mu, \sigma^2)$, with the UI prior, $\mu \sim \mathrm{N}(\mu_0, \sigma^2)$, and, without loss of generality, take $\mu_0 = 0$. Next, take the improper prior, $\pi(\sigma^2) \propto \frac{1}{\sigma^2}$, which is IG$(0, 0)$. The argument does not depend on this prior, and a proper conjugate inverse-gamma (IG) prior could be substituted here. The conditional posterior is $\mu|\sigma^2, \boldsymbol{y} \sim \mathrm{N}\left(\frac{n}{n+1}\bar{y}, \frac{\sigma^2}{n+1}\right)$. The fraction $\frac{n}{n+1}$ shrinks the posterior mean slightly toward 0, or more generally the prior mean. With a reasonable amount of data, the influence of the prior mean will be limited, so if interest focuses solely on the point estimate, the effect of the conjugate prior is limited. However, the limited effect on the posterior mean may lead one to overlook the effect on the uncertainty estimates. Consider the marginal posterior for σ^2,

$$\sigma^2|\boldsymbol{y} \sim \mathrm{IG}\left(\frac{n}{2}, \frac{1}{2}\left(ns^2 + \frac{\sum y_i^2}{n+1}\right)\right),$$

where $s^2 \equiv \frac{\sum (y_i - \bar{y})^2}{n+1}$, which gives us a point estimate,

$$\mathrm{E}(\sigma^2|\boldsymbol{y}) = \frac{ns^2 + \frac{\sum y_i^2}{n+1}}{n-2}.$$

This posterior is similar to the posterior for σ^2 under a noninformative prior for μ except for the term, $\frac{\sum y_i^2}{n+1}$, which can be consequential when the data are far from zero, and inflates the estimate for σ^2. The expression in the second parameter of the IG posterior can be expressed as $ns^2 + \frac{\sum y_i^2}{n+1} = \sum (y_i - \frac{n}{n+1}\bar{y})^2 + (\frac{n}{n+1}\bar{y})^2$, where the first term accounts for the squared deviations of the observations from the posterior mean of μ and the second term is the deviation of the prior pseudo-observation ($\mu_0 = 0$ in this case) from the posterior mean for μ. While the second term does not appear asymptotically, in finite samples, as I will show next, the estimate for σ^2 can be seriously inflated. Of more serious concern in many applications, an inflated estimate of σ^2 inflates $\mathrm{V}(\mu|\sigma^2, \boldsymbol{y}) = \frac{\sigma^2}{n+1}$, and therefore the marginal variance of μ,

$$\mathrm{V}(\mu|\boldsymbol{y}) = \frac{1}{n+1}\left(\frac{ns^2 + \frac{\sum y_i^2}{n+1}}{n-2}\right),$$

which contains the extra term, $\frac{\sum y_i^2}{n+1}$, which is not present under an improper prior for μ.

Consider a simple example of 100 observations with $\bar{y} = 10$ and $\sum(y_i - \frac{n}{n+1}\bar{y})^2 = 100 \cdot 1$. If $\mu_0 = 0$, the pseudo-observation at zero introduced into the model by the UI prior is an outlier with respect to the data. The posterior mean for μ is $\frac{100}{101} \cdot 10 \approx 10$, which is reasonable. But the estimate of the second parameter in the IG distribution for σ^2 is inflated by $\frac{1}{2}\left(\frac{100}{101} \cdot 10\right)^2$ over that of an improper prior. The resulting mean of the IG posterior for σ^2 is 1.4^2, which is much larger than the mean squared deviation of 1^2. Correspondingly, $\mathrm{V}(\mu|\boldsymbol{y}) = \left(\frac{1.4^2}{101}\right) \approx \left(\frac{1.4}{10}\right)^2$ is inflated by a factor of 1.4^2 over the estimate from an improper prior, $\frac{\sigma^2}{n} = \left(\frac{1}{10}\right)^2$. With smaller deviations of the observations from their mean or with a larger value of $\bar{Y}$, the effect would be more pronounced. We see that the effect of the single outlying pseudo-observation from the UI prior can be substantial.

I do not want to overstate the situation. If we take the conjugate prior, $\mu \sim \mathrm{N}(0, \sigma^2/n_0)$, with n_0 small, the effect on the uncertainty estimates diminishes as $n_0 \to 0$, so in the general conjugate prior, we can resolve the problem by using small n_0. Furthermore, in simple models, one will probably recognize when the prior mean is poorly chosen. However, I will show that the UI form of the conjugate prior is of particular interest in important models in which the problem arises in practice and cannot always be easily solved.

3 The unit information prior and free-knot spline modelling

The practical importance of this feature of the conjugate prior arises because the UI prior has been suggested as a default prior and has been used in successful Bayesian models — in particular, the nonparametric regression model, Bayesian Adaptive Regression Splines (BARS). In this model, the UI prior is specifically chosen because of its role in model selection. Furthermore, the prior is placed on the coefficients of the B-spline basis with varying numbers and locations of knots, which makes it difficult to choose a reasonable prior mean or to know when the prior mean is poorly chosen.

3.1 The UI prior in BARS

DiMatteo et al. (2001) specify a free-knot spline nonparametric regression model known as BARS, with an unknown number and location of knots. Conditional on the number, k, and location of knots, $\boldsymbol{\xi}$, they specify the regression function,

$$f(x) = \sum_{j=1}^{k+2} \beta_j b_j(x),$$

where $b_j(\cdot)$ is the jth B-spline basis function and $\boldsymbol{\beta}$ is a vector of basis coefficients. Letting B denote the basis matrix formed from the basis functions and suppressing its dependence on $(k, \boldsymbol{\xi})$, the UI prior for $\boldsymbol{\beta}$ is

$$\boldsymbol{\beta} \sim \mathrm{N}_{k+2}(\mathbf{0}, n\sigma^2(B^TB)^{-1}),$$

where σ^2 is the error variance parameter, with improper prior, $\pi(\sigma^2) \propto 1/\sigma^2$. DiMatteo et al. (2001) fit the model by reversible-jump MCMC to account for the change in model dimension as knots are added and deleted.

Similar calculations to those done in the previous section for the simple normal mean problem can be performed here, for the case of normal data, conditional on the number and location of knots. We see that the posterior mean for $\boldsymbol{\beta}$, $\bar{\boldsymbol{\beta}}$, is shrunk toward zero relative to the MLE, $\hat{\boldsymbol{\beta}}$: $\bar{\boldsymbol{\beta}} = \mathrm{E}(\boldsymbol{\beta}|k, \boldsymbol{\xi}, \boldsymbol{y}) = \frac{n}{n+1}(B^TB)^{-1}B^T\boldsymbol{y} = \frac{n}{n+1}\hat{\boldsymbol{\beta}}$. Once again, this has limited practical impact with reasonable sample size. The posterior distribution for the error variance is

$$\sigma^2|k, \boldsymbol{\xi}, \boldsymbol{y} \sim \mathrm{IG}\left(\frac{n}{2}, \frac{1}{2}(\boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{y}^TB\bar{\boldsymbol{\beta}})\right),$$

with a posterior mean of

$$\mathrm{E}(\sigma^2|k, \boldsymbol{\xi}, \boldsymbol{y}) = \frac{\boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{y}^TB\bar{\boldsymbol{\beta}}}{n-2} = \frac{\sum \bar{r}_i^2 + \frac{1}{n+1}\boldsymbol{y}^T\bar{\boldsymbol{f}}}{n-2}, \tag{3}$$

where $\bar{r}_i = y_i - \bar{f}_i$ is the residual from the ith posterior mean fitted value, $\bar{f}_i$. If we were to use an improper prior for $\boldsymbol{\beta}$, $\pi(\boldsymbol{\beta}) \propto 1$, rather than the UI prior, the posterior mean would be

$$\mathrm{E}(\sigma^2|k, \boldsymbol{\xi}, \boldsymbol{y}) = \frac{\boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{y}^TB\hat{\boldsymbol{\beta}}}{n-2} = \frac{\sum \hat{r}_i^2}{n-2}$$

Relative to the expected value from an improper prior, in which the estimate of the error variance is based solely on the squared residuals, the posterior mean from the UI prior is inflated by $\frac{1}{n+1}\boldsymbol{y}^T\bar{\boldsymbol{f}} = \frac{1}{n}\bar{\boldsymbol{f}}^T\bar{\boldsymbol{f}}$. Returning to (3), the inflation is of concern when the second term in the numerator, which involves the sum of squares of the function values, $\bar{\boldsymbol{f}}^T\bar{\boldsymbol{f}}$, is large relative to the first term, which involves the residuals. This occurs when the signal to noise ratio is large and the sample size is small; it can happen even if $\int f(x)dx \approx 0$. The basic problem with the UI prior in this situation is that some basis coefficients need to be large in order to represent a function for which $\boldsymbol{f}^T\boldsymbol{f}$ is large, but the prior specifies that the coefficients have zero mean and prior variance based on σ^2. If the noise variance is small, this introduces a conflict in the estimation of σ^2 between the small residual variation about the function estimate and the large deviations between the estimated coefficients and their prior mean of zero. In the full free-knot model, the number and location of knots are also sampled during the model fitting, but the effect of the UI prior on the estimation of the error variance — and therefore the uncertainty in the basis coefficients and resulting regression function — remains.

This can be material in practice. DiMatteo et al. (2001) demonstrate the success of BARS on three test datasets, the first of which has a mean function that is not near

zero, combined with small error variance. As shown in Paciorek (2003, Sec. 4.6.1), while the posterior mean of the regression function for this first dataset is estimated well, the estimate of the error variance is inflated by a factor of almost 2 (estimated to be 1.5 compared to the true value of 0.81). If draws of the error variance are used in computing uncertainty estimates for the regression function, the problem carries through to one's uncertainty about the estimated function.

In this example, one could solve the problem and retain the Bayesian estimation of σ^2 by subtracting off the mean of the data. However, this simple solution does not work when the function is centered on zero but has a high signal to noise ratio. Consider the function $f(x) = 10 \cdot \sin(2\pi x)$ for $n = 30$ observations on a grid over $x \in (0, 1)$ and $\sigma^2 = 1$. Fitting this for 1000 simulations using a fixed five-knot B-spline basis allows me to estimate $\mathrm{E}(\sigma^2|\boldsymbol{y})$, averaging over the 1000 simulations, as 2.59, compared to the true value of 1. Similarly, fitting one simulated dataset using BARS gives an estimate for σ^2 of 2.32. This example has been carefully chosen as a scenario in which the Bayesian error variance estimate is inflated, with large signal to noise ratio and small n, but it illustrates that subtracting off the mean of the data does not completely solve the problem. As we will see next, avoiding the UI prior raises other complications.

3.2 The UI prior and model selection

The UI prior is used in BARS not only for the convenience of a conjugate prior, which allows one to integrate $\boldsymbol{\beta}$ and σ^2 out of the model and sample only $(k, \boldsymbol{\xi})$ in the MCMC, but more importantly because it plays a critical role in model selection. BARS moves between models of different dimensions by adding and deleting knots via a reversible-jump algorithm. The ratio of marginal likelihoods, which is the Bayes factor (BF), plays the key role in determining whether to change model dimension. Consider the ratio when the proposed model contains one more knot than the current model, $k^* = k + 1$,

$$\begin{aligned}
\mathrm{BF} = \frac{f(\boldsymbol{y}|k^*, \boldsymbol{\xi}^*)}{f(\boldsymbol{y}|k, \boldsymbol{\xi})} &= \frac{1}{\sqrt{n+1}} \left(\frac{\boldsymbol{y}^T (I - n(n+1)^{-1} B (B^T B)^{-1} B) \boldsymbol{y}}{\boldsymbol{y}^T (I - n(n+1)^{-1} B^* (B^{*T} B^*)^{-1} B^*) \boldsymbol{y}} \right)^{\frac{n}{2}} \\
&= \frac{1}{\sqrt{n+1}} \left(\frac{\sum \bar{r}_i^2 + \frac{1}{n} \bar{\boldsymbol{f}}^T \bar{\boldsymbol{f}}}{\sum \bar{r}_i^{*2} + \frac{1}{n} \bar{\boldsymbol{f}}^{*T} \bar{\boldsymbol{f}}^*} \right)^{\frac{n}{2}} \\
&= \frac{1}{\sqrt{n+1}} \left(\frac{\sum \hat{r}_i^2 + \frac{1}{n+1} \hat{\boldsymbol{f}}^T \hat{\boldsymbol{f}}}{\sum \hat{r}_i^{*2} + \frac{1}{n+1} \hat{\boldsymbol{f}}^{*T} \hat{\boldsymbol{f}}^*} \right)^{\frac{n}{2}}, \qquad (4)
\end{aligned}$$

where $B^* = B_{k^*, \boldsymbol{\xi}^*}$ (i.e., the proposed B-spline basis matrix), and where the last equality is expressed in terms of maximum likelihood estimates. Ignoring the additional UI-induced term, $\frac{1}{n} \bar{\boldsymbol{f}}^T \bar{\boldsymbol{f}}$, in both numerator and denominator, the BF is approximately equal to $\exp(-\mathrm{BIC}/2)$ with the dimensionality penalty being $\sqrt{n+1}$. The $\sqrt{n+1}$ term, when considered on the log scale, corresponds to the BIC penalty, $\log n$, on the log likelihood ratio for models differing by one parameter. In addition to the theoretical appeal of the BIC-type penalty induced by the UI prior, this particular penalty has been

successful in simulations and ongoing applied work (DiMatteo et al. 2001; Kass et al. 2003; Wallstrom et al. 2004).

The value of c in the general conjugate prior, $\boldsymbol{\beta} \sim \mathrm{N}(0, c\sigma^2(B^TB)^{-1})$, determines the penalty, $\sqrt{c+1}$, that replaces $\sqrt{n+1}$ in the BF (4). Particular choices of c correspond to AIC and the risk inflation criteria (RIC) (Foster and George 1994), so choice of c may depend on one's utilities for model selection (Clyde 2001). However, AIC corresponds to $c < n$, while RIC corresponds to $c = p^2$, where $p = k + 2$ is the varying number of basis functions, and may also result in $c < n$. Since $c < n$ gives a more informative prior than the UI prior, neither criterion seems useful for addressing the estimation of σ^2 in BARS. In a similar context to BARS, but based on knot selection from a fixed set of knots, Smith and Kohn (1996) found that $c = 100$ worked well, with their results insensitive to $10 < c < 1000$.

An alternative approach that attempts to address the effect of c on the estimation of σ^2 would be to make the conjugate prior for $\boldsymbol{\beta}$ more diffuse, taking $c > n$. However, increasing c increases the penalty on larger models in (4) relative to the BIC-type penalty of the BF resulting from the UI prior. It is well-known that BF calculations require proper priors and that diffuse priors favor smaller models (Kass and Raftery 1995; Gelman et al. 2003), because the marginal likelihood with diffuse priors averages the conditional likelihood over extreme values of the parameters. It may be possible to choose a value of c that produces a penalty term that performs well in practice — perhaps better than $c = n$ — and also provides a reasonable estimate of σ^2. However, it is difficult to know what this value should be, particularly given that $c = n$ has performed well and corresponds to BIC. In summary, there is no avoiding the dependence of the Bayes factor and model selection in this context on the prior for $\boldsymbol{\beta}$ and the resulting potential conflict in choosing c between the model selection criterion and estimation of σ^2. In the next section I suggest some alternative modifications to the UI prior.

3.3 Modifying the UI prior

Subtracting off the mean of the data will solve the problem in many cases; I suggest this as an initial general solution. To determine if the UI prior is still biasing the estimate of the error variance and the posterior variance of the function estimate, I suggest comparing the estimate of σ^2 from the posterior to a classical estimate of σ^2, for example, computing a classical estimate at every MCMC iteration, $\tilde{\sigma}_t^2 = \sum(y_i - f_{t,i})^2/(n-p)$, $t = 1, \ldots, T$, or more simply $\tilde{\sigma}^2 = \sum(y_i - \bar{f}_i)^2/(n-\bar{p})$ where the overbar indicates the posterior mean estimates. If the estimates are sufficiently different, which should only occur when the signal to noise ratio is high and the number of observations small, one could consider the following approaches.

Ideally we could choose a prior mean, $\boldsymbol{\beta_0}$, that is reasonable, but since the prior is conditional on $(k, \boldsymbol{\xi})$, this is difficult to do. An ad hoc alternative that has an empirical Bayes flavor and aims to retain the approximate penalty of $\sqrt{n+1}$ is to fix the error variance term in the prior variance for $\boldsymbol{\beta}$, taking $\boldsymbol{\beta} \sim \mathrm{N}(\mathbf{0}, s^2 n(B^TB)^{-1})$, where s^2 is a reasonable estimate of σ^2. This might be an estimate based on $\tilde{\sigma}^2$ from an initial model

run. We no longer have a joint conjugate model for $(\boldsymbol{\beta}, \sigma^2)$, but we can still integrate $\boldsymbol{\beta}$ out of the model, adding a sampling step for σ^2 to the MCMC. This approach gives approximately the $\sqrt{n+1}$ penalty, because the marginal likelihood, $\frac{f(\boldsymbol{y}|k,\boldsymbol{\xi},\sigma^2)}{f(\boldsymbol{y}|k^*,\boldsymbol{\xi}^*,\sigma^2)}$, has a penalty of $\sqrt{s^2 n/\sigma^2 + 1}$, which for $s^2 \approx \sigma^2$ is $\sqrt{n+1}$.

An alternative, which is the current approach in BARS (Wallstrom et al. 2005), is to not be fully Bayesian where σ^2 is concerned, using a plug-in estimate based on the residuals. One can use a classical estimate of σ^2 at each iteration and use these values when estimating the uncertainty in $\boldsymbol{\beta}$.

George and Foster (2000), in the context of variable selection in the normal linear model, suggest the use of empirical Bayes or fully Bayes analysis for c, allowing the data to help determine the model selection penalty term. One difficulty with the empirical Bayes approach applied to BARS is that the marginal likelihood used to estimate c involves an intractable integral over the locations of the knots; the uncountable number of sets of knot locations also makes numerical maximization difficult. To avoid such difficulties, George and Foster (2000) suggest conditioning on the model (i.e., $(k, \boldsymbol{\xi})$ in BARS), and using $\hat{c}_{k,\boldsymbol{\xi}} = \max(\hat{\boldsymbol{\beta}}_{k,\boldsymbol{\xi}}^T B_{k,\boldsymbol{\xi}}^T B_{k,\boldsymbol{\xi}} \hat{\boldsymbol{\beta}}_{k,\boldsymbol{\xi}}/(\sigma^2 p) - 1, 0)$, which would correspond to changing the penalty in (4) at every MCMC iteration, as well as sampling σ^2 within the chain. They show that the conditional approach is asymptotically equivalent to using BIC. Finally, George and Foster (2000) suggest placing a diffuse prior on c and sampling from its posterior during the MCMC. These approaches deserve consideration; the key issue is how well they balance model selection with estimation of σ^2 in situations in which $c \leq n$ can cause the estimate of σ^2 to be distorted.

References

Clyde, M. (2001). "Discussion of 'The practical implementation of Bayesian model selection' by Chipman, H., E.I. George, and R.E. McCulloch." In Lahiri, P. (ed.), *Model Selection*, 117–124. Institute of Mathematical Statistics Lecture Notes. 381

DiMatteo, I., Genovese, C., and Kass, R. (2001). "Bayesian curve-fitting with free-knot splines." *Biometrika*, 88: 1055–1071. 376, 377, 378, 379, 381

Foster, D. P. and George, E. I. (1994). "The risk inflation criterion for multiple regression." *The Annals of Statistics*, 22: 1947–1975. 381

Gelman, A., Carlin, J., Stern, H., and Rubin, D. (2003). *Bayesian Data Analysis*. Boca Raton: Chapman & Hall CRC. 375, 381

George, E. I. and Foster, D. P. (2000). "Calibration and empirical Bayes variable selection." *Biometrika*, 87: 731–747. 376, 382

Jeffreys, H. (1967). *Theory of Probability*. Oxford: Oxford University Press, 3d, rev. edition. 376

Kass, R. and Raftery, A. (1995). "Bayes factors." *Journal of the American Statistical Association*, 90: 773–795. 381

Kass, R., Ventura, V., and Cai, C. (2003). "Statistical smoothing of neuronal data." *Network: Computation in Neural Systems*, 14: 5–15. 381

Kass, R. E. and Wasserman, L. (1995). "A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion." *Journal of the American Statistical Association*, 90: 928–934. 376

Paciorek, C. (2003). "Nonstationary Gaussian Processes for Regression and Spatial Modelling." Ph.D. thesis, Carnegie Mellon University, Pittsburgh, Pennsylvania. 380

Pauler, D. K. (1998). "The Schwarz criterion and related methods for normal linear models." *Biometrika*, 85: 13–27. 376

Smith, M. and Kohn, R. (1996). "Nonparametric regression using Bayesian variable selection." *Journal of Econometrics*, 75: 317–343. 381

Wallstrom, G., Kass, R., Miller, A., Cohn, J., and Fox, N. (2004). "Automatic correction of ocular artifact in the EEG: a comparison of regression-based and component-based methods." *International Journal of Psychophysiology*, 53: 105–119. 381

Wallstrom, G., Liebner, J., and Kass, R. (2005). "An implementation of Bayesian Adaptive Regression Splines (BARS) with S and R wrappers." *Journal of Statistical Software, under revision*. 382

Zellner, A. (1986). "On assessing prior distributions and Bayesian regression analysis with g-prior distributions." In Goel, P. and Zellner, A. (eds.), *Bayesian Inference and Decision Techniques*, 233–243. Amsterdam: Elsevier Science Publishers. 376

Zellner, A. and Siow, A. (1980). "Posterior odds ratios for selected regression hypotheses." In Bernardo, J., DeGroot, M., Lindley, D., and Smith, A. (eds.), *Bayesian Statistics – Proceedings of the First Valencia International Meeting*, 585–603. Valencia, Spain: University Press. 376

Acknowledgments

The author thanks Rob Kass for insightful comments and discussion.