CENTRE FOR APPLIED ECONOMIC RESEARCH

WORKING PAPER
(2006/02)

Bayesian Subset Selection and Model Averaging using a Centred and Dispersed Prior for the Error Variance

By E. Cripps, R.Kohn and D. Nott

ISSN 13 29 12 70
ISBN 0 7334 2323 X
www.caer.unsw.edu.au

Faculty of Commerce and Economics | UNSW THE UNIVERSITY OF NEW SOUTH WALES

BAYESIAN SUBSET SELECTION AND MODEL AVERAGING USING A CENTRED AND DISPERSED PRIOR FOR THE ERROR VARIANCE

Edward Cripps *[1], Robert Kohn[2], David Nott[3]

University of New South Wales

Summary

This article proposes a new data-based prior distribution for the error variance in a Gaussian linear regression model, when the model is used for Bayesian variable selection and model averaging. For a given subset of variables in the model, this prior has a mode that is an unbiased estimator of the error variance but is suitably dispersed to make it uninformative relative to the marginal likelihood. The advantage of this empirical Bayes prior for the error variance is that it is centred and dispersed sensibly and avoids the arbitrary specification of hyperparameters. The performance of the new prior is compared to that of a prior proposed previously in the literature using several simulated examples and two loss functions. For each example our paper also reports results for the model that orthogonalizes the predictor variables before performing subset selection. A real example is also investigated. The empirical results suggest that for both the simulated and real data, the performance of the estimators based on the prior proposed in our article compares favourably with that of a prior used previously in the literature.

Key words: empirical Bayes prior; Markov chain Monte Carlo; nonparametric regression; orthogonal predictors.

*Author to whom correspondence should be addressed.

[1]School of Economics, University of New South Wales, Sydney, NSW, 2052, Australia. email ecripps@unsw.edu.au

[2]Schools of Economics and Banking and Finance, Faculty of Economics, University of New South Wales, Sydney, NSW, 2052, Australia.

[3]Department of Statistics, University of New South Wales, Sydney, NSW, 2052, Australia.

1. Introduction

This article is concerned with subset selection and model averaging for linear, Gaussian regression models. The approach is Bayesian and by model averaging we mean taking a weighted average of regression models, where each model is defined by the subset of independent variables that it contains, and the weight for each model is its posterior probability. Subset selection in linear regression is an important theoretical and applied problem and a Bayesian analysis is given by a number of authors, (e.g. Mitchell & Beauchamp, 1988). George & McCulloch (1993) were the first to propose a Bayesian sampling method that enabled statisticians to consider problems with a large number of variables. Raftery, Madigan & Hoeting (1997) emphasize that if prediction is an important goal of the analysis, then it may be better to estimate the regression function as a weighted average of different subset estimators, rather than choosing a single 'optimal' subset of variables. See also the later discussions by George & McCulloch (1997) and Hoeting *et al.* (1999).

The contribution of our article is to present a new prior specification for the error variance. The prior specification for the complete regression model is hierarchical, with a vector of binary indicator variables specifying which covariates are in the model. The prior for the regression coefficients, conditional on the error variance and the independent variables included in the regression (specified by a set of indicator variables), is described first. The mean of this prior is centred at an unbiased estimate of the parameters that are specified as active by the set of indicator variables. It is proper, but uninformative relative to the likelihood, and is approximately the same as the sample size increases. The prior for the error variance is conditioned on the indicator variables only. It is a proper prior, with a mode that is the unbiased estimator of the error variance for a given subset of the regression coefficients, but it is uninformative relative to the information about the error variance contained in the marginal likelihood obtained by integrating out the regression coefficients. Conditional on a hyperparameter, the indicator variables have *a priori* independent Bernoulli distributions.

Our prior specification for the regression coefficients and the indicator variables is the same as that of Kohn, Smith & Chan (2001), and is similar to that used in previous work by Smith & Kohn (1996). However, the prior specification for the error variance seems different to previous approaches, and we argue that it has some conceptual and practical advantages over them. Usually, the prior for the error variance is taken as an inverse gamma distribution, with a mode close to 0 and parameters that do not depend on the number of variables in the model. We believe that any prior for the error variance that is skewed needs to be centred properly to be effective and this is the aim of our article.

We study the performance of our approach empirically using both simulated

and real data. Two loss functions are used to assess separately the performance of our prior on a point estimate of the error variance and on the predictive density.

Several authors (e.g. Clyde, Desimone & Parmigiani, 1996) have suggested orthogonalizing the variables before carrying out variable selection in order to speed up the computation. For completeness, our article also presents results when the variables are orthogonalized and reports on the effectiveness of orthogonalizing at estimating the predictive distribution. All the computations in the article are carried out using Markov chain Monte Carlo simulation.

The following simulated examples are considered in our article; (a) Predictors that are highly multicollinear; (b) Models where the number of predictors is large relative to the sample size; (c) ANOVA models with main effects and interaction effects; (d) Nonparametric regression problems where the unknown regression function is expressed in regression form as a linear combination of basis functions. This application was the focus of the work of Smith & Kohn (1996) and Kohn *et al.* (2001). The real data set investigated is the US crime data examined in Vandaele (1978). The simulation results and US crime data results show that the performance of the new prior specification compares favorably with the results obtained using the prior in Kohn *et al.* (2001), particularly when the number of predictors is large compared to the sample size.

The paper is organized as follows. Section 2 presents the model, the priors, the sampling scheme, the method of estimation, and the loss functions used to judge performance. Section 3 describes the various regression functions used in the simulation and the real example and compares our results to those using the prior in Kohn *et al.* (2001). Section 3 also presents a new efficient method of selecting knots for nonparametric regression problems. Section 4 concludes the paper.

2. Model, Orthogonalized Variables and Sampling Scheme

2.1. Model Description

Let $\boldsymbol{Y}$ be a vector of responses and $\boldsymbol{X}$ an $n \times k$ design matrix having ones in the first column and with the remaining columns containing the $(k-1)$ predictor variables. We assume the model

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}, \qquad \boldsymbol{e} \stackrel{\mathrm{d}}{=} \mathrm{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}) \,. \tag{1}$$

Let $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_k)$ be a vector of binary variables such that the ith column of $\boldsymbol{X}$, i.e., the ith predictor variable, is included in the regression if $\gamma_i = 1$, and is excluded if $\gamma_i = 0$. Let $\boldsymbol{X}_\gamma$ be the matrix obtained from $\boldsymbol{X}$ by including column i of $\boldsymbol{X}$ if $\gamma_i = 1$ and let $\boldsymbol{\beta}_\gamma$ be the corresponding subvector of $\boldsymbol{\beta}$. All the regression models contain an intercept, which means that the first column of $\boldsymbol{X}$ is always retained and γ_1 is identically 1. The vector $\boldsymbol{\gamma}$ indexes all regression models, or equivalently, all subsets of variables. Conditional on $\boldsymbol{\gamma}$, (1) becomes

$$\boldsymbol{Y} = \boldsymbol{X}_\gamma \boldsymbol{\beta}_\gamma + \boldsymbol{e}\,.$$

A prior distribution on the parameters $(\beta_\gamma, \sigma^2, \gamma)$ is specified in a hierarchical manner as

$$p(\boldsymbol{\beta}_\gamma|\sigma^2, \boldsymbol{\gamma})p(\sigma^2|\boldsymbol{\gamma})p(\boldsymbol{\gamma})\,,$$

where each of the densities is described below.

2.2. Prior for the Regression Coefficients

Before describing the prior for the regression coefficients we repeat the important point made by Kohn *et al.* (2001), which is that for any regression coefficient whose prior has a discrete component (usually zero) and a continuous component, the distribution of the continuous component must be proper. If it is improper, then this coefficient will always take its discrete value. In our paper, this means that all the regression coefficients except for the intercept must have proper continuous components. Our prior for $\boldsymbol{\beta}_\gamma$ assumes the intercept is always included in our model. All remaining predictor variables are mean corrected and we set the first column of $\boldsymbol{X}_\gamma$ to be a vector with all elements equal to $1/\sqrt{n}$ such that

$$\boldsymbol{X}_\gamma^\top \boldsymbol{X}_\gamma = \begin{pmatrix} 1 & 0 \\ 0 & \boldsymbol{A}_\gamma \end{pmatrix}\,.$$

The prior for $\boldsymbol{\beta}_\gamma$ is

$$p(\boldsymbol{\beta}_\gamma|\sigma^2, \boldsymbol{\gamma}) \sim N(\boldsymbol{b}_\gamma, \sigma^2 \boldsymbol{V}_\gamma) \tag{2}$$

where $\boldsymbol{b}_\gamma = (\boldsymbol{X}_\gamma^\top \boldsymbol{X}_\gamma)^{-1} \boldsymbol{X}_\gamma^\top \boldsymbol{Y}$ and

$$\boldsymbol{V}_\gamma = \begin{pmatrix} c_1 & 0 \\ 0 & c_2 \boldsymbol{A}_\gamma^{-1} \end{pmatrix}.$$

We set $c_2 = n$ in our model such that the prior has a covariance matrix that stays approximately the same as n increases. For all elements of $\boldsymbol{\beta}_\gamma$ except the intercept this is equivalent to writing

$$p(\boldsymbol{\beta}_\gamma|\sigma^2, \boldsymbol{\gamma}) \propto p(\boldsymbol{Y}|\boldsymbol{\beta}_\gamma, \sigma^2, \boldsymbol{\gamma})^{\frac{1}{n}}$$

and is the same prior used in Kohn *et al.* (2001). That is, the prior for $\boldsymbol{\beta}_\gamma$ is centred consistent with the likelihood and is an unbiased estimator of $\boldsymbol{\beta}_\gamma$, conditional on $\boldsymbol{\gamma}$ and σ^2, but has c_2 times the variance. Also, we set $c_1 = n^2$ such that the prior for β_1 has the same location as in Kohn *et al.* (2001) but is more diffuse since we do not perform any variable selection for the intercept.

2.3. Prior for the error variance

We define q_γ as the number of independent variables that are in the model specified by γ, i.e.

$$q_\gamma = \sum_{i=1}^{k} \gamma_i \,.$$

The marginal likelihood for σ^2 and $\boldsymbol{\gamma}$, with $\boldsymbol{\beta}_\gamma$ integrated out, is

$$p(\boldsymbol{Y}|\sigma^2, \boldsymbol{\gamma}) \propto (\sigma^2)^{-n/2} \exp\left(-\frac{\mathrm{SSE}(\boldsymbol{\gamma})}{2\sigma^2}\right) ,$$

where

$$\mathrm{SSE}(\gamma) = \boldsymbol{Y}^\top \boldsymbol{Y} - \boldsymbol{Y}^\top \boldsymbol{X}_\gamma (\boldsymbol{X}_\gamma^\top \boldsymbol{X}_\gamma)^{-1} \boldsymbol{X}_\gamma^\top \boldsymbol{Y} \,,$$

is the residual sum of squares for the model containing those predictors defined by $\boldsymbol{\gamma}$. This marginal likelihood is proportional to an inverse gamma density with parameters $(n/2)-1$ and $\mathrm{SSE}(\boldsymbol{\gamma})/2$, which has mode $\mathrm{SSE}(\boldsymbol{\gamma})/n$ and variance

$$\frac{\mathrm{SSE}(\gamma)^2}{(n-4)^2(n/2-3)} \qquad \text{for } n > 6.$$

We choose the prior for σ^2 as inverse gamma with shape and scale parameters a_σ and b_σ specified as follows. For a given $\boldsymbol{\gamma}$, let

$$a_\sigma = \frac{\kappa}{2} - 1 \qquad \text{and} \qquad b_\sigma = \frac{\kappa \mathrm{SSE}(\boldsymbol{\gamma})}{2(n-q_\gamma)} \,,$$

so the prior for σ^2 is

$$p(\sigma^2|\boldsymbol{\gamma}) \propto (\sigma^2)^{-\frac{\kappa}{2}} \exp\Big(-\frac{\kappa \mathrm{SSE}(\boldsymbol{\gamma})}{2(n-q_\gamma)\sigma^2}\Big) \tag{3}$$

with mode

$$\frac{\mathrm{SSE}(\boldsymbol{\gamma})}{(n-q_\gamma)} ,$$

and variance

$$\frac{\mathrm{SSE}(\boldsymbol{\gamma})^2}{(n-q_\gamma)^2} \frac{2\kappa^2}{(\kappa-4)^2(\kappa-6)} , \quad \text{for} \quad \kappa > 6.$$

The mode is an unbiased estimator of σ^2 for given $\boldsymbol{\gamma}$, and setting $\kappa = 7$ ensures the prior has finite variance and is much less informative than the likelihood. We also ran the model with $\kappa = 9, 10$, and 11. The results were insensitive to the choice of these values for κ. Thus, we can view (3) as an empirical Bayes prior for σ^2.

The prior for σ^2 is usually taken as an uninformative inverse gamma prior (e.g. George & McCulloch, 1993; Kohn *et al.*, 2001) that does not depend on $\boldsymbol{\gamma}$; for example, see (1).

2.4. Prior for the vector of binary indicator variables

As in Kohn *et al.* (2001) we specify the prior for $\boldsymbol{\gamma}$ as

$$p(\boldsymbol{\gamma}|\pi) = \pi^{q_\gamma - 1}(1-\pi)^{k-q_\gamma}, \qquad \text{with} \qquad 0 \leq \pi \leq 1\,,$$

i.e. the $\gamma_i, i = 2, \ldots, k$ are assumed to be independent with $p(\gamma_i = 1|\pi) = \pi$. For flexibility and convenience we place a beta hyperprior on π such that

$$\begin{aligned} p(\boldsymbol{\gamma}) &= \int p(\boldsymbol{\gamma}|\pi)p(\pi)d\pi \\ &= \frac{B(q_\gamma + a_\pi - 1, k - q_\gamma + b_\pi)}{B(a_\pi, b_\pi)}. \end{aligned} \tag{4}$$

The marginal likelihood for γ, with β_γ and σ^2 integrated out, is

$$\begin{aligned} p(\boldsymbol{Y}|\boldsymbol{\gamma}) &= \int p(\boldsymbol{Y}|\sigma^2, \boldsymbol{\gamma})p(\sigma^2|\boldsymbol{\gamma})d\sigma^2 \\ &\propto (c_1+1)^{-\frac{1}{2}}(c_2+1)^{-\frac{(q_\gamma-1)}{2}}\left(\frac{\mathrm{SSE}(\boldsymbol{\gamma})}{2(n-q_\gamma)}\right)^{-\frac{n}{2}}(n+\kappa-q_\gamma)^{-\frac{n+\kappa-2}{2}} \times K(\kappa) \end{aligned}$$

where

$$K(\kappa) = (2\pi)^{-n/2}\Gamma(\kappa/2-1)^{-1}\Gamma\left(\frac{n+\kappa-2}{2}\right)\kappa^{(\frac{\kappa}{2}-1)}\,,$$

and Γ is the gamma function.

The posterior density of $\boldsymbol{\gamma}$ is

$$p(\boldsymbol{\gamma}|\boldsymbol{Y}) \propto p(\boldsymbol{Y}|\boldsymbol{\gamma})p(\boldsymbol{\gamma}),$$

and it is this density that is used in the Markov chain Monte Carlo simulation scheme.

2.5. Orthogonalized regressors

Several authors have suggested orthogonalizing the design matrix $\boldsymbol{X}$ to speed up the variable selection computations (e.g. Clyde *et al.*, 1996; Liang, Truong & Wong, 2001). Although there may be some difficulty in interpreting the results of variable selection on orthogonalized variables when the original variables are not orthogonal, we believe that it is of some interest to study the effectiveness of model averaging at estimating the predictive distribution. If orthogonalizing variables gives estimates of the predictive density that are as good or nearly as good as working with the original variables, then orthogonalizing would be the method of choice because of its faster computation. Therefore, for completeness, our empirical work also applies the priors for $\boldsymbol{\beta}, \sigma^2$ and $\boldsymbol{\gamma}$ to the orthogonalized variables when the original variables are not orthogonal.

To be specific, when the original variables are not orthogonal we begin with a design matrix $\boldsymbol{X}$ and transform it to $\boldsymbol{W} = \boldsymbol{X}\boldsymbol{T}$ so that $\boldsymbol{W}$ is a $n \times k$ orthonormal matrix with entries $1/\sqrt{n}$ in the first column. For conciseness, we rewrite the model (1) as

$$\boldsymbol{Y} = \boldsymbol{W}\boldsymbol{\beta} + \boldsymbol{e}\,, \tag{5}$$

noting that the β in (5) is different to that in (1).

2.6. Sampling Scheme

We use the following Gibbs sampler to explore the parameter space of $\boldsymbol{\gamma}$.

Step 0: Randomly choose an initial value $\boldsymbol{\gamma}^{[0]} = (\gamma_1^{[0]}, \ldots, \gamma_k^{[0]})$, with $\gamma_1^{[0]} = 1$.

Step 1: for $g = 1, 2, \ldots$, successively generate $p(\gamma_i | \boldsymbol{Y}, \gamma_{i \neq j})$, $i = 2, \ldots, k$ to obtain $\boldsymbol{\gamma}^{[g]}$. Step 1 is performed a number of times and in two stages. The first stage is a warm up period to allow the sampler to converge to draws from the joint posterior distribution. The second is the sampling period and the values of $\boldsymbol{\gamma}$ generated during this period are used for inference.

In our simulations, iterates of σ^2 and $\boldsymbol{\beta}$, which we write as $(\sigma^2)^{[g]}$ and $\boldsymbol{\beta}^{[g]}$, are also generated concurrently during the sampling period, but are not part of the sampling scheme in the sense that they do not affect the convergence or mixing properties of the Markov chain. The iterates are generated from the conditional densities $p(\sigma^2 | \boldsymbol{Y}, \boldsymbol{\gamma}^{[g]})$ and $p(\boldsymbol{\beta}_\gamma | \boldsymbol{Y}, (\sigma^2)^{[g]}, \boldsymbol{\gamma}^{[g]})$.

3. Simulation study

3.1. Alternative prior

This section uses six examples to compare the predictive performance of the prior in our paper with the following prior, which is based on Kohn *et al.* (2001). The priors for $\boldsymbol{\beta}_\gamma$ and $\boldsymbol{\gamma}$ are the same as (2) and (4). The prior for σ^2 is an inverse gamma, with

$$p(\sigma^2) \propto \left(\sigma^2\right)^{-(1+a_\sigma)} \exp(-b_\sigma/\sigma^2) \tag{1}$$

where $a_\sigma = 10^{-10}$ and $b_\sigma = 0.001$. This is a proper, but uninformative, prior for σ^2, and does not depend on $\boldsymbol{\gamma}$. The prior (1) is very close to the usual noninformative and improper prior for σ^2 , $p(\sigma^2) \propto 1/\sigma^2$, (e.g. George & McCulloch, 1997).

3.2. Loss functions

We use two loss functions to assess the effect of the new data-based prior for the error variance. The Squared Error loss function (SQE) assesses the effect on a point estimate of the regression error variance and the Average Kullback-Leibler Divergence (AKLD) loss function assesses the effect on the whole predictive density. These two loss functions are described now.

1. *The squared error loss function.* Let $\hat{\sigma}^2$ be the posterior mean of σ^2 and σ_T^2 its true value. We define the SQE as

$$SQE = (\hat{\sigma}^2 - \sigma_T^2)^2$$

2. *The Kullback-Leibler divergence.* Suppose that the true model generating the data is

$$y = f_T(\boldsymbol{x}) + e \ , \quad e \stackrel{\mathrm{d}}{=} \mathrm{N}(0, \sigma_T^2) \ ,$$

i.e. $f_T(\boldsymbol{x})$ and σ_T^2 are the true regression function and true error variance. Hence,

$$p_T(y|\boldsymbol{x}) = \frac{1}{\sqrt{2\pi\sigma_T^2}} \exp\Big(-\frac{1}{2\sigma_T^2}\big(y - f_T(\boldsymbol{x})\big)^2 \Big) \ .$$

Let $p(y|\boldsymbol{x}, \boldsymbol{Y})$ be the estimated predictive density of y given $\boldsymbol{x}$ and the data $\boldsymbol{Y}$. The Kullback-Leibler divergence between $p(y|\boldsymbol{x}, \boldsymbol{Y})$ and $p_T(y|\boldsymbol{x})$ for a

given $\boldsymbol{x}$ is defined as (Kullback & Leibler, 1951)

$$KL\big(p(\cdot|\boldsymbol{x},\boldsymbol{Y}),p_T(\cdot|\boldsymbol{x})\big) = \int p_T(y|\boldsymbol{x}) \log\left(\frac{p(y|\boldsymbol{x},\boldsymbol{Y})}{p_T(y|\boldsymbol{x})}\right) dy$$
$$= \frac{1}{\sqrt{\pi}} \int \phi(z) \log\left(\frac{p\big(\sigma_T\sqrt{2}z + f_T(\boldsymbol{x})|\boldsymbol{x},\boldsymbol{Y}\big)}{\phi(z)/\sqrt{2\pi\sigma_T^2}}\right) dz$$

where

$$\phi(z) = \exp(-z^2)$$

and we evaluate this integral using Gauss-Hermite integration over 10 points.

We note that $KL\big(p(\cdot|\boldsymbol{x},\boldsymbol{Y}),p_T(\cdot|\boldsymbol{x})\big) \leq 0$ and that it is equal to 0 if and only if $p(y|\boldsymbol{x},\boldsymbol{Y}) = p_T(y|\boldsymbol{x})$ for all y (Rao, 1973, pp 58-59). Thus, the closer KL is to 0 for a given abscissa $\boldsymbol{x}$, the closer is the predictive density to the true density. We also note that Kullback-Leibler divergence is often defined as the negative of $KL\big(p(\cdot|\boldsymbol{x},\boldsymbol{Y}),p_T(\cdot|\boldsymbol{x})\big)$.

In our simulations, we compute the predictive density and the Kullback-Leibler divergence at a number of abscissae $\boldsymbol{x}_l, l = 1,\ldots,L$, and average the Kullback-Leibler divergences over these abscissae. We write this average as AKLD and it is this value which we use as our loss function to assess performance in terms of the predictive density.

In practice, we cannot compute the predictive density exactly because it is necessary to average over all possible models $\boldsymbol{\gamma}$ and this is infeasible if the number of variables k is moderate to large. Instead, we estimate the predictive density from the Markov chain Monte Carlo output as follows. Let $\boldsymbol{\beta}^{[j]}, {\sigma^2}^{[j]}, \boldsymbol{\gamma}^{[j]}, j = 1,\ldots,M$ be the iterates of $\boldsymbol{\beta}, \sigma^2$ and $\boldsymbol{\gamma}$ during the sampling period. Then,

$$p(y|\boldsymbol{x},\boldsymbol{Y}) = \sum_{\boldsymbol{\gamma}} \iint p(y|\boldsymbol{x},\boldsymbol{\beta},\sigma^2,\boldsymbol{\gamma}) p(\boldsymbol{\beta},\sigma^2,\boldsymbol{\gamma}|\boldsymbol{Y}) d\boldsymbol{\beta}_\gamma d\sigma^2$$
$$\approx \frac{1}{M}\sum_{j=1}^{M} p(y|\boldsymbol{x},\boldsymbol{\beta}^{[j]},(\sigma^2)^{[j]},\boldsymbol{\gamma}^{[j]})$$
$$= \frac{1}{M}\sum_{j=1}^{M} \frac{1}{\big(2\pi(\sigma^2)^{[j]}\big)^{\frac{1}{2}}} \exp\left(-\frac{1}{2(\sigma^2)^{[j]}}(y - \boldsymbol{x}'\boldsymbol{\beta}_\gamma^{[j]})^2\right).$$

To compare two models, which we call Models A and B for convenience, in terms of a certain loss function, $LOSS$, with Model A treated as the base model, we use the percentage change in the $LOSS$ in going from A to B, i.e.,

$$D(A,B) = \frac{LOSS(B) - LOSS(A)}{LOSS(A)} \times 100. \tag{2}$$

If $D(A,B) > 0$ then Model A outperforms Model B and if $D(A,B) < 0$ then Model B outperforms Model A.

3.3. Regression test functions

Six linear regression functions were considered in the study. For each regression function a new design matrix was generated similar to the original design matrix used to construct the sample data in order to evaluate the AKLD. In all the examples the first column of the design matrix was a vector of ones.

Example 1: The first regression function is similar to that used by Raftery *et al.* (1997) and Fernandez, Ley & Steel (2001). A 50×16 design matrix was generated as follows. Columns 2 to 11 were generated from independent standard normal distributions. The last 5 columns were constructed as linear combinations of $(\boldsymbol{X}_2, \ldots, \boldsymbol{X}_6)$ with noise, i.e.,

$$(\boldsymbol{X}_{12}, \ldots, \boldsymbol{X}_{16}) = (\boldsymbol{X}_2, \ldots, \boldsymbol{X}_6) \times (0.3, 0.5, 0.7, 0.9, 1.1)' \times (1, 1, 1, 1, 1) + \boldsymbol{E},$$

where $\boldsymbol{E}$ is an 50×5 matrix drawn from independent standard normal distributions. The generation of the last five columns resulted in moderate correlation between the groups $(\boldsymbol{X}_2, \ldots, \boldsymbol{X}_6)$ and $(\boldsymbol{X}_{12}, \ldots, \boldsymbol{X}_{16})$ and substantial correlation within $(\boldsymbol{X}_{12}, \ldots, \boldsymbol{X}_{16})$.

The response $\boldsymbol{Y}$ was generated as

$$\boldsymbol{Y} = 4\boldsymbol{X}_1 + 2\boldsymbol{X}_2 - \boldsymbol{X}_6 + 1.5\boldsymbol{X}_8 + \boldsymbol{X}_{12} + 0.5\boldsymbol{X}_{14} + \boldsymbol{e}, \;\; \boldsymbol{e} \stackrel{\mathrm{d}}{=} \mathrm{N}(\boldsymbol{0}, 2.5^2\boldsymbol{I}) \,.$$

Example 2: The second regression function was constructed by generating a 10×9, design matrix, $\boldsymbol{X}$. Columns 2 to 9 were drawn from independent standard normal deviates. The response $\boldsymbol{Y}$ was generated as

$$Y = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}, \;\; \boldsymbol{e} \stackrel{\mathrm{d}}{=} \mathrm{N}(\boldsymbol{0}, 2.5^2\boldsymbol{I}) \,,$$

where the first 8 elements of $\boldsymbol{\beta}$ were generated as independent uniform random variables on (-5,5) and the 9th element was zero. The maximum and minimum absolute values of the first 8 elements of the generated $\boldsymbol{\beta}$ were 4.11 and 0.53.

Example 3: The third regression function was constructed by generating a 50×41 design matrix $\boldsymbol{X}$. Columns 2 to 41 were drawn from independent standard normal deviates. The response $\boldsymbol{Y}$ was generated as

$$Y = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}, \;\; \boldsymbol{e} \stackrel{\mathrm{d}}{=} \mathrm{N}(\boldsymbol{0}, 2.5^2\boldsymbol{I}) \,,$$

where the first 31 elements of $\boldsymbol{\beta}$ were generated as independent uniform random variables on (-5,5), and the remaining elements of $\boldsymbol{\beta}$ were zero. The maximum and minimum absolute values of the first 31 elements of the generated $\boldsymbol{\beta}$ were 4.90 and 0.14.

Example 4: The fourth regression function was constructed by generating a 100×80 design matrix $\boldsymbol{X}$. Columns 2 to 80 were drawn from independent standard normal deviates. The response $\boldsymbol{Y}$ was generated as

$$Y = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}, \;\; \boldsymbol{e} \stackrel{\mathrm{d}}{=} \mathrm{N}(\boldsymbol{0}, 2.5^2\boldsymbol{I}) \,,$$

where the first 71 elements of $\boldsymbol{\beta}$ were generated as independent uniform random variables on (-5,5), and the remaining elements of $\boldsymbol{\beta}$ were zero. The maximum and minimum absolute values of the first 71 elements of the generated $\boldsymbol{\beta}$ were 4.94 and 0.22.

Example 5:

The fifth regression function was constructed by generating a 50×22 design matrix $\boldsymbol{X}$. Columns $\boldsymbol{X}_2, \ldots, \boldsymbol{X}_7$ were dummy variables where $p(x_{ij} = 1) = 0.5$ for $j = 2, \ldots, 7$, $i = 1, \ldots, 50$ and x_{ij} is the ith element of column $\boldsymbol{X}_j$. Columns $\boldsymbol{X}_8, \ldots, \boldsymbol{X}_{22}$ represent all the possible two-way interaction effects between $\boldsymbol{X}_2, \ldots, \boldsymbol{X}_7$. The response $\boldsymbol{Y}$ was generated as

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}, \;\; \boldsymbol{e} \stackrel{\mathrm{d}}{=} \mathrm{N}(\boldsymbol{0}, 2.5^2\boldsymbol{I}),$$

where $\beta_i = 0$ for $i = 6, 7, 10, 11, 12, 14, 15, 16, 18, 19, 20, 21$, and the remaining elements of $\boldsymbol{\beta}$ were generated uniformly on the interval (-10,10). The maximum and minimum absolute values of the non-zero elements of the generated $\boldsymbol{\beta}$ were 9.55 and 0.58.

Example 6: We follow the work of Kohn *et al.* (2001) and estimate a surface nonparametrically. Consider the bivariate surface described by the mixture of normal densities,

$$f(\boldsymbol{x}) = 1 + \mathrm{N}(\boldsymbol{x}; \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) + \mathrm{N}(\boldsymbol{x}; \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2) \,,$$

where $\mathrm{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the bivariate normal density with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, evaluated at the abscissa x. Following Kohn *et al.* (2001), we took $\boldsymbol{\mu}_1 = (0.25, 0.75)'$, $\boldsymbol{\mu}_2 = (0.75, 0.25)'$,

$$\boldsymbol{\Sigma}_1 = \begin{pmatrix} 0.05 & 0.01 \\ 0.01 & 0.05 \end{pmatrix} \quad \text{and} \quad \boldsymbol{\Sigma}_2 = \begin{pmatrix} 0.1 & 0.01 \\ 0.01 & 0.1 \end{pmatrix}$$

and generated

$$Y = f(\boldsymbol{x}) + e, \;\; e \stackrel{\mathrm{d}}{=} \mathrm{N}(0, 2.5^2).$$

The test data and knot selection are described in subsection 3.4. below.

3.4. Knot Selection and Basis Function Construction

Before presenting the empirical results we first describe the construction of the basis functions and an easy and efficient method of knot selection, which avoids the clustering approach in Kohn *et al.* (2001).

Knot selection in the bivariate case is more difficult than in the univariate case because the abscissae cannot be ordered. Kohn *et al.* (2001) use a clustering algorithm to choose abscissae, whereas Holmes & Mallick (1998) use all the abscissae. However, clustering algorithms can be computationally expensive, while choosing all abscissae may result in an unnecessarily large number of knots. We now outline a knot selection strategy based on dividing the predictor space into cells. Without loss of generality, suppose that the abscissae all lie in the unit square, and that the sample size is n, so that there are n abscissae. We partition the unit square into intervals of area δ^2, giving a grid of squares labeled Δ_i, $i = 1, \ldots, \frac{1}{\delta^2}$. For each square Δ_i, we take its midpoint as a knot if Δ_i contains at least one abscissa. No other knots are selected and redundant knots are omitted.

For the simulated data from regression function 6, we generated $n = 50$ abscissae in the unit square. Figure 1 shows bivariate data generated over the unit square for $n = 50$ with the knots selected using $\delta = 0.1$, giving 100 squares, and 39 knots were selected. It is clear from Figure 1 that the proposed knot selection method provides good coverage of the abscissae. We also generated 50 test data points from a uniform distribution on the interval $[0.05, 0.95]^2$. The knots were selected only on the basis of the sample data.

Once the knots were selected, a bivariate thin plate spline basis was constructed as follows. Let the collection of knots selected by the above method be denoted as $(\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_l)$, and let $\boldsymbol{X}$ be the covariate matrix. Then, for regression function 6, the ith row of $\boldsymbol{X}$ is

$$\boldsymbol{X}_{i.} = \left(1, x_{i1}, x_{i2}, ||\boldsymbol{x}_i - \boldsymbol{\zeta}_1||^2 \log(||\boldsymbol{x}_i - \boldsymbol{\zeta}_1||^2), \ldots, ||\boldsymbol{x}_i - \boldsymbol{\zeta}_l||^2 \log(||\boldsymbol{x}_i - \boldsymbol{\zeta}_l||^2)\right) ,$$

where $\boldsymbol{x}_i = (x_{i1}, x_{i2})$ is the ith abscissa. For the test data the thin plate splines were

$$\boldsymbol{X}_{g.} = \left(1, x_{g1}, x_{g2}, ||\boldsymbol{x}_g - \boldsymbol{\zeta}_1||^2 \log(||\boldsymbol{x}_g - \boldsymbol{\zeta}_1||^2), \ldots, ||\boldsymbol{x}_g - \boldsymbol{\zeta}_l||^2 \log(||\boldsymbol{x}_g - \boldsymbol{\zeta}_l||^2)\right) ,$$

where $\boldsymbol{x}_g = (x_{g1}, x_{g2})$ is the gth abscissa for the test data.

3.5. Simulation results

For each of the simulated regressions we examined the performance of the priors (3) and (1) for σ^2 when the design matrix was non-orthogonalized and when the design matrix was orthogonalized. We also compared the results of the orthogonalized design matrix versus the non-orthogonalized design matrix. For each comparison we examined the percentage difference in the AKLD, D, defined by (2). We note that in (2), if Model A outperforms Model B then $D > 0$. Details of the exact models constructed for comparison are given below. We include boxplots of D for 50 replications of each simulated example and a one-sided hypothesis test such that

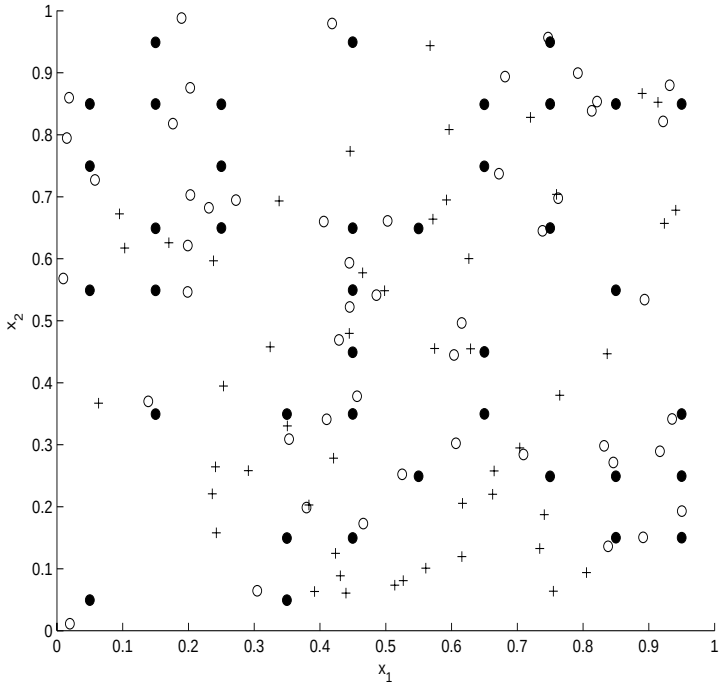

Figure 1. Plot of the sample data, test data and knots selected using $\delta = 0.1$. The sample data are the empty dots, the test data are the crosses and the knots are the filled dots.

$$\mathrm{H}_0 : E(D) = 0 \text{ and } \mathrm{H}_\mathrm{A} : E(D) > 0. \tag{3}$$

3.5.1. Comparison of Priors: original design matrix

This section compares the two priors for σ^2 described in the paper for the non-orthogonalized design matrix. When calculating the loss function D the prior for σ^2 is (3) for Model A and (1) for Model B. The prior for $\boldsymbol{\beta}$ is (2) and the prior for $\boldsymbol{\gamma}$ is (4). Figure 2 presents the boxplots of D for all six examples for the two loss functions. Table 1 contains the p-values of the t-statistics for the hypothesis (3) for the two loss functions. For the AKLD loss function, Figure 2 shows that the prior (3) outperformed the prior (1) for all the examples. Table 1 shows that the p-values for the null hypothesis that $E(D) = 0$ were less than 5% for all the examples except example 2, where the null hypothesis could not be rejected. For the SQE loss function, Figure 2 shows that the prior (3) outperformed (1) for examples 2 to 6, and particularly for examples 2 to 4 where the number of predictors is large

relative to the sample size. Table 1 shows that the p-values for the null hypothesis that $E(D) = 0$ were less than 0.05 for examples 2 to 4, less than 0.10 for examples 5 and 6, and the null hypothesis could not be rejected for example 1. We conclude that, for the non-orthogonalized case, the prior proposed in our article outperformed the prior specified by Kohn *et al.* (2001) for estimating the predictive density and the error variance.

Table 1

P-values for the hypothesis (3) for the non-orthogonal case for the AKLD loss function (top row) and the SQE loss function (bottom row). The prior for σ^2 is (3) for Model A and (1) for Model B

Example	1	2	3	4	5	6
AKLD P-value	0.06	0.20	0.00	0.00	0.00	0.00
SQE P-value	0.14	0.04	0.00	0.00	0.09	0.06

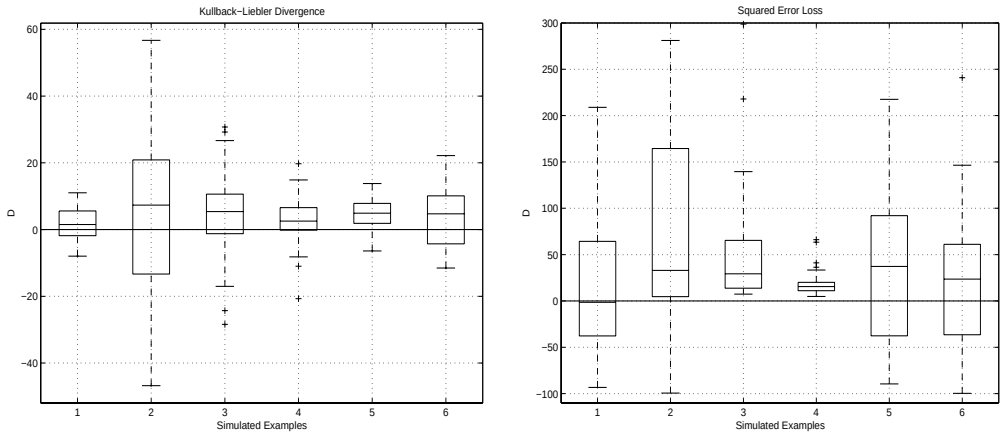

Figure 2. The non-orthogonal case. The left panel shows the results for the AKLD loss function and the right panel shows the results for the SQE loss function. From left to right in both panels the boxplots of D are for examples 1 to 6 respectively. The prior for σ^2 is (3) for Model A and (1) for Model B.

3.5.2. Comparison of Priors: othogonalized design matrix

This section compares the priors (3) and (1) for the case of orthogonalized variables, with model A corresponding to (3) and model B to (1). The priors for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are (2) and (4) for both Model A and Model B. Figure 3 presents the boxplots of D for all six examples for the two loss functions. Table 2 contains the p-values of the statistics for the hypothesis (3) for the two loss functions. For the AKLD loss function, Figure 3 shows that the prior (3) outperformed the prior (1) for examples 2, 3, 4 and 6 and the results were similar for examples 1 and 5. Table 2 shows that the p-values for the null hypothesis that $E(D) = 0$ were less than 0.05 for examples 2, 3, 4 and 6 and the null hypothesis could not be rejected for examples 1 and 5. For the SQE loss function, Figure 3 shows that the prior (3) outperformed (1) for examples 2, 3, 4 and 6 but the prior (1) outperformed the prior (3) for examples 1 and 5. Table 2 shows that the p-values for the null hypothesis that $E(D) = 0$ were less than 0.05 for examples 2, 3, 4 and 6 and the null hypothesis could not be rejected for examples 1 and 5. Again, the improvement was most significant in examples 2, 3 and 4 where the number of predictors was large relative to the sample size. We conclude that, for the orthogonalized case, the prior proposed in our article outperformed the prior specified by Kohn *et al.* (2001) for estimating the predictive density as well as the error variance when the number of predictors was large relative to sample size.

Table 2

P-values for the hypothesis (3) for the orthogonal case for the AKLD loss function (top row) and the SQE loss function (bottom row). The prior for σ^2 is (3) for Model A and (1) for Model B

Example	1	2	3	4	5	6
AKLD P-value	0.60	0.00	0.00	0.00	0.19	0.00
SQE P-value	0.34	0.00	0.00	0.00	0.35	0.02

3.5.3. Comparison of original versus orthogonalized design matrix

This section compares the results of using the original design matrix versus the orthogonalized design matrix for examples 1, 5 and 6, where the covariates were correlated. We did not consider examples 2 to 4 because for

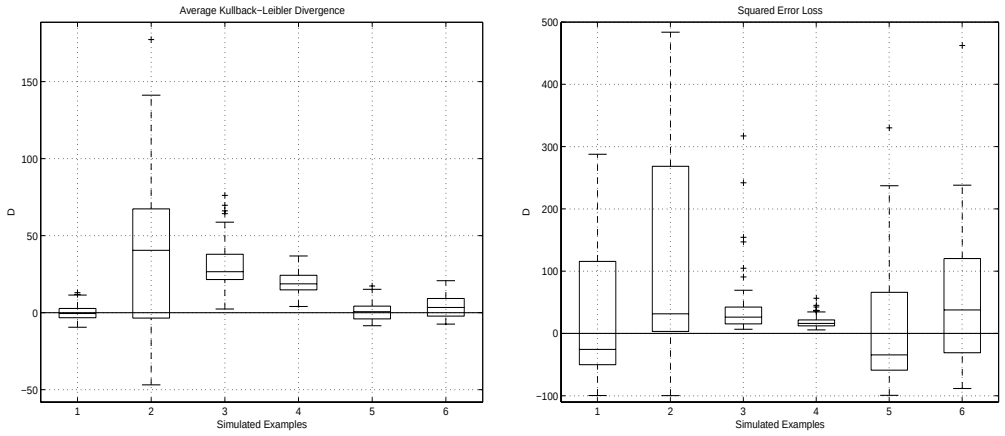

Figure 3. The orthogonal case. The left panel shows the results for the AKLD loss function and the right panel shows the results for the SQE loss function. From left to right in both panels the boxplots of D are for examples 1 to 6 respectively. The prior for σ^2 is (3) for Model A and (1) for Model B.

these examples the columns were generated independently and were almost orthogonal. The results are reported in terms of the AKLD loss function only. Figure 4 presents the boxplots of the percentage difference in AKLD defined in (2), i.e. D, between the orthogonalized and non-orthogonalized cases, for the two priors for σ^2. For each boxplot, Model A corresponds to the non-orthogonalized case and Model B to the orthogonalized case, and the same prior for σ^2 was used for both models. The figure shows that the non-orthogonal case outperformed the orthogonal case for all three examples so that if speed is not a consideration, then the original (non-orthogonalized) design matrix should be used. Testing hypothesis (3) gave p-values that were approximately zero. We conclude that when the covariates are correlated it is better not to orthogonalize if performance is judged by the predictive density.

3.6. Comparison of priors using a real example

We now compare the two priors for σ^2 discussed in our article on the real data described in Vandaele (1978). This data aims to describe how the crime rate depends on the other variables provided in the study. There are 47 observations in the study and the variables are:

1. Crime: rates of crime in a particular category per head of population
2. M: The percentage of males aged 14-24

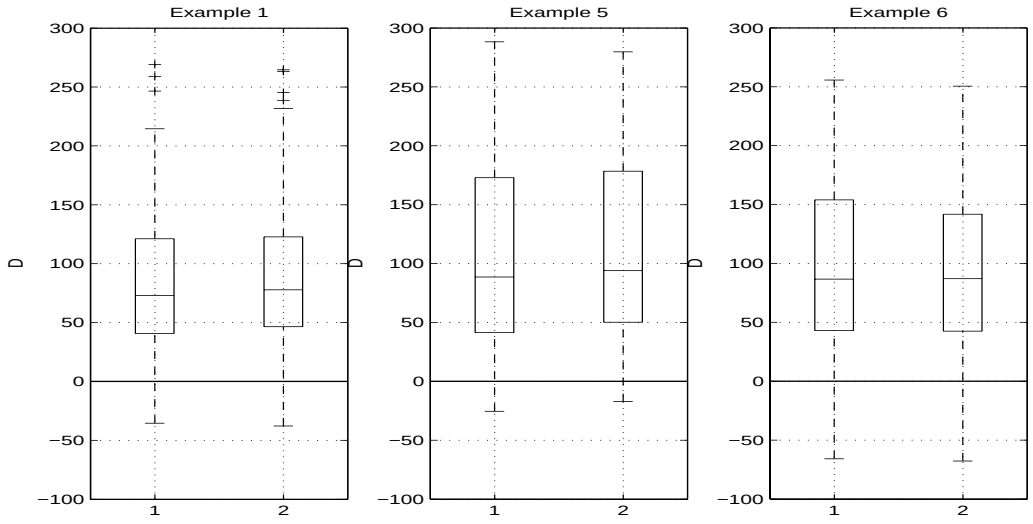

Figure 4. The orthogonal vs non-orthogonal case. Boxplots of D. In each boxplot, Model A corresponds to the non-orthogonal case and Model B to the orthogonal case, and the priors for Models A and B are the same. From left to right the panels correspond to examples 1, 5 and 6 respectively. In each panel the prior for σ^2 is (1) for the left boxplot, indicated by 1 on the horizontal axis, and (3) for the right boxplot, indicated by 2 on the horizontal axis. For each boxplot, values of D greater than 300% are omitted so as not to distort the rest of the boxplot.

3. S: Indicator variable for Southern states (0 = No, 1 = Yes)
4. Ed: Mean number of years of schooling
5. Po1: Police expenditure in 1960
6. Po2: Police expenditure in 1959
7. LF: Labor force participation rate
8. MF: The number of males per 1000 females
9. Pop: State population
10. NW: The number of non-whites per 1000 people

11. U1: Unemployment rate of urban males of age 14-24
12. U2: Unemployment rate of urban males of age 35-39
13. GDP: Gross domestic product per head
14. Ineq: Income inequality
15. Prob: Probability of imprisonment
16. Time: Average time served in state prisons

We assumed that the relationship between the dependent variable and the predictor variables could be described by a linear regression with Gaussian errors. The predictor variables Po1, Po2, Pop and NW were right skewed and we took the natural logarithms of these variables. In the real example the true error variance and the true predictive density were unknown, so we used the following strategy to mimic the approach taken in the simulated examples. We fitted a regression model to the data using the priors (2), (3) and (4) and identified the model having the highest marginal likelihood observed in the MCMC run. We then treated the estimated values of $\boldsymbol{\beta}$ and σ^2 associated with this model as the true parameter values and denoted these estimates as $\boldsymbol{\beta}_*$ and σ^2_* respectively. The dependent variable Y was generated as

$$Y = \boldsymbol{X}\boldsymbol{\beta}_* + e, \;\; e \stackrel{\mathrm{d}}{=} \mathrm{N}(0, \sigma^2_*). \tag{4}$$

Twenty four training data points and 23 test data points were generated for each of 50 replications. For each replication the parameters in (4) were estimated using the prior (3) for Model A and (1) for Model B. For both models the priors for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ were (2) and (4). Figure 5 shows the boxplots of D for the AKLD and SQE loss functions. The figure suggests that the prior (3) for σ^2 outperformed the prior (1) under both loss functions. We also tested the null hypothesis $E(D) = 0$ against the alternative that $E(D) > 0$. The p-values for the AKLD and SQE loss functions were both less than 2%. We concluded for this example that the prior for σ^2 in this paper outperformed the prior specified in Kohn *et al.* (2001) both in terms of estimating the predictive density and for point estimates of the error variance.

4. Conclusions

Our article presents a data-based prior for the error variance in a linear regression model with Gaussian errors, when that model is used for Bayesian variable selection and model averaging. The new prior is centred at an unbiased estimate of σ^2 given the variables included in the model and is made

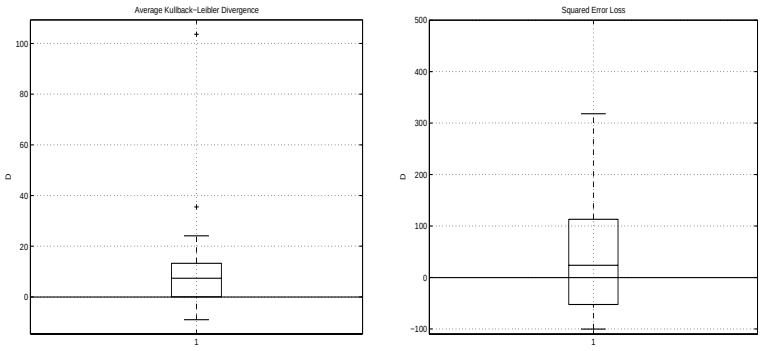

Figure 5. The US Crime Data Set. The boxplot of D. Model A has the prior (3) for σ^2. Model B has the prior (1) for σ^2.

suitably uninformative. The prior is simple to use and largely avoids the specification of arbitrary hyperparameters as in equation (1). The performance of the new prior compared favorably on simulated data to the traditional uninformative prior for σ^2, e.g. equation (1), when the loss function is the Kullback-Leibler divergence between the estimated predictive density and the true density and also when the loss function is the squared error loss between the posterior mean of the error variance and the true error variance. This was true in particular when the number of predictors in the regression was large compared to the sample size.

The article makes two further contributions. First, the simulation results suggested that it is always better to work with the original variables when carrying out variable selection, rather than orthogonalizing them. Second, a novel technique for knot selection in nonparametric regression problems is also presented that is simple, covers the observed predictor space effectively and avoids the need for clustering algorithms.

Acknowledgement

We thank a referee, an associate editor and the technical editor for improving the clarity of the presentation. The work of all authors was supported by an Australian Research Council grant on mixture models. The work of Cripps was also supported by an Australian Postgraduate Award.

References

Clyde, M., Desimone, M. & Parmigiani, G. (1996). Prediction via

orthogonalized model mixing. *J. Amer. Statist. Assoc.* **91**, 1197–1208.

Fernandez, C., Ley, E. & Steel, M.F.J. (2001). Benchmark priors for Bayesian model averaging. *J. Econometrics* **100**, 381–427.

George, E.I. & McCulloch, R.E. (1993). Variable Selection via Gibbs Sampling. *J. Amer. Statist. Assoc.* **88**, 881–889.

George, E.I. & McCulloch, R.E. (1997). Approaches for Bayesian Variable Selection. *Statistica Sinica* **7**, 339–373.

Hoeting, J.A., Madigan, D., Raftery, A.E. & Volinsky, C.T. (1999). Bayesian Model Averaging: A Tutorial (with discussion). *Statist. Sci.* **14**, 382–417. Corrected version at http://www.stat.washington.edu/www/research/online/hoeting1999.pdf.

Holmes, C.C. & Mallick, B.K. (1998). Radial basis functions of variable dimension. *Neural Computation* **10**, 1217–1233.

Kohn, R., Smith, M. & Chan, D. (2001). Nonparametric regression using linear combinations of basis functions. *Statistics and Computing* **11**, 313–322.

Kullback, S. & Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics* **22**, 79–86.

Liang, F., Truong, Y. K. & Wong, W. H. (2001). Automatic Bayesian model averaging for linear regression and application in Bayesian curve fitting. *Statistica Sinica* **11**, 1005–1029.

Mitchell, T.J. & Beauchamp, J.J. (1988). Bayesian variable selection in linear regression. *J. Amer. Statist. Assoc.* **83**, 1023–1036.

Raftery, A.E., Madigan, D.M. & Hoeting, J.A. (1997). Bayesian Model Averaging for Linear Regression Models. *J. Amer. Statist. Assoc.* **94**, 179–191.

Rao, C. R. (1973). *Linear statistical inference and its applications*, 2nd edn. New York: John Wiley.

Smith, M. & Kohn, R. (1996). Nonparametric regression using Bayesian variable selection. *J. Econometrics* **75**, 317–342.

Vandaele, W. (1978). Participation in illegitimate activities; Ehrlich revisited. In *Deterrence and Incapacitation*, eds Blumstein, A., Cohen, J., Nagin, D., pp 270-335. Washington, DC: National Academy of Sciences Press.