Computational Statistics and Data Analysis 54 (2010) 3227–3241

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Bayesian projection approaches to variable selection in generalized linear models

David J. Nott *, Chenlei Leng

Department of Statistics and Applied Probability, National University of Singapore, Singapore 117546, Singapore

ARTICLE INFO

Article history:
Received 29 January 2009
Received in revised form 29 January 2010
Accepted 29 January 2010
Available online 10 February 2010

Keywords:
Bayesian variable selection
Kullback–Leibler projection
Lasso
Non-negative garotte
Preconditioning

ABSTRACT

A Bayesian approach to variable selection which is based on the expected Kullback–Leibler divergence between the full model and its projection onto a submodel has recently been suggested in the literature. For generalized linear models an extension of this idea is proposed by considering projections onto subspaces defined via some form of L_1 constraint on the parameter in the full model. This leads to Bayesian model selection approaches related to the lasso. In the posterior distribution of the projection there is positive probability that some components are exactly zero and the posterior distribution on the model space induced by the projection allows exploration of model uncertainty. Use of the approach in structured variable selection problems such as ANOVA models is also considered, where it is desired to incorporate main effects in the presence of interactions. Projections related to the non-negative garotte are able to respect the hierarchical constraints. A consistency result is given concerning the posterior distribution on the model induced by the projection, showing that for some projections related to the adaptive lasso and non-negative garotte the posterior distribution concentrates on the true model asymptotically.

1. Introduction

Bayesian approaches to model selection and describing model uncertainty have become increasingly popular in recent years. In this paper we extend a method of variable selection considered by Dupuis and Robert (2003) and related to earlier suggestions by Goutis and Robert (1998) and Mengersen and Robert (1996). Dupuis and Robert (2003) consider an approach to variable selection where models are selected according to a relative explanatory power, where relative explanatory power is defined using the expected Kullback–Leibler divergence between the full model and its projection onto a submodel. The goal is to find the most parsimonious model which achieves an acceptable loss of explanatory power compared to the full model.

We consider an extension of the method of Dupuis and Robert (2003) where instead of considering a projection onto a subspace defined by a set of active covariates we consider projection onto a subspace defined by some other form of constraint on the parameter in the full model. Certain choices of the constraint (an L_1 constraint as used in the lasso of Tibshirani (1996), for example) lead to exact zeros for some of the coefficients. The kinds of projections we consider also have computational advantages, in that parsimony is controlled by a single continuous parameter and we avoid the search over a large and complex model space. Searching the model space in traditional Bayesian model selection approaches with large numbers of covariates is a computationally daunting task. In contrast, with our method we handle model uncertainty in a continuous way through an encompassing model, and can exploit existing fast lasso type algorithms to calculate

* Corresponding author. Tel.: +65 6516 2744; fax: +65 6872 3919.
E-mail addresses: standj@nus.edu.sg (D.J. Nott), stalc@nus.edu.sg (C. Leng).

0167-9473/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2010.01.036

projections for samples from the posterior distribution in the encompassing model. This allows sparsity and exploration of model uncertainty while preserving approximately posterior predictive behaviour based on the full model. Furthermore, the method is easy to implement given a combination of existing Bayesian software and software implementing lasso type fitting methods. While the method is more computationally intensive than calculating a solution path for a classical shrinkage approach like the lasso, this is the price to be paid for exploring model uncertainty and the computational demands of the method are certainly less than Bayesian approaches which search the model space directly. Our idea can also be applied to structured variable selection problems such as those arising in ANOVA models where we might wish to include interaction terms only in the presence of the corresponding main effects. A plug-in version of our approach is also related to the preconditioning method of Paul et al. (2008) for feature selection in "large p, small n" regression problems. A key advantage of our approach is simplicity in prior specification. One is only required to specify a prior on the parameter in the full model, and not a prior on the model space or a prior on parameters for every submodel. Nevertheless, the posterior distribution on the model space induced by the projection can be used in a similar way to the posterior distribution on the model space in a traditional Bayesian analysis for exploring model uncertainty and different interpretations of the data.

There are many alternative Bayesian strategies for model selection and exploring model uncertainty to the one considered here. Bayes factors and Bayesian model averaging (Kass and Raftery, 1995; Hoeting et al., 1999; Fernández et al., 2001) are the traditional approaches to addressing issues of model uncertainty in a Bayesian framework. As already mentioned, prior specification can be very demanding for these approaches, although general default prior specifications have been suggested (Berger and Pericchi, 1996; O'Hagan, 1995). In our later examples we focus on generalized linear models, and Raftery (1996) suggests some reference priors for Bayesian model comparison in this context. Structuring priors hierarchically and estimating hyperparameters in a data driven way is another way to reduce the complexity of prior specification, and this can work well (George and Foster, 2000). Various Bayesian predictive criteria for selection have also been suggested (Laud and Ibrahim, 1995; Gelfand and Ghosh, 1998; Spiegelhalter et al., 2002). These approaches generally do not require specification of a prior on the model—however, prior specification for parameters in all models is still required and this can be quite demanding if there are a large number of models to be compared. Decision theoretic strategies which attempt to take account of the costs of data collection for covariates have also been considered (Lindley, 1968; Brown et al., 1999; Draper and Fouskakis, 2000). Bayesian model averaging can also be combined with model selection as in Brown et al. (2002). The projection method of Dupuis and Robert (2003) that we extend here is related to the Bayesian reference testing approach of Bernardo and Rueda (2002) and the predictive method of Vehtari and Lampinen (2004). There are also less formal approaches to Bayesian model comparison including posterior predictive checks (Gelman et al., 1996) which are targeted according to the uses that will be made of a model. We see one application of the methods we describe here as being to suggest a small set of candidate simplifications of the full model which can be examined by such means as to their adequacy for specific purposes.

The projection methods we use here for model selection are related to the lasso of Tibshirani (1996) and its many later extensions. There has been some recent work on incorporating the lasso into Bayesian approaches to model selection. Tibshirani (1996) pointed out the Bayesian interpretation of the lasso as a posterior mode estimate in a model with independent double exponential priors on regression coefficients. Park and Casella (2008) consider the Bayesian lasso, where estimators other than the posterior mode are considered—their estimators do not provide automatic variable selection via the posterior mode but convenient computation and inference are possible within their framework. Yuan and Lin (2005) consider a hierarchical prior formulation in Bayesian model comparison and a certain analytical approximation to posterior probabilities connecting the lasso with the Bayes estimate. Recently Griffin and Brown (2007) have also considered alternatives to double exponential prior distributions on the coefficients to provide selection approaches related to the adaptive lasso of Zou (2006).

The structure of the paper is as follows. In the next section we briefly review the method of Dupuis and Robert (2003) and consider our extension of their approach. Computational issues and predictive inference are considered in Section 3, and then a consistency result relating to the posterior distribution on model space induced by the projection is proved in Section 4. We describe applications to structured variable selection in Section 5 and Section 6 considers connections between our method and the preconditioning approach to selection of Paul et al. (2008) in "large p, small n" regression problems. Section 7 considers some examples and simulation studies and Section 8 concludes.

2. Projection approaches to model selection

2.1. Method of Dupuis and Robert

Dupuis and Robert (2003) consider a method of model selection based on the Kullback–Leibler divergence between the true model and its projection onto a submodel. Suppose we are considering a problem of variable selection in regression, where M_F denotes the full model including all covariates and M_S is a submodel with a reduced set of covariates. Write $f(\boldsymbol{y}|\boldsymbol{\theta}_F, M_F)$ and $f(\boldsymbol{y}|\boldsymbol{\theta}_S, M_S)$ for the corresponding likelihoods of the models with parameters $\boldsymbol{\theta}_F$ and $\boldsymbol{\theta}_S$. Let $\boldsymbol{\theta}'_S = \boldsymbol{\theta}'_S(\boldsymbol{\theta}_F)$ be the projection of $\boldsymbol{\theta}_F$ onto the submodel M_S. That is, $\boldsymbol{\theta}'_S$ is the value for $\boldsymbol{\theta}_S$ for which $f(\boldsymbol{y}|\boldsymbol{\theta}_S, M_S)$ is closest in Kullback–Leibler divergence to $f(\boldsymbol{y}|\boldsymbol{\theta}_F, M_F)$, so that

$$\boldsymbol{\theta}'_S = \arg\min_{\boldsymbol{\theta}_S} \int \log \frac{f(\boldsymbol{x}|\boldsymbol{\theta}_F, M_F)}{f(\boldsymbol{x}|\boldsymbol{\theta}_S, M_S)} f(\boldsymbol{x}|\boldsymbol{\theta}_F, M_F) \mathrm{d}\boldsymbol{x}.$$

Let

$$\delta(M_S, M_F) = \int\int \log \frac{f(\boldsymbol{x}|\boldsymbol{\theta}_F, M_F)}{f(\boldsymbol{x}|\boldsymbol{\theta}'_S, M_S)} f(\boldsymbol{x}|\boldsymbol{\theta}_F, M_F) \mathrm{d}\boldsymbol{x} p(\boldsymbol{\theta}_F|y) \mathrm{d}\boldsymbol{\theta}_F,$$

be the posterior expected Kullback–Leibler divergence between the full model and its Kullback–Leibler projection onto the submodel M_S. The relative loss of explanatory power for M_S is

$$d(M_S, M_F) = \frac{\delta(M_S, M_F)}{\delta(M_0, M_F)},$$

where M_0 denotes the model with no covariates and Dupuis and Robert (2003) suggest model selection by choosing the subset model most parsimonious for which $d(M_S, M_F) < c$ where c is an appropriately small constant. If there is more than one model of the minimal size satisfying the bound, then the one with the smallest value of $\delta(M_S, M_F)$ is chosen. Dupuis and Robert (2003) show that $\delta(M_0, M_F)$ can be interpreted as measuring the explanatory power of the full model, and using an additivity property of projections they show that $d(M_S, M_F) < c$ guarantees that our chosen submodel S has explanatory power at least $100(1 - c)\%$ of the explanatory power of the full model. This interpretation is helpful in choosing c. For a predictive quantity Δ we further suggest approximating the predictive density $p(\Delta|\boldsymbol{y})$ for a chosen subset model S by

$$p_S(\Delta|\boldsymbol{y}) = \int p(\Delta|\boldsymbol{\theta}'_S, \boldsymbol{y}) p(\boldsymbol{\theta}'_S|\boldsymbol{y}) \mathrm{d}\boldsymbol{\theta}'_S, \tag{1}$$

where $p(\boldsymbol{\theta}'_S|\boldsymbol{y})$ is the posterior distribution of $\boldsymbol{\theta}'_S$. For example, suppose that we are considering a linear model with $\boldsymbol{\theta}_F = (\boldsymbol{\beta}_F, \sigma_F^2)$ where $\boldsymbol{\beta}_F$ denotes coefficients and σ_F^2 is the error variance and that $\boldsymbol{\theta}'_S = (\boldsymbol{\beta}'_S, \sigma_S^{2'})$ is the projection of $\boldsymbol{\theta}_F$ onto M_S. Then if Δ is a future (as yet unobserved) response for which we know the covariates $\boldsymbol{x}$ and if $\boldsymbol{x}_S$ denotes the subset of these covariates active in model M_S then

$$p(\Delta|\boldsymbol{\theta}'_S, \boldsymbol{y}) = N(\boldsymbol{x}_S^T \boldsymbol{\beta}'_S, \sigma_S^{2'}).$$

For other predictive quantities of interest $p(\Delta|\boldsymbol{\theta}'_S, \boldsymbol{y})$ will be different, and calculation of this conditional density might involve some analytical work or even numerical simulation.

2.2. Extension of the method

To be concrete suppose we are considering variable selection for generalized linear models. Although the ideas we describe are quite general, the computational advantages of focusing on the case of generalized linear models will become apparent later and this is essential to the practicality of our approach. Write $y_1, \ldots, y_n$ for the responses with $E(y_i) = \mu_i$ and suppose that each y_i has a distribution from the exponential family

$$f\left(y_i; \theta_i, \frac{\phi}{A_i}\right) = \exp\left(\frac{y_i\theta_i - b(\theta_i)}{\phi/A_i} + d\left(y_i; \frac{\phi}{A_i}\right)\right),$$

where $\theta_i = \theta_i(\mu_i)$ is the natural parameter, ϕ is a scale parameter, the A_i are known weights and $b(\cdot)$ and $d(\cdot)$ are known functions. For a smooth invertible link function $g(\cdot)$ we have $\eta_i = g(\mu_i) = \boldsymbol{x}_i^T\boldsymbol{\beta}$ where $\boldsymbol{x}_i$ is a p-vector of covariates and $\boldsymbol{\beta}$ is a p-dimensional parameter vector. Writing $\boldsymbol{X}$ for the design matrix with ith row $\boldsymbol{x}_i^T$ and $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)^T$ (the values of $\boldsymbol{\eta}$ are called the linear predictor values) we have $\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta}$. We write $f(\boldsymbol{y}; \boldsymbol{\beta})$ for the likelihood. Now let $\boldsymbol{\beta}$ be fixed and suppose we wish to find for some subspace S of the parameter space the Kullback–Leibler projection onto S. The subspace S might be defined by a subset of "active" covariates as in Dupuis and Robert (2003) but here we consider subspaces such as

$$S = S(\lambda) = \left\{\boldsymbol{\beta} : \sum_{j=1}^{p} |\beta_j| \le \lambda\right\}, \tag{2}$$

$$S = S(\boldsymbol{\beta}^*, \lambda) = \left\{\boldsymbol{\beta} : \sum_{j=1}^{p} |\beta_j|/|\beta_j^*| \le \lambda\right\}, \tag{3}$$

where $\boldsymbol{\beta}^*$ is a parameter value that supplies weighting factors in the constraint or

$$S = S(\lambda, \eta) = \left\{\boldsymbol{\beta} : \sum_{j=1}^{p} |\beta_j| + \eta \sum_{j=1}^{p} \beta_j^2 \le \lambda\right\}. \tag{4}$$

In considering these projections, we assume that the covariates have been transformed to a common scale so that the regression coefficients are also on the same scale. The choice (2) leads to procedures related to the lasso of Tibshirani (1996), (3) relates to the adaptive lasso of Zou (2006) and (4) is related to the elastic net of Zou and Hastie (2005). In (3) we have allowed the space that we are projecting onto to depend on some parameter $\boldsymbol{\beta}^*$, and later we will allow $\boldsymbol{\beta}^*$ to be the parameter in the full model that we are projecting so that the subspace that we are projecting onto is adapting with the parameter. That is, later on we consider generating a sample from the posterior distribution of the parameter in the full model, $\boldsymbol{\beta}^{(1)}, \ldots, \boldsymbol{\beta}^{(s)}$ say, and then we consider projecting $\boldsymbol{\beta}^{(1)}$ onto $S(\boldsymbol{\beta}^{(1)}, \lambda)$, $\boldsymbol{\beta}^{(2)}$ onto $S(\boldsymbol{\beta}^{(2)}, \lambda)$, and so on so that the space we project onto depends on the parameter. In the original adaptive lasso of Zou (2006) typically β_j^* will be chosen

as some preliminary point estimate of β_j, such as the ordinary least squares estimate. In a later section we also consider projections which are related to Breiman's (1995) non-negative garotte and which allow for structured variable selection in the presence of hierarchical relationships among predictors. The close connection between the adaptive lasso and the non-negative garotte is discussed in Zou (2006).

Now let $\boldsymbol{\beta}_S$ be a parameter in the subspace S. In the development below we consider the scale parameter ϕ as known—we consider an unknown scale parameter later. The Kullback–Leibler divergence between $f(\boldsymbol{y};\boldsymbol{\beta})$ and $f(\boldsymbol{y};\boldsymbol{\beta}_S)$ is

$$E_{\beta}\left(\log\frac{f(\boldsymbol{Y};\boldsymbol{\beta})}{f(\boldsymbol{Y};\boldsymbol{\beta}_S)}\right), \tag{5}$$

where E_{β} denotes the expectation with respect to $f(\boldsymbol{y};\boldsymbol{\beta})$. Writing $\mu_i(\boldsymbol{\beta})$ and $\theta_i(\boldsymbol{\beta})$ for the mean and natural parameter for y_i when the parameter is $\boldsymbol{\beta}$ (5) is given by

$$\begin{aligned}
&E_{\beta}\left(\sum_{i=1}^{n}\frac{Y_i\theta_i(\boldsymbol{\beta})-b(\theta_i(\boldsymbol{\beta}))}{\phi/A_i}-\sum_{i=1}^{n}\frac{Y_i\theta_i(\boldsymbol{\beta}_S)-b(\theta_i(\boldsymbol{\beta}_S))}{\phi/A_i}\right)\\
&=\sum_{i=1}^{n}\frac{\mu_i(\boldsymbol{\beta})\theta_i(\boldsymbol{\beta})-b(\theta_i(\boldsymbol{\beta}))}{\phi/A_i}-\sum_{i=1}^{n}\frac{\mu_i(\boldsymbol{\beta})\theta_i(\boldsymbol{\beta}_S)-b(\theta_i(\boldsymbol{\beta}_S))}{\phi/A_i}\\
&=-\log f(\mu(\boldsymbol{\beta});\boldsymbol{\beta}_S)+C,
\end{aligned}$$

where $f(\mu(\boldsymbol{\beta});\boldsymbol{\beta}_S)$ is the likelihood evaluated at $\boldsymbol{\beta}_S$ with data $\boldsymbol{y}$ replaced by fitted means $\boldsymbol{\mu}(\boldsymbol{\beta})$ when the parameter is $\boldsymbol{\beta}$ and C represents terms not depending on $\boldsymbol{\beta}_S$ and hence irrelevant when minimizing over $\boldsymbol{\beta}_S$. So minimization with respect to $\boldsymbol{\beta}_S$ subject to a constraint just corresponds to minimization of the negative log-likelihood subject to a constraint but with data $\boldsymbol{\mu}(\boldsymbol{\beta})$ instead of $\boldsymbol{y}$. Dupuis and Robert (2003) observed that in the case where the subspace S is defined by a set of active covariates calculation of the Kullback–Leibler projection can be done using standard software for calculation of the maximum likelihood estimator in generalized linear models: one simply "fits to the fit" using the fitted values for the full model instead of the responses $\boldsymbol{y}$ in the fitting for a subset model. Clearly for some choices of the response distribution the data $\boldsymbol{y}$ might be integer valued, but commonly generalized linear modelling software does not check this condition so that replacement of the data $\boldsymbol{y}$ with the fitted means for the full model can usually be done.

In our case, suppose we wish to calculate the projection onto the subspace (2). We must minimize

$$-\log f(\boldsymbol{\mu}(\boldsymbol{\beta});\boldsymbol{\beta}_S)\quad\text{subject to}\quad\sum_{j=1}^{p}|\beta_{S,j}|\le\lambda,$$

where $\beta_{S,j}$ denotes the jth element of $\boldsymbol{\beta}_S$. This is equivalent to minimization of

$$-\log f(\boldsymbol{\mu}(\boldsymbol{\beta});\boldsymbol{\beta}_S)+\zeta\sum_{j=1}^{p}|\beta_{S,j}|,$$

for some $\zeta>0$. Here the calculation just involves the use of the lasso of Tibshirani (1996) where in the calculation the responses are replaced by the fitted values $\boldsymbol{\mu}(\boldsymbol{\beta})$. In the case of a Gaussian linear model, the whole solution path over values of δ can be calculated with computational effort equivalent to a single least squares fit (Osborne et al., 2000; Efron et al., 2004). Efficient algorithms are also available for generalized linear models (Park and Hastie, 2007). Note that because the relationship between λ and δ depends on $\boldsymbol{\beta}$, generally we calculate the whole solution path over δ in order to calculate the projection onto the subspace defined by the constraint. One can consider other constraints apart from an L_1 constraint. For instance, (3) leads to minimization of

$$-\log f(\boldsymbol{\mu}(\boldsymbol{\beta});\boldsymbol{\beta}_S)+\zeta\sum_{j=1}^{p}|\beta_{S,j}|/|\beta_j|,$$

for $\zeta>0$ if we choose $\boldsymbol{\beta}^*=\boldsymbol{\beta}$ which gives a certain adaptive lasso estimator (Zou, 2006) obtained by fitting with data $\boldsymbol{y}$ replaced by $\boldsymbol{\mu}(\boldsymbol{\beta})$. Also, (4) leads to minimization of

$$-\log f(\boldsymbol{\mu}(\boldsymbol{\beta});\boldsymbol{\beta}_S)+\zeta\sum_{j=1}^{p}|\beta_{S,j}|+\gamma\sum_{j=1}^{p}\beta_{S,j}^2,$$

for positive constants ζ and γ which is related to the elastic net of Zou and Hastie (2005).

Later we focus on the lasso and adaptive lasso type projections for which the parameter λ needs to be chosen. One way to do this is to follow a similar strategy to the one employed in Dupuis and Robert (2003). Writing $M_S=M_S(\lambda)$ for the model subject to the restriction (2) or (3), we choose λ as small as possible subject to $d(M_S,M_F)<c$. Note that choosing the single parameter λ (through specification of c) is much easier than searching over subsets as in Dupuis and Robert (2003), and that $d(M_S,M_F)$ increases monotonically as λ decreases. An alternative to choosing λ based on relative explanatory power would be to directly choose the observed sparsity in the model: that is, to choose λ so that the posterior mean of the number of active components in the projection is equal to some specified value. The relative loss of explanatory power can also be

reported for this choice. Another possibility is to avoid choosing λ at all, but instead to simply report the characteristics of the models appearing on the solution path over different samples from the posterior distribution in the full model.

So far we have considered the scale parameter ϕ to be known. For binomial and Poisson responses $\phi = 1$, but we also wish to consider Gaussian linear models with unknown variance $\phi = \sigma^2$. Calculation of projections is still straightforward in the Gaussian linear model with unknown variance parameter. Now suppose we have mean and variance parameter $\boldsymbol{\beta}$ and σ^2, and consider some subspace S. It is easily shown that the projection $\boldsymbol{\beta}'_S$ of $\boldsymbol{\beta}$ onto S is independent of the value of σ^2, and that once the projection $\boldsymbol{\beta}'_S$ is calculated, the projection $\sigma_S^{2'}$ is

$$\sigma_S^{2'} = \sigma^2 + \frac{\left(\boldsymbol{\mu}(\boldsymbol{\beta}) - \boldsymbol{\mu}(\boldsymbol{\beta}'_S)\right)^T \left(\boldsymbol{\mu}(\boldsymbol{\beta}) - \boldsymbol{\mu}(\boldsymbol{\beta}'_S)\right)}{n}.$$

3. Computation and predictive inference

We have already discussed computation of the Kullback–Leibler projection onto subspaces of certain forms in generalized linear models. Hence generating from the posterior distribution of the projection is easily done—we simply generate a sample from the posterior distribution $p(\boldsymbol{\beta}|\boldsymbol{y})$ of the parameter, $\boldsymbol{\beta}^{(1)}, \ldots, \boldsymbol{\beta}^{(s)}$ say, and then for each of these parameter values we calculate the corresponding projections $\boldsymbol{\beta}^{(1)'}, \ldots, \boldsymbol{\beta}^{(s)'}$. Note that the pattern of sparsity of the projection is different for different samples from the posterior distribution, so that the posterior distribution of the projection provides one way of exploring model uncertainty.

In algorithmic form, our method works as follows.

1. Generate a sample $\boldsymbol{\beta}^{(1)}, \ldots, \boldsymbol{\beta}^{(s)}$ from the posterior distribution of the parameter in the full model. This can be done by MCMC for example.
2. Choose the value c for the acceptable explanatory loss.
3. For each $i = 1, \ldots, s$, calculate the projections $\boldsymbol{\beta}^{(i)'}(\lambda)$ of $\boldsymbol{\beta}^{(i)}$ for all $\lambda > 0$. Standard lasso algorithms can be used for calculating the whole solution path.
4. Choose λ as small as possible subject to $d(M_s(\lambda), M_F) < c$. Write this value of λ as $\lambda^* = \lambda^*(c)$. Note that the choice of λ^* follows from the choice of c (i.e. λ^* and c do not need to be separately specified).
5. For $i = 1, \ldots, s$, record the models selected by the projections $\boldsymbol{\beta}^{(i)'}(\lambda^*)$ (i.e. which coefficients are nonzero in the projections with $\lambda = \lambda^*$).

Note that for a large enough value of λ the constraint is not active and $\boldsymbol{\beta}^{(i)'}(\lambda) = \boldsymbol{\beta}^{(i)}$ so that calculating the projection for all $\lambda > 0$ in step 2 above will in general not present any difficulty. In the case of a linear model with an L_1 constraint the solution paths are piecewise linear (Osborne et al., 2000; Efron et al., 2004) providing a further simplification. Note also that generating a sample from $p(\boldsymbol{\beta}|\boldsymbol{y})$ and looking at the nonzero coefficients for the projection onto $M_s(\lambda^*)$ defines a distribution on the model space. We refer to this distribution as the posterior distribution on the model space induced by the projection. The ensemble of models generated at step 5 represents a sample from this distribution. In the algorithm above we consider choosing λ through specifying an acceptable upper limit c on $d(M_s(\lambda), M_F)$ but the other possibilities mentioned in Section 2, such as achieving a desired average model size, can also easily be implemented for choosing λ.

For prediction, we approximate the predictive density (1) by

$$\frac{1}{s}\sum_{i=1}^{s} p(\Delta|\boldsymbol{\beta}^{(i)'}, \boldsymbol{y})$$

where $p(\Delta|\boldsymbol{\beta}^{(i)'}, y)$ denotes the predictive distribution for Δ given the parameter value $\boldsymbol{\beta}^{(i)'}$ and data $\boldsymbol{y}$. We can write $\gamma'_j = I(\beta'_j \neq 0)$ for the indicator of whether or not the jth component of the projection of $\boldsymbol{\beta}$ is nonzero, and $\boldsymbol{\gamma}' = (\gamma'_1, \ldots, \gamma'_p)^T$. Then we can write $p_S(\Delta|\boldsymbol{y})$ in a different form to (1), namely

$$p_S(\Delta|\boldsymbol{y}) = \sum_{\boldsymbol{\gamma}'} p(\boldsymbol{\gamma}'|\boldsymbol{y}) p(\Delta|\boldsymbol{\gamma}', \boldsymbol{y}),$$

where

$$p(\Delta|\boldsymbol{\gamma}', \boldsymbol{y}) = \int p(\Delta|\boldsymbol{\gamma}', \boldsymbol{\beta}', \boldsymbol{y}) p(\boldsymbol{\beta}'|\boldsymbol{\gamma}', \boldsymbol{y}) \mathrm{d}\boldsymbol{\beta}'.$$

These expressions for predictive densities are formally similar to those arising in Bayesian model averaging, where different values for the indicators $\boldsymbol{\gamma}'$ define different models. Of course, the posterior distribution on $\boldsymbol{\gamma}'$ cannot be interpreted in quite the same way as the posterior distribution on the model space in a formal Bayesian approach to model comparison, but we still believe that examining $p(\boldsymbol{\gamma}'|\boldsymbol{y})$, the posterior distribution on the model space induced by the projection, can be helpful for exploring different interpretations of the data in our approach. We illustrate this in the examples below.

4. Consistent model selection

Let $\boldsymbol{\beta}^0$ denote the true parameter, and for any $\boldsymbol{\beta}$ write $\mathcal{A}(\boldsymbol{\beta}) = \{k : \beta_k \neq 0\}$ so that for instance $\mathcal{A}(\boldsymbol{\beta}^0)$ is the set of nonzero coefficients for the true parameter. Suppose that $\boldsymbol{\beta}$ is some fixed parameter value and consider $\boldsymbol{\beta}'_S$ which minimizes

$$-\log p(\boldsymbol{\mu}(\boldsymbol{\beta});\boldsymbol{\beta}_S) \quad \text{subject to } \sum_{j=1}^{p}|\beta_{S,j}|/|\beta_j| \le \lambda, \tag{6}$$

with respect to $\boldsymbol{\beta}_S$. That is, we consider in this section Kullback–Leibler projections for subspaces of the form (3) related to the adaptive lasso of Zou (2006). A similar result to the one below can be proved for some projections related to the non-negative garotte in view of the close connection between the adaptive lasso and the non-negative garotte (Zou, 2006). We will examine projections related to the non-negative garotte when we look at structured variable selection problems later.

Considering only the case of a generalized linear model, the minimization problem above is equivalent to minimization of

$$\sum_{i=1}^{n}\left\{\mu_i(\boldsymbol{\beta})\theta_i(\boldsymbol{\beta}_S)+b(\theta_i(\boldsymbol{\beta}_S))\right\}+\gamma\sum_{j=1}^{p}|\beta_{S,j}|/|\beta_j|, \tag{7}$$

where there is a natural one to one correspondence between γ and λ in (6) and for simplicity we are considering the case where the observation specific weights ϕ/A_i are all equal. Actually, as mentioned earlier, the relationship between γ and δ depends on $\boldsymbol{\beta}$, but this can be ignored in what follows: if we take $\lambda = \#\mathcal{A}(\boldsymbol{\beta}^0) + O_p(1/\sqrt{n})$, where $\#\mathcal{A}(\boldsymbol{\beta}^0)$ is the number of the entries in $\mathcal{A}(\boldsymbol{\beta}^0)$, the consistency result for (6) can be similarly established.

If $\boldsymbol{\beta}$ is a sample from the posterior distribution then under general conditions it is a root-n consistent estimator, and we know that for any $\epsilon \in [0,1]$ there exists C not depending on n such that with $\mathcal{N}_n = \{\boldsymbol{\alpha} : \|\boldsymbol{\alpha}-\boldsymbol{\beta}^0\| \le C/\sqrt{n}\}$, $\Pr(\boldsymbol{\beta}\in\mathcal{N}_n) \ge 1-\epsilon$ where the probability is calculated with respect to the distribution

$$q(\boldsymbol{\beta}) = \int p(\boldsymbol{\beta}|\boldsymbol{y})p(\boldsymbol{y}|\boldsymbol{\beta}^0)\mathrm{d}y.$$

We will show that

$$\lim_{n\to\infty}\Pr(\mathcal{A}(\boldsymbol{\beta}'_S)=\mathcal{A}(\boldsymbol{\beta}^0) \quad \text{for every } \boldsymbol{\beta}\in\mathcal{N}_n)=1,$$

for a suitable sequence of values γ_n for γ and hence

$$\lim_{n\to\infty}\Pr(\mathcal{A}(\boldsymbol{\beta}'_S)=\mathcal{A}(\boldsymbol{\beta}^0))=1.$$

That is, the posterior distribution of the projection indicates the correct model with probability one as $n\to\infty$ for a suitable choice of the sequence of parameters γ_n defining the projection.

We assume the following regularity conditions, which are the same as in Zou (2006).

1. The Fisher information $I(\boldsymbol{\beta}^0)$ is positive definite;
2. There is a large enough open set $\mathcal{O}$ containing the true parameter $\boldsymbol{\beta}^0$ such that $\forall\boldsymbol{\beta}\in\mathcal{O}$,

$$|b'''(\boldsymbol{x}^T\boldsymbol{\beta})| \le M(x) < \infty,$$

and

$$E[M(\boldsymbol{x})|x_ix_jx_k|] < \infty,$$

for any i, j, k.

Theorem 1. *For $\boldsymbol{\beta}\in\mathcal{N}_n$, if $\gamma_n/\sqrt{n}\to 0$ and $\gamma_n\to\infty$ as $n\to\infty$, $\boldsymbol{\beta}'_S$ which minimizes (7) is consistent in variable selection and is $\sqrt{n}$-consistent.*

The proof, which is an easy adaptation of a similar result in Zou (2006), is given in an online technical report (Nott and Leng, 2009).

5. Structured variable selection

In the last section we considered variable selection in generalized linear models: as before, write $\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta}$ where $\boldsymbol{\eta}$ is the vector of linear predictor values, $\boldsymbol{X}$ is a design matrix and $\boldsymbol{\beta}$ is a parameter vector. In this section we will combine the structured variable selection approach of Yuan et al. (2007) which uses the non-negative garotte of Breiman (1995) with our projection approach to variable selection.

Consider the model in which $\boldsymbol{\eta} = X(\boldsymbol{\beta}^*\circ\boldsymbol{\theta})$ where $\boldsymbol{\beta}^* = (\beta_1^*,\ldots,\beta_p^*)^T$ and $\boldsymbol{\theta} = (\theta_1,\ldots,\theta_p)^T$ are p-vectors of parameters with $\theta_j\ge 0$, $\sum_{j=1}^p\theta_j \le p$ and where $\circ$ denotes element by element multiplication of two vectors. We write $\boldsymbol{\beta}^*$ instead of $\boldsymbol{\beta}$ to emphasize that $\boldsymbol{\beta}^*$ is a different parameter in a different model to the original one, although our original model can be recovered by setting $\boldsymbol{\beta}^* = \boldsymbol{\beta}$ and $\boldsymbol{\theta}$ a p-vector of ones. We can consider the projection of this parameter onto the subspace

$$S = S(\boldsymbol{\beta},\lambda) = \left\{(\boldsymbol{\beta}^*,\theta) : \boldsymbol{\beta}^* = \boldsymbol{\beta}, \sum_{j=1}^{p}\theta_j\le\lambda, \theta_j\ge 0, j=1,\ldots,p\right\}.$$

To calculate the projection we need to minimize $-\log p(\boldsymbol{\mu}(\boldsymbol{\beta});\boldsymbol{\beta}\circ\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ subject to $\theta_j\ge 0$ and $\sum_{j=1}^p\theta_j\le\lambda$. For the Gaussian case, this is just Breiman's non-negative garotte applied to the fitted values $\boldsymbol{\mu}(\boldsymbol{\beta})$ rather than the data $\boldsymbol{y}$.

Table 1
Predictors for low birth weights dataset.

Predictor	Description
age	Age of mother in years
lwt	Weight of mother (lbs) at least menstrual period
raceblack	Indicator for race = black (0/1)
raceother	Indicator for race other than white or black (0/1)
smoke	Smoking status during pregnancy (0/1)
ptd	Previous premature labors (0/1)
ht	History of hypertension (0/1)
ui	Has uterine irritability (0/1)
ftv1	Indicator for one physician visit in first trimester (0/1)
ftv2+	Indicator for two or more physician visits in first trimester (0/1)

The minimization problem is easily solved. See Yuan et al. (2007) for computational details in the slightly more complicated situation of structured fitting of generalized linear models. As pointed out by Zou (2006), the non-negative garotte is very closely related to the adaptive lasso.

In solving the minimization problem above we may find that some of the θ_j are zero. This allows variable selection in the original model for which we have replaced the parameter $\boldsymbol{\beta}$ with $\boldsymbol{\beta}^* \circ \boldsymbol{\theta}$. Yuan et al. (2007) also suggested a way in which hierarchical structure can be incorporated in the non-negative garotte, and we make use of this idea here. Following their notation, for the ith predictor (corresponding to the ith column of $\boldsymbol{X}$) we write $\mathcal{D}_i$ for the set of predictors which are so-called parents of i. Under the strong heredity principle (Chipman, 1996) all the predictors in $\mathcal{D}_i$ must be included in the model before the ith predictor is included. Under the weak heredity principle at least one of the predictors in $\mathcal{D}_i$ must be included before the ith predictor is included. Yuan et al. (2007) suggest the constraints $\theta_i \leq \theta_j$ for $j \in \mathcal{D}_i$ to enforce the strong heredity principle and $\theta_i \leq \sum_{j\in\mathcal{D}_i} \theta_j$ to enforce the weak heredity principle. The first constraint $\theta_i \leq \theta_j$ ensures that none of the $\theta_j, j \in \mathcal{D}_i$ can be zero unless θ_i is also zero, so that inclusion of predictor i implies inclusion of all predictors in $\mathcal{D}_i$. For the second constraint, we can only have $\theta_j = 0$ for every $j \in \mathcal{D}_i$ if $\theta_i = 0$, so it is only possible to exclude all predictors in $\mathcal{D}_i$ if predictor i is also excluded. Hence the constraints enforce strong and weak heredity respectively, and furthermore the linear nature of the constraints ensures that computations are still tractable.

6. "Large p, small n" problems and preconditioning

Paul et al. (2008) suggest that in "large p, small n" regression problems with more predictors than observations it is beneficial to separate the problem of obtaining good predictions from that of variable selection. With this in mind, they suggest a two step procedure where first a good predictor $\hat{\boldsymbol{y}}$ is found for the mean response, and then in a second stage a model selection and fitting procedure such as the lasso is applied with the responses $\boldsymbol{y}$ replaced by $\hat{\boldsymbol{y}}$. They show that such a procedure can perform better than application of the fitting and selection procedure to the raw outcome $\boldsymbol{y}$.

Suppose, for example, that we have fitted a linear model with $p > n$ and obtained an estimate $\hat{\boldsymbol{\beta}}$ of the coefficients (not sparse). When using the lasso as the variable selection method, Paul et al. (2008) would consider minimization of

$$(X\boldsymbol{\beta} - X\hat{\boldsymbol{\beta}})^T(X\boldsymbol{\beta} - X\hat{\boldsymbol{\beta}}) + \zeta \sum_{j=1}^{p} |\beta_j|,$$

with respect to $\boldsymbol{\beta}$ to obtain a sparse estimate of the coefficient that is "close" to $\hat{\boldsymbol{\beta}}$. Here $\zeta > 0$ is some tuning parameter and X is the design matrix. As we have seen in Section 2, the minimization problem above is exactly the minimization problem we solve to find the Kullback–Leibler projection of the coefficients onto a subspace such as (2). The discussion here extends to other variable selection approaches such as the elastic net. We can see that the method of Paul et al. (2008) corresponds to projecting a point estimate of $\boldsymbol{\beta}$, whereas our approach consists of projecting samples from the posterior rather than just a point estimate.

7. Examples and simulations

7.1. Low birthweight data

We consider application of our approach to the low birthweight data of Hosmer and Lemeshow (1989). The data are concerned with 189 births at a US hospital. We consider a logistic regression model for a response which is a binary indicator for birthweight being less than 2.5 kg. The predictors in the model are shown in Table 1. These predictors had been shown to be associated with low birthweight in past studies and it was desired to find out which of the predictors were important for the medical centre where the data were collected. In our analysis we leave the binary predictors unchanged, but centre and scale the other predictors to have mean zero and variance one. We fit the full model with a prior on the coefficients that is normal, with mean vector 0 and covariance matrix $3\boldsymbol{I}$ where $\boldsymbol{I}$ denotes the identity matrix. This is a fairly noninformative prior on the scale of the probabilities: note that making the prior variances of coefficients very large would correspond to a very informative prior on the probability scale where high prior probability is placed on coefficient values corresponding to

Table 2
Posterior means and standard deviations of coefficients for full model fitted to low birthweight data.

Predictor	Posterior mean	Posterior standard deviation
age	−0.21	0.21
lwt	−0.48	0.22
raceblack	1.06	0.52
raceother	0.65	0.43
smoke	0.68	0.41
ptd	1.31	0.47
ht	1.69	0.67
ui	0.64	0.46
ftv1	−0.49	0.46
ftv2	0.11	0.44

Table 3
Two most frequently appearing models of each size in solution path for the projection together with relative frequency of each model within all appearances of model of the same size (Prob/Size). Zeros and ones in the columns labelled by the predictors show inclusion and exclusion for different models (rows).

Model size	Predictor										Prob/size
	age	lwt	black	other	smoke	ptd	ht	ui	ftv1	ftv2	
1	0	0	0	0	0	1	0	0	0	0	0.48
1	0	1	0	0	0	0	0	0	0	0	0.17
2	0	1	0	0	0	1	0	0	0	0	0.24
2	0	0	0	0	0	1	1	0	0	0	0.10
3	0	1	0	0	0	1	1	0	0	0	0.13
3	0	1	1	0	0	1	0	0	0	0	0.06
4	0	1	0	0	1	1	1	0	0	0	0.07
4	0	1	0	0	0	1	1	0	1	0	0.06
5	0	1	1	0	1	1	1	0	0	0	0.05
5	0	1	1	0	0	1	1	1	0	0	0.05
6	0	1	1	1	1	1	1	0	0	0	0.06
6	1	1	1	0	1	1	1	0	0	0	0.05
7	0	1	1	1	1	1	1	1	0	0	0.10
7	1	1	1	1	1	1	1	0	0	0	0.09
8	1	1	1	1	1	1	1	1	0	0	0.13
8	0	1	1	1	1	1	1	1	1	0	0.13
9	1	1	1	1	1	1	1	1	1	0	0.29
9	1	1	1	1	1	1	1	1	0	1	0.19

most of the fitted values being close to zero or one. Coefficient estimates (posterior means) and posterior standard deviations obtained by fitting the full model are shown in Table 2. These results were obtained using the MCMCpack package in R (Martin and Quinn, 2007). To obtain the results reported we ran the MCMC scheme for 1000 "burn in" and 10 000 sampling iterations.

We consider our projection approach to selection with the lasso type constraint (2) as well as the adaptive lasso type constraint (3). For each sample from the posterior, the whole solution path was calculated for the projection as the parameter λ was varied, and all the distinct models on the path were recorded. Thus for each sample from the posterior, we should have roughly 11 distinct models (including the null and the full models) because there are 10 covariates. We say "roughly" 11 distinct models because it is possible for a variable to leave the model as the regularization parameter is increased in the solution path, but this is not very common in practice. Table 3 shows the two most frequently appearing models of each size across solution paths for all samples from the posterior, together with the relative frequency with which this model appears amongst models with the same number of covariates. The table only reports results for the adaptive lasso. The reason why we only report results for the adaptive lasso is shown in Fig. 1, which gives the relative loss of explanatory power as a function of the posterior expected number of variables selected in the projection. We show in the figure results for both the $N(0, 3\boldsymbol{I})$ prior on coefficients, as well as for a flat prior $p(\boldsymbol{\beta}) \propto 1$. The results are very similar, as are the best models of each size (results not shown). The solid lines in the graphs are for the lasso projection and the dashed lines for the adaptive lasso—it can be seen that for the adaptive lasso there is a reduced loss of explanatory power compared to the lasso for a given level of parsimony.

Examining the models in Table 3, the indicators for number of first trimester physician visits (ftv), one of the indicators for race and age appear to be the least important covariates. This is consistent with other published analyses of this dataset such as in Venables and Ripley (2002). They consider stepwise variable selection using AIC in a main effects model including all the covariates, which results in exclusion of the dummy variables coding for ftv and age. They also consider inclusion of second order interactions and note that there is some evidence for an interaction between ftv and age. Raftery and Zheng (2003) also consider some reference Bayesian model averaging approaches to the analysis of this dataset. Their conclusions concerning the important variables (based on marginal posterior probabilities of inclusion) are similar to ours, although it should be noted that their model is different with first trimester physician visits treated as a continuous covariate rather

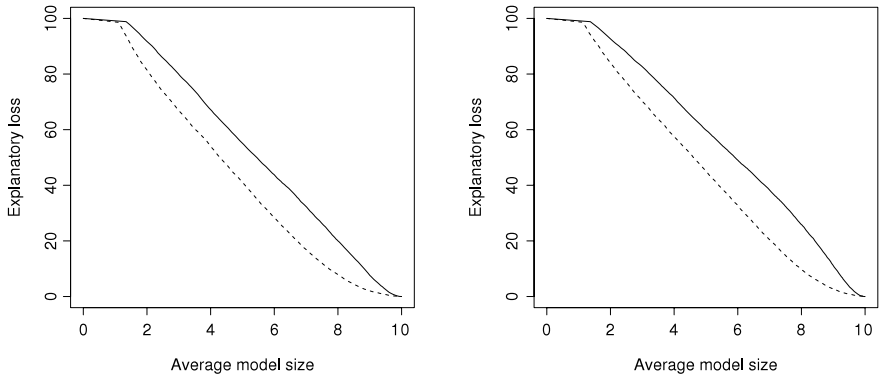

Fig. 1. Plots of relative loss of explanatory power versus posterior expected model size for low birth weight example. The solid line is for the lasso projection and the dashed line is for the adaptive lasso. The left figure is for the $N(0, 3I)$ prior and the right figure is for the uniform prior.

Table 4
Best two models of each size and explanatory loss $d(M_S, M_F)$ for method of Dupuis and Robert and the low birthweight data. Zeros and ones in the columns labelled by the predictors show inclusion and exclusion for different models (rows).

Model size	Predictor										$d(M_S, M_F)$
	age	lwt	black	other	smoke	ptd	ht	ui	ftv1	ftv2	
1	0	0	0	0	0	1	0	0	0	0	0.72
1	0	1	0	0	0	0	0	0	0	0	0.87
2	1	0	0	0	0	1	0	0	0	0	0.60
2	0	1	0	0	0	1	0	0	0	0	0.60
3	0	1	0	0	0	1	1	0	0	0	0.43
3	1	0	0	0	0	1	1	0	0	0	0.51
4	0	1	0	0	0	1	1	0	1	0	0.33
4	0	1	1	0	0	1	1	0	0	0	0.35
5	0	1	1	0	0	1	1	0	1	0	0.25
5	0	1	0	0	0	1	1	1	1	0	0.26
6	0	1	1	0	0	1	1	1	1	0	0.19
6	1	1	1	0	0	1	1	0	1	0	0.20
7	0	1	1	1	1	1	1	1	0	0	0.13
7	1	1	1	0	0	1	1	1	1	0	0.13
8	0	1	1	1	1	1	1	1	1	0	0.07
8	1	1	1	1	1	1	1	1	0	0	0.08
9	1	1	1	1	1	1	1	1	1	0	0.02
9	0	1	1	1	1	1	1	1	1	1	0.05

than being coded through two indicator variables as in our analysis, which follows Venables and Ripley (2002). We also applied the original method of Dupuis and Robert (2003) in this example. Table 4 shows the best two models of each size for their approach. The conclusions concerning the important variables are similar. We note that the search of the model space required here for Dupuis and Robert's method is computationally intensive—Dupuis and Robert (2003) note that forward and backward stepwise searches through the model space to find a model with satisfactory explanatory power often lead to the same model, but this is not the case in this example. The orderings of predictors obtained by forward additions from the null model and backward deletions from the full model are quite different, necessitating a more exhaustive search of the model space for most choices of the satisfactory level of explanatory loss. However, even in this situation there is still some potential for computational savings by a clever search strategy, as explained in Dupuis and Robert (2003).

7.2. Structured variable selection example

Our next example concerns a variable selection problem with hierarchical structure. The data are simulated following a similar example discussed in Yuan et al. (2007). The purpose of considering this example is to show that the method described in Section 5 which incorporates hierarchical constraints into variable selection is beneficial. In particular, we consider a model satisfying the strong heredity principle, and then show that a projection approach to variable selection which imposes strong heredity outperforms an approach which does not impose this constraint. By outperforms here we mean that for a given level of parsimony (a given value for the posterior expected number of nonzero components of the projection) we have a greater posterior probability for the model chosen via the projection to encompass the true model, with a relatively small loss of explanatory power due to imposing the constraint. When we say "encompass the true model" we mean that all variables in the true model are chosen, but possibly some additional irrelevant variables are also selected.

In the example of Yuan et al. (2007) three predictors X_1, X_2 and X_3 are simulated following a multivariate normal distribution with mean zero and $\mathrm{Cov}(X_i, X_j) = \rho^{|i-j|}$ for values ρ of $-0.5, 0$ and 0.5. There are $n = 50$ observations simulated

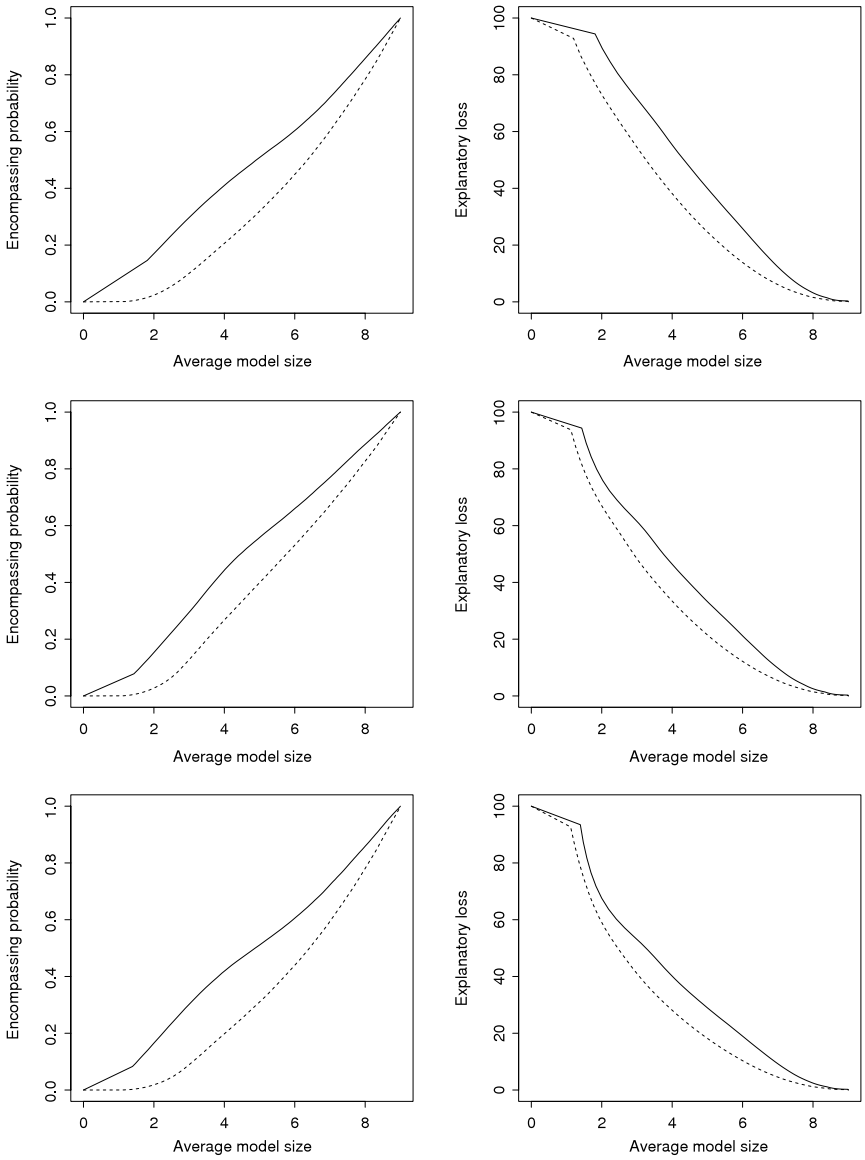

Fig. 2. Plots of posterior probability of encompassing the true model versus average model size (left column) and explanatory loss versus average model size (right column). The parameter ρ takes values of -0.5 (top) 0 (middle) and 0.5 (bottom). Solid line is for strong heredity and broken line no constraint.

and 100 different datasets are considered for each value of ρ. We consider fitting a model that includes X_1, X_2, X_3 and all second order interaction terms (nine possible terms in all—no intercept is fitted). The true model used to generate Y is

$$Y = 3X_1 + 2X_2 + 1.5X_1X_2 + \epsilon,$$

where $\epsilon \sim N(0, 9)$. Note that this model respects the strong heredity principle—for the interaction term, the corresponding main effects are also included. We use two variants of our non-negative garotte approach to fitting the data. The first variant respects the strong heredity principle, and the second variant does not impose any constraint.

We considered a grid of 100 equally spaced values for λ between 0 and 9 in the constraint $\sum_{j=1}^{p} \theta_j \leq \lambda$ as described in Section 5. Using the usual noninformative prior in the Bayesian linear model on the regression coefficients and variance parameter of $p(\boldsymbol{\beta}, \sigma^2) \propto \sigma^{-2}$, we can simulate directly from the posterior distribution without the need for iterative methods (see, for instance Gelman et al., 2003). For each simulated dataset we generated 1000 samples from the posterior distribution. For each value of λ in the grid and each draw from the posterior distribution, we calculated projections (with and without the strong heredity constraint) recording the number of active variables, whether or not the projection encompassed the true model and the Kullback–Leibler divergence between the full model and the projections.

Plotting the posterior expected values of these quantities against one another for the grid of values of λ gives a sense of the trade off between parsimony, predictive accuracy and identification of the important variables. Fig. 2 shows plots of

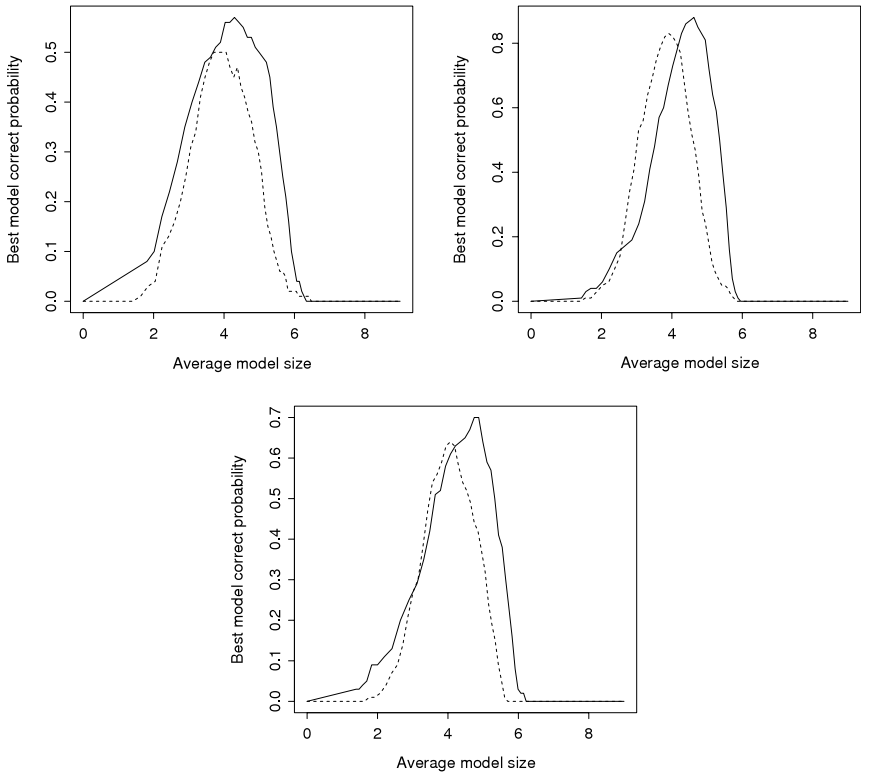

Fig. 3. Plots of frequency of selection of the correct model versus average model size for model selected by projection of posterior mean. Solid line is for strong heredity and broken line no constraint. The parameter ρ takes values of -0.5 (top left), 0 (top right) and 0.5 (bottom).

the probability of encompassing the true model and of the explanatory loss versus posterior expected number of variables selected in the projected model for the two projection methods (imposing strong heredity, solid line, and no constraint, broken line). For a given level of parsimony it can be seen that the projection method which imposes the hierarchical constraint has a higher posterior probability of encompassing the true model so that enforcing the strong heredity principle is helpful for obtaining more parsimonious models and for identifying the important variables. There is also a loss of explanatory power by imposing the constraint, but this loss is relatively small in situations such as this one where the constraint is appropriate. We also considered the situation where a single model is chosen by finding the Kullback–Leibler projection of a point estimate of the parameter (the posterior mean). Fig. 3 shows the performance of this model selector. The frequency of selections of the correct model across the simulation replicates is plotted against average model size. The solid line shows the estimate enforcing strong heredity and the dotted line the unconstrained method. It can be seen that certainly for the "best" choices of λ (in terms of maximizing the chance of choosing the correct model) for each case and method we obtain better results with the approach enforcing strong heredity.

7.3. *"Large p, small n" regression*

We now consider some simulations for the "large p, small n" case where there are more predictors than observations. We consider generating 100 datasets with $n = 20$ and 40 predictors. The datasets follow a linear model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\boldsymbol{\epsilon} \sim N(0, 5^2\boldsymbol{I})$. Below we write $\boldsymbol{x}_{i\cdot}$ for the ith row of the design matrix $\boldsymbol{X}$.

1. Example 1: set $\beta_j = 0, j = 1, \ldots, 10, j = 21, \ldots, 30$, $\beta_j = 2, j = 11, \ldots, 20, j = 31, \ldots, 40$. We have $\boldsymbol{x}_{i\cdot} \sim N(0, \boldsymbol{I})$.
2. Example 2: set $\beta_j = 4, j = 1, \ldots, 5$, $\beta_j = 0, j = 6, \ldots, 40$. We generate $\boldsymbol{x}_{i\cdot} \sim N(0, \boldsymbol{\Sigma})$ with $\Sigma_{jj} = 1, j = 1, \ldots, 40$ and $\Sigma_{ij} = 0.5i \neq j$.

In the first example there is no multicollinearity, but 20 active predictors. In the second example there is moderate multicollinearity but only 5 active predictors.

Since the number of predictors is double the number of observations in both examples, here we are considering a "large p, small n" situation. For a Bayesian analysis of the data with an encompassing model we consider the Bayesian lasso of Park and Casella (2008). They consider the following priors on parameters. If an intercept term β_0 is included, this is given a flat prior $p(\beta_0) \propto 1$ and this parameter can be integrated out of the model analytically. Conditional on the variance σ^2, the β_j

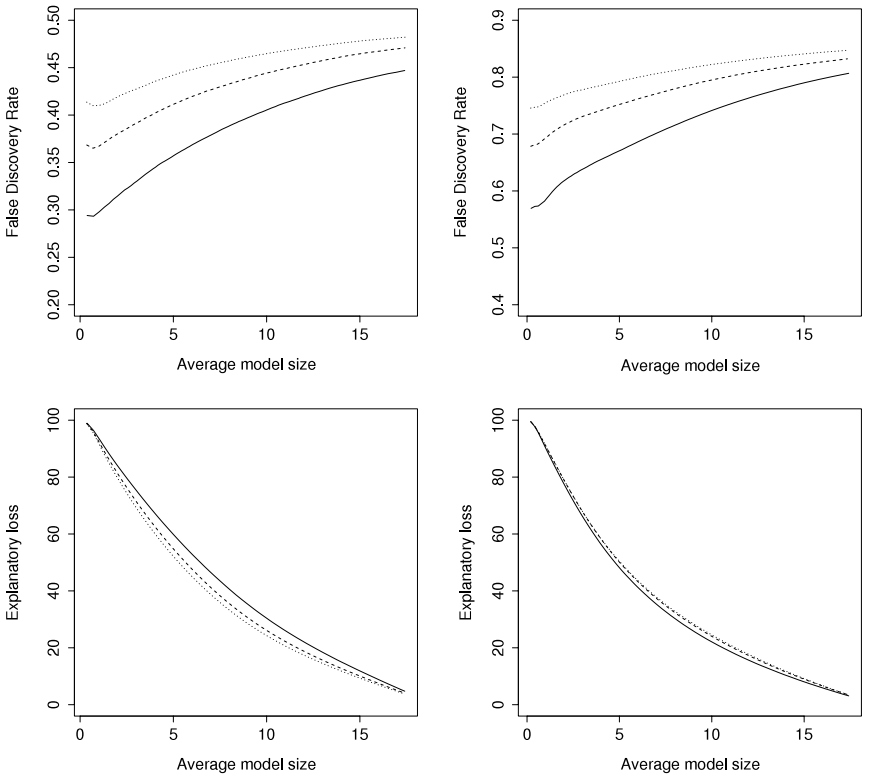

Fig. 4. Plots of false discovery rate versus average model size (top row) and explanatory loss versus average model size (bottom row) for examples 1 (left) and 2 (right) and $\rho = 5, 10, 15$ (solid line, dashed line, dotted line respectively). Plotted points correspond to a grid of values for the constraint λ.

are conditionally independent in their prior with

$$p(\beta_j|\sigma^2) = \frac{\rho}{2\sigma^2}\exp\left(-\frac{\rho|\beta_j|}{\sqrt{\sigma^2}}\right),$$

where $\rho > 0$ is a shrinkage parameter. Finally an inverse gamma prior can be used for σ^2, where we use $IG(0.01, 0.01)$. A hyperprior can be placed on ρ, or it can be estimated by marginal maximum likelihood as outlined in Park and Casella (2008) or by cross-validation. To investigate prior sensitivity we consider a number of different fixed values for ρ here of 5, 10 and 15. Park and Casella (2008) outline an efficient MCMC scheme for computations.

We consider projections based on the adaptive lasso, and Fig. 4 shows plots of the false discovery rate (where by this we mean the average number of variables incorrectly selected divided by average number of variables selected) versus average model size. The averages are over 100 simulation replicates. Quite a large model would need to be chosen to encompass all the active predictors. Note that if we use the classical lasso to do selection then the number of predictors chosen by the projection cannot be more than the number of observations. In our approach, even if we project onto the subspace (2) related to the lasso which results in projections involving no more than n active covariates, we note that different variables are selected for different parameters in the full model and averaging over different projections for prediction results in the use of more than n covariates in prediction. Fig. 4 also shows the explanatory loss as a function of the average model size. Model uncertainty is considerable here, and we believe that the distribution on the model space defined by the projection is extremely valuable for exploring model uncertainty. Fig. 5 shows for the first simulation replicate in each example with $\rho = 10$ the marginal posterior probabilities of the variables being nonzero in the projection. The projections in the figure correspond to average model size of 13 (example 1) and 10 (example 2) corresponding to approximately 20% explanatory loss in both cases. The lines show the mean values for these probabilities within the active and inactive groups. It is clear that there is some useful information in the posterior distribution of the projection for distinguishing active from inactive variables. It is important to realize that Fig. 5 is examining posterior probabilities of selection in the projection for a single replicate, not the frequentist behaviour of selection across replicates—such frequentist behaviour is summarized by the false discovery rates of Fig. 4.

Fig. 6 shows the behaviour of the model selector which selects a single model based on projection of the posterior mean. The average number of variables correctly selected is plotted versus average model size. We have also considered simulations identical to those already described but with $n = 40$. We do not report all the results here, but not surprisingly model selection performance is improved for this case. Fig. 7 shows the model selection performance of the model chosen by projection of the posterior mean, similar to Fig. 6.

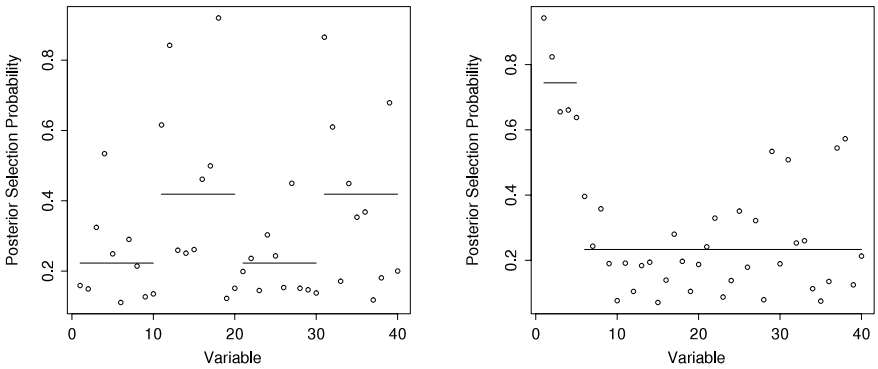

Fig. 5. Plots of marginal posterior probability of inclusion versus variable for first simulation replicate for example 1 (left) and 2 (right). Probabilities are for projections with average model size 13 (left) and 10 (right) corresponding in both cases to approximately 20% explanatory loss. The lines show the mean posterior probabilities of inclusion among the active and inactive groups.

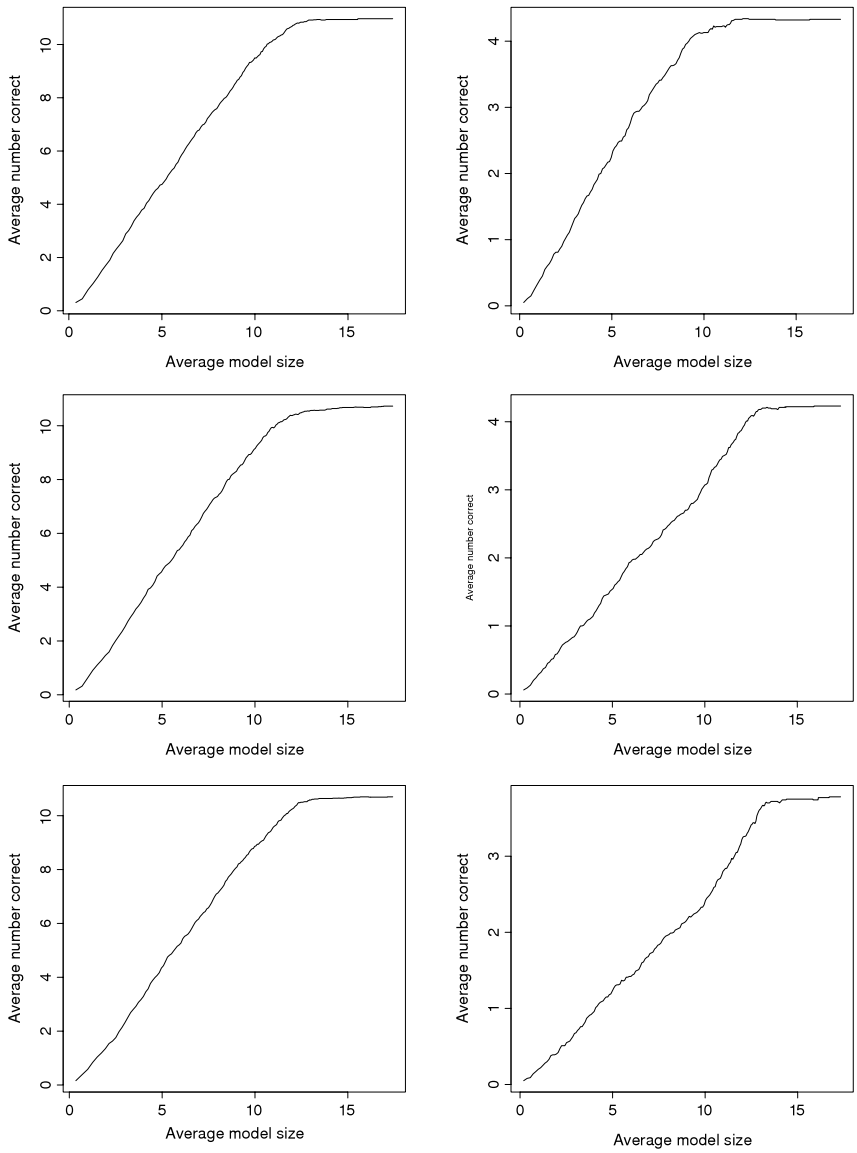

Fig. 6. Plots of number of variables correctly selected versus average model size for model selected by projection of posterior mean. Left column is for example 1, right column for example 2, and rows correspond to $\rho = 5$ (top), 10 (middle) and 15 (bottom).

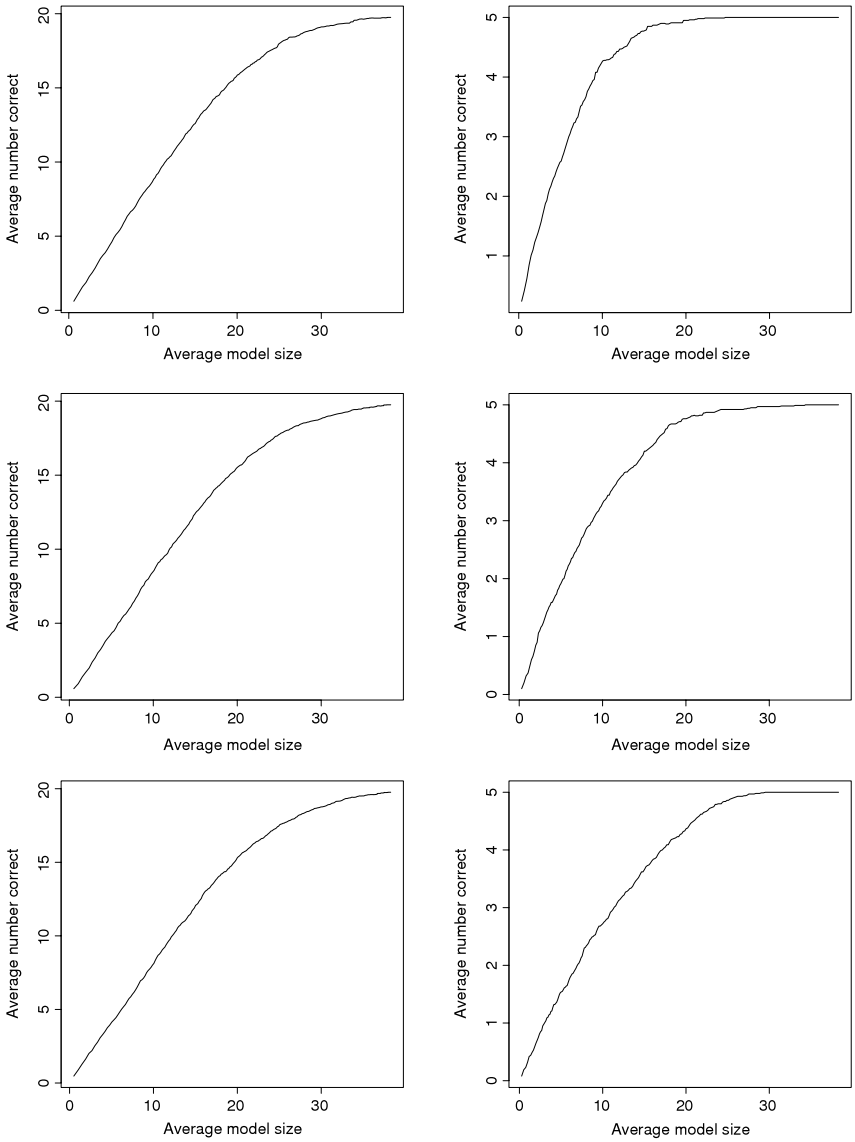

Fig. 7. Plots of number of variables correctly selected versus average model size for model selected by projection of posterior mean in simulations with $n = 40$. Left column is for example 1, right column for example 2, and rows correspond to $\rho = 5$ (top), 10 (middle) and 15 (bottom).

8. Conclusion

We have discussed the use of Kullback–Leibler projections related to the lasso as a tool for the exploration of model uncertainty. There are many possible extensions to our suggested framework. One interesting possibility which we are currently pursuing is the use of projections related to versions of the lasso for selection on batches of parameters and random effects.

Acknowledgements

David Nott was supported by a Singapore MOE Grant R-155-000-068-133. Leng is supported by National University of Singapore research grants.

References

Berger, J., Pericchi, L., 1996. The intrinsic Bayes factor for model selection and prediction. J. Amer. Statist. Assoc. 91, 109–122.
Bernardo, J.M., Rueda, R., 2002. Bayesian hypothesis testing: A reference approach. Internat. Statist. Rev. 70, 351–372.
Breiman, L., 1995. Better subset regression using the non-negative garotte. Technometrics 3, 373–384.
Brown, P.J., Fearn, T., Vannucci, M., 1999. The choice of variables in multivariate regression: A non-conjugate Bayesian decision theory approach. Biometrika 86, 635–648.

Brown, P.J., Vannucci, M., Fearn, T., 2002. Bayes Model averaging with selection of regressors. J. Roy. Statist. Soc. Ser. B 64, 519–536.
Chipman, H., 1996. Bayesian variable selection with related predictors. Canad. J. Statist. 24, 17–36.
Draper, D., Fouskakis, D., 2000. A case study of stochastic optimization in health policy: Problem formulation and preliminary results. J. Global Optim. 18, 399–416.
Dupuis, J.A., Robert, C.P., 2003. Variable selection in qualitative models via an entropic explanatory power. J. Statist. Plann. Inference 111, 77–94.
Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., 2004. Least angle regression (with discussion). Ann. Statist. 32, 407–499.
Fernández, C., Ley, E., Steel, M.F.J., 2001. Benchmark priors for Bayesian model averaging. J. Econometrics 100, 381–427.
Gelfand, A.E., Ghosh, S.K., 1998. Model choice: A minimum posterior predictive loss approach. Biometrika 85, 1–11.
Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B., 2003. Bayesian Data Analysis, 2nd ed. CRC Press, London.
Gelman, A., Meng, X.-L., Stern, H., 1996. Posterior predictive assessment of model fitness via realized discrepancies. Statist. Sinica 6, 733–807.
George, E.I., Foster, D.P., 2000. Calibration and empirical Bayes variable selection. Biometrika 87, 731–747.
Goutis, C., Robert, C.P., 1998. Model choice in generalised linear models: A Bayesian approach via Kullback–Leibler projections. Biometrika 85, 29–37.
Griffin, J.E., Brown, P.J., 2007. Bayesian adaptive lassos with non-convex penalization. Technical report. Available at: http://www.kent.ac.uk/ims/personal/jeg28/BALasso.pdf.
Hoeting, J.A., Madigan, D., Raftery, A.E., Volinsky, C.T., 1999. Bayesian model averaging: A tutorial (with Discussion). Statist. Sci. 14, 382–401. (Correction: vol. 15, pp. 193–195. Corrected version. Available at: http://www.stat.washington.edu/www/research/online/hoeting1999.pdf).
Hosmer, D.W., Lemeshow, S., 1989. Applied Logistic Regression. Wiley, New York.
Kass, R.E., Raftery, A.E., 1995. Bayes factors. Journal of the American Statistical Association 90, 773–795.
Laud, P.W., Ibrahim, J.G., 1995. Predictive model selection. Journal of the Royal Statistical Society, Series B 57, 247–262.
Lindley, D.V., 1968. The choice of variables in multiple regression (with discussion). J. R. Stat. Soc. Ser. B 30, 31–66.
Martin, A., Quinn, M., 2007. The MCMCpack package (version 0.9-1). R package manual. Available at http://cran.r-project.org/doc/packages/MCMCpack.pdf.
Mengersen, K., Robert, C.P., 1996. Testing for mixtures: A Bayesian entropy approach. In: Berger, J.O., Bernardo, J.M., Dawid, A.P., Lindley, D.V., Smith, A.F.M. (Eds.), Bayesian Statistics, vol. 5. Oxford University Press, pp. 255–276.
Nott, D.J., Leng, C., 2009. Bayesian projection approaches to variable selection and exploring model uncertainty. Technical report. Available at: http://arxiv.org/PS_cache/arxiv/pdf/0901/0901.4605v1.pdf.
O'Hagan, A., 1995. Fractional Bayes factors for model comparison (with discussion). J. R. Stat. Soc. Ser. B 56, 99–138.
Osborne, M.R., Presnell, B., Turlach, B.A., 2000. A new approach to variable selection in least squares problems. IMA J. Numer. Anal. 20, 389–403.
Park, T., Casella, G., 2008. The Bayesian lasso. J. Amer. Statist. Assoc. 103, 681–686.
Park, M.-Y., Hastie, T., 2007. An L1 regularization-path algorithm for generalized linear models. J. Roy. Statist. Soc. Ser. B 69, 659–677.
Paul, D., Bair, E., Hastie, T., Tibshirani, R., 2008. Pre-conditioning for feature selection and regression in high-dimensional problems. Ann. Statist. 36, 1595–1618.
Raftery, A.E., 1996. Approximate Bayes factors and accounting for model uncertainty in generalized linear models. Biometrika 83, 251–266.
Raftery, A.E., Zheng, Y., 2003. Discussion: Performance of Bayesian model averaging. J. Amer. Statist. Assoc. 98, 931–938.
Spiegelhalter, D.J., Best, N.G., Carlin, B.P., van der Linde, A., 2002. Bayesian measures of model complexity and fit (with discussion). J. R. Stat. Soc. Ser. B 64, 583–639.
Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B 58, 267–288.
Vehtari, A., Lampinen, J., 2004. Model Selection via predictive explanatory power. Report B38. Laboratory of Computational Engineering, Helsinki University of Technology.
Venables, W.N., Ripley, B.D., 2002. Modern Applied Statistics with *S*, fourth ed. Springer, New York.
Yuan, M., Joseph, V.R., Zou, H., 2007. Structured variable selection and estimation. Technical report. Available at: http://www2.isye.gatech.edu/~myuan/YuanPub.html.
Yuan, M., Lin, Y., 2005. Efficient empirical Bayes variable selection and estimation. J. Amer. Statist. Assoc. 100, 1215–1225.
Zou, H., 2006. The adaptive lasso and its oracle properties. J. Amer. Statist. Assoc. 101, 1418–1429.
Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. J. R. Stat. Soc. Ser. B 67, 301–320.