Cross-validation prior choice in Bayesian probit regression with many covariates

D. Lamnisos*, J. E. Griffin† and M. F. J. Steel*

May 11, 2009

Abstract

This paper examines prior choice in probit regression through a predictive cross-validation criterion. In particular, we focus on situations where the number of potential covariates is far larger than the number of observations, such as in gene expression data. Cross-validation avoids the tendency of such models to fit perfectly. We choose the hyperparameter in the ridge prior, c, as the minimizer of the log predictive score. This evaluation requires substantial computational effort, and we investigate computationally cheaper ways of determining c through importance sampling. Various strategies are explored and we find that $K-$fold importance densities perform best, in combination with either mixing over different values of c or with integrating over c through an auxiliary distribution.

Keywords: Bayesian variable selection, cross-validation, gene expression data, importance sampling, predictive score, ridge prior.

1 Introduction

We are interested in modelling binary variables $\boldsymbol{y} = (y_1, \ldots y_n)'$, which take the values 0 or 1. For example, we may want to find genes that discriminate between two disease states using samples taken from patients in the first disease

*Department of Statistics, University of Warwick, Coventry, CV4 7AL, U.K. and † Institute of Mathematics, Statistics and Actuarial Science, University of Kent, Canterbury, CT2 7NF, U.K. Correspondence to M. Steel, Email: M.F.Steel@stats.warwick.ac.uk, Tel.: +44(0)24-76523369, Fax: +44(0)24-76524532

state ($y_i = 1$) or the second one ($y_i = 0$). Typically, the number of measured gene expressions (covariates) will be much larger than the number of samples. The popular probit model assumes that y_i is modelled as $y_i \sim \text{Bernoulli}(\Phi(\eta_i))$, where $\boldsymbol{\eta} = (\eta_1, \eta_2, \ldots, \eta_n)' = \alpha\mathbf{1} + \mathbf{X}\boldsymbol{\beta}$. Here $\mathbf{X}$ is an $n \times p$ matrix whose (i, j)-th entry is the measurement of the j-th covariate for the i-th individual, Φ is the cumulative distribution function of a standard normal random variable, $\boldsymbol{\eta}$ is a vector of linear predictors, $\mathbf{1}$ represents a $n \times 1$-dimensional vector of ones, α is the intercept and $\boldsymbol{\beta}$ represents a $p \times 1$-dimensional vector of regression coefficients. We specifically consider situations where $p >> n$ and assume that the covariates have been centred.

We wish to model the response $\boldsymbol{y}$ in terms of a (small) subset of the p explanatory variables. Models are identified with the choice of a particular subset of covariates. The 2^p possible subset choices are indexed by the vector $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)$ where $\gamma_j = 0$ or 1 according to whether the j-th predictor is included or excluded from the model. The number of variables included in a model is denoted by $p_\gamma = \sum_{j=1}^{p} \gamma_j$. Exclusion of a variable means that the corresponding element of $\boldsymbol{\beta}$ is zero. Thus, a model indexed by $\boldsymbol{\gamma}$ containing p_γ variables is defined by

$$\begin{aligned} y_i|\alpha, \boldsymbol{\beta}_\gamma, \mathbf{x}_{\gamma i} &\sim \text{Bernoulli}(\Phi(\eta_i)) \\ \boldsymbol{\eta} &= \alpha\mathbf{1} + \mathbf{X}_\gamma\boldsymbol{\beta}_\gamma \end{aligned}$$

where $\mathbf{X}_\gamma$ is a $n \times p_\gamma$ matrix whose columns are the included variables and $\boldsymbol{\beta}_\gamma$ is a $p_\gamma \times 1$-dimensional vector of regression coefficients. We denote the model parameters by $\boldsymbol{\theta}_\gamma = (\alpha, \boldsymbol{\beta}_\gamma')' \in \boldsymbol{\Theta}_\gamma$. To deal with the uncertainty regarding the inclusion of covariates, we put a prior on $\boldsymbol{\gamma}$ and adopt a formal Bayesian framework for inference, which naturally leads to methods for model selection or for Bayesian Model Averaging (BMA) (see *e.g.* Hoeting *et al.*, 1999 and Fernández *et al.*, 2001). To complete the Bayesian model, we need a prior distribution for the intercept α, the regression coefficients $\boldsymbol{\beta}_\gamma$ and the model $\boldsymbol{\gamma}$ which usually has the following structure $\pi(\alpha, \boldsymbol{\beta}_\gamma, \boldsymbol{\gamma}) = \pi(\boldsymbol{\beta}_\gamma|\boldsymbol{\gamma})\pi(\alpha)\pi(\boldsymbol{\gamma})$. For the intercept α we adopt a $N(0, h)$ prior as in Sha et al. (2004) and Brown and Vanucci (1998). Since the covariates have been centred, α represents the overall mean of the linear predictors and is regarded as a common parameter to all models. Thus, a non-informative improper prior can also be used for α, as *e.g.* in Fernández et al. (2001). The prior distribution for the regression coefficients $\boldsymbol{\beta}_\gamma$ is the so-called ridge prior

$$\pi(\boldsymbol{\beta}_\gamma|\boldsymbol{\gamma}) \quad \sim \quad N_{p_\gamma}(\mathbf{0}, c\mathbf{I}_{p_\gamma}), \tag{1}$$

where $N_q(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ represents a q-dimensional normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, and $\mathbf{I}_q$ is the $q \times q$ identity matrix. This commonly used prior (see Denison *et al*, 2002) implies prior independence between the coefficients. Alternatively, a g-prior where the prior covariance matrix in (1) is given by $c(\mathbf{X}_\gamma^T \mathbf{X}_\gamma)^{-1}$ could be used, as in Bottolo and Richardson (2007). Finally, we assume that each regressor is included independently with probability w, which implies that

$$\pi(\boldsymbol{\gamma}) = w^{p_\gamma}(1-w)^{p-p_\gamma} \tag{2}$$

and p_γ is binomially distributed $\text{Bin}(p, w)$.

The choice of the hyperparameters w and c is critical for posterior inference on model space since w plays the main role in inducing a size penalty and c regularises the regression coefficients. The value of the hyperparameter c in (1) has an important effect on BMA in probit regression with $p >> n$. More specifically, the value of c influences both the variables that appear in the best models and their posterior inclusion probabilities. Bottolo and Richardson (2007), Ley and Steel (2009) and Liang et al. (2008) discuss analogous results for the g-prior in linear regression. Ley and Steel (2009) also highlight the role of the hyperparameter c as a model size penalty in that context.

As discussed in Sha et al. (2004), c determines the amount of shrinkage of the probit regression coefficients when $p >> n$. Therefore, probit models with large regression coefficients (in absolute value) are favoured when c is large. In addition, Bayesian model selection chooses models that perfectly fit the data when there is less regularisation (large c). However, in practice, a perfect model fit is often associated with poor out-of-sample prediction.

In this work, we focus on the choice of the hyperparameter c. We use the log predictive score introduced by Good (1952) as a measure of predictive performance. Prediction is not necessarily the main goal in itself, but is a good indicator of successful variable selection. The log predictive score is defined through the cross-validation density $\pi(y_i|\boldsymbol{y}_{-i}, c)$, where $\boldsymbol{y}_{-i} = (y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n)$ is the response vector without the ith observation. Fully Bayesian analysis would place a prior distribution on c. Inference for c allows considering multiple values of c which might lead to improved predictive performance. However, a prior on c is hard to choose. Furthermore, Cui and George (2008) found that empirical Bayes approaches that provide an adaptive choice for the $g-$prior hyperparameter in Bayesian linear regression outperform fully Bayesian analysis that places a prior on c. Our results

with these data point in the same direction: using a diffuse proper prior on c we can end up with considerably worse prediction than using an "optimal" choice of c, even though the prior chosen has ample mass close to the optimal value. Ongoing research will report further on this issue. The value of c that minimizes the log predictive score is the preferred choice for c. Alternative proper score functions for categorical variables, discussed by Gneiting and Raftery (2007) are also examined, and lead to quite similar minimizers. Since cross-validation densities are employed to determine c, we should be able to successfully partition not just the sample, but also the population into the different groups. This approach can be viewed as empirical Bayes as the response vector $\boldsymbol{y}$ is used to determine the preferred value of c.

The main aim of this work is to estimate accurately and efficiently the log predictive score and thus to identify its minimizer. The cross-validation density $\pi(y_i|\boldsymbol{y}_{-i}, c)$, the main component of all predictive scores, does not have a closed analytic expression in our context and therefore we suggest two novel importance samplers to estimate it. In comparison to the direct Markov chain Monte Carlo (MCMC) methodology for each observation and value of c, importance sampling makes repeated use of the same sample, generated from an importance density, to estimate $\pi(y_i|\boldsymbol{y}_{-i}, c)$ for different i and c, leading to computational gains.

The accuracy and efficiency in estimating the log predictive score of various importance samplers are evaluated and compared using some gene expression datasets obtained from DNA microarray studies. We propose guidelines for the implementation of these samplers that optimize the efficiency and make them more or less automatic procedures. The proposed methods lead to accurate estimates of the optimal value for c with a very considerable saving in computational effort. Matlab code to implement our samplers is freely available at
`http://www.warwick.ac.uk/go/msteel/steel_homepage/software/`

2 Influence of the hyperparameter c in BMA

It is well known that the amount of regularisation can have an important impact on many statistical procedures. Here we illustrate that it is a particularly critical issue in our context, since the value of the hyperparameter c crucially affects BMA in probit regression with $p >> n$. In order to show this we first consider a study of links between gene expression and cases of rheumatoid arthritis or osteoarthritis. This

study was analysed by Sha et al. (2003) and has $n = 31$ and $p = 755$. We identify the genes that appear in the ten models with the highest posterior probability for different values of c, ranging from 1 to 100. Throughout the paper, we use the data augmentation algorithm of Holmes and Held (2006) to sample from $\pi(\boldsymbol{\theta}_\gamma, \boldsymbol{\gamma}|\boldsymbol{y}, c)$ through MCMC. We generated five independent chains with $2,000,000$ iterations, the burn-in period $100,000$ and the thinning of 10 resulted in an MCMC sample size T of $190,000$. The posterior probability of model $\boldsymbol{\gamma}$ is computed as the relative frequency of model $\boldsymbol{\gamma}$ in the MCMC output.

Table 1 reports the genes of the Arthritis dataset that appeared in the ten best (*i.e.* highest posterior probability) models for each c. Genes indexed 170, 258 and 290 appeared for all c and genes 489, 584 and 729 appeared for five out of six values of c. However, many genes are only identified for specific values of c. This indicates substantial differences in variable selection for different values of c.

c	Genes included in the ten best models
1	20 83 145 **170** 225 **258** **290** 324 332 395 473 498 665 707 728 740 742
5	43 44 83 145 **170** **258** **290** 324 473 489 498 539 584 729 740
10	43 44 83 **170** **258** **290** 324 421 461 489 539 584 646 729
30	44 49 **170** **258** **290** 324 389 392 395 421 461 489 584 646 665 729
50	43 44 **170** 208 **258** **290** 389 421 461 489 532 539 584 646 729 754
100	89 **170** 208 **258** **290** 389 395 421 489 532 584 585 616 671 729 754

Table 1: Genes of the Arthritis dataset included in the ten best models of the union of the five chains for different values of c. Boxed genes are selected for all c

Figure 1 shows the estimated posterior gene inclusion probabilities and the corresponding scatter-plots of the log estimated posterior gene inclusion probabilities of the Arthritis dataset for $c = 1, 100$. There are some substantial differences in posterior inclusion probabilities for both datasets. Gene 290 in the Arthritis dataset has posterior inclusion probability 0.45 when $c = 1$ and 0.2 for $c = 100$. On the other hand, gene 258 has posterior inclusion probability 0.15 when $c = 1$ and 0.4 for $c = 100$. It is obvious from the scatter-plots that many log posterior gene inclusion probabilities are quite different for the pairs $c = 1, 100$. Again, these results indicate differences in variable selection for different values of c.

As well as affecting the posterior inclusion probabilities, the hyperparameter c regularises the amount of shrinkage of the included regression coefficients and this is

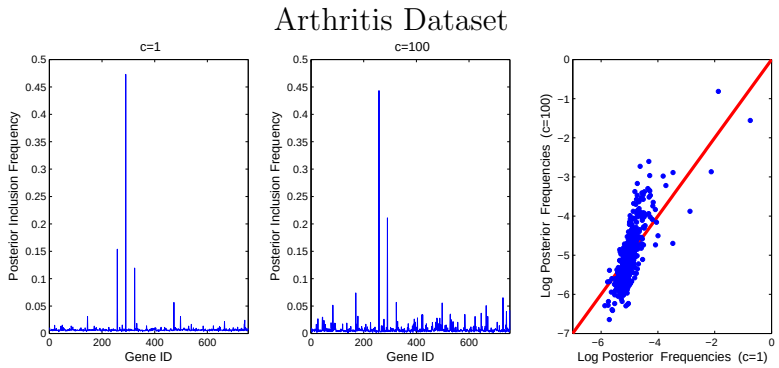

Figure 1: Estimated posterior gene inclusion probabilities and scatter-plot of the log estimated posterior gene inclusion probabilities of the Arthritis dataset for different values of c

illustrated in the left panel of Figure 2, which shows the posterior density function of the ratio of the sum of regression coefficients (in absolute value) to the model size for different values of c, *i.e.*

$$||\boldsymbol{\beta}||_1 = \frac{1}{p_{\gamma^{(t)}}} \sum_{j=1}^{p_{\gamma^{(t)}}} |\beta_{\gamma_j{}^{(t)}}| \qquad t = 1, \ldots, T,$$

where $\beta_{\gamma_j{}^{(t)}}$ are the components of the regression coefficient vector $\boldsymbol{\beta}_{\gamma^{(t)}}$. There is more probability mass at larger values of $||\boldsymbol{\beta}||_1$ when c is large. Therefore, the probit models with large regression coefficients (in absolute value) are favoured for large c. The existence of these models is possible in the large p setting because of the large number of potential models. This results in different best models and consequently in different variable selection as c varies.

Large absolute values of $||\boldsymbol{\beta}||_1$ are often associated with overfitting. To illustrate this problem it is useful to express the probit model as a latent variable model by introducing auxiliary variables $z_1, \ldots, z_n$ such that

$$\begin{aligned} y_i &= \begin{cases} 1 & \text{if } z_i > 0 \\ 0 & \text{otherwise.} \end{cases} \\ \boldsymbol{z} &= \alpha + \mathbf{X}_\gamma \boldsymbol{\beta}_\gamma + \boldsymbol{\varepsilon} \text{ with } \boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \mathbf{I}_n), \end{aligned} \tag{3}$$

where y_i is now deterministic conditional on the sign of the stochastic auxiliary variable z_i. The right panel of Figure 2 displays the posterior mean, $\hat{z}_i$, of the auxiliary variable z_i for each individual of the Arthritis dataset. The first seven

individuals have response $y_i = 1$ and the other twenty-four have $y_i = 0$. Clearly, the absolute value of $\hat{z}_i$ is larger for all i when there is less regularisation (large c). These fitted values are in the tails of the standard normal distribution for $c \geq 30$, indicating that the fitted probabilities $\Phi(\hat{z}_i)$ are very close to 1 when $y_i = 1$ and very close to 0 otherwise. In other words, the visited models discriminate perfectly the n observations into the disease groups. Therefore, when $p >> n$, BMA selects probit models that fit the data perfectly when there is less regularisation (large c) on the regression coefficients. However, perfect model fit typically leads to high out-of-sample prediction error. Thus we need to carefully consider the specification of c, avoiding a perfect fit to the data.

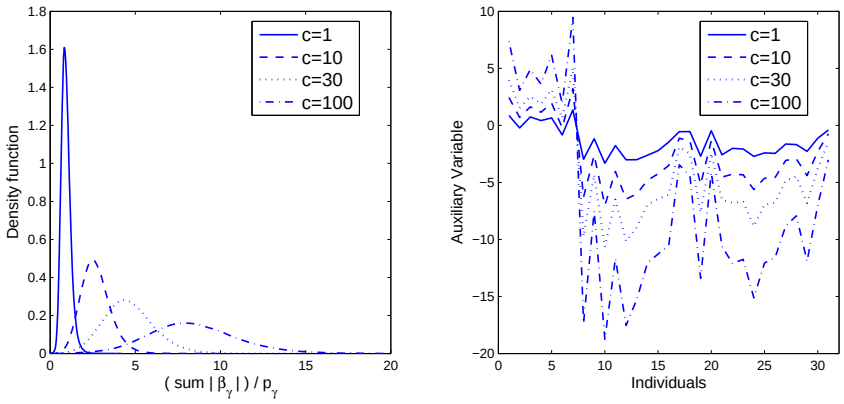

Figure 2: Arthritis data: The left panel displays the posterior density function of the ratio of the sum of regression coefficients (in absolute value) to the model size for different values of the hyperparameter c. The right panel shows the mean of the fitted auxiliary variable for each individual i for different values of c

A direct consequence of perfect fit is the uncertainty about the intercept α since we can make moderate changes to α while leaving all $\hat{z}_i$ in the tails of the standard normal distribution, retaining the same fit of the probit model. However, this does not happen for small values of c, where a small change in α would appreciably affect the fit. Figure 3 displays the posterior density function of the intercept for different values of the hyperparameter c. The absolute value of the posterior mode and the variance of α clearly increases with c.

These feature of the inference are common to many gene expression data sets. For example, we performed the same inference for the Colon Tumour dataset described by Alon et al. (1999), which contains $n = 62$ observations of tumour and normal colon groups with $p = 1224$ and found similar results.

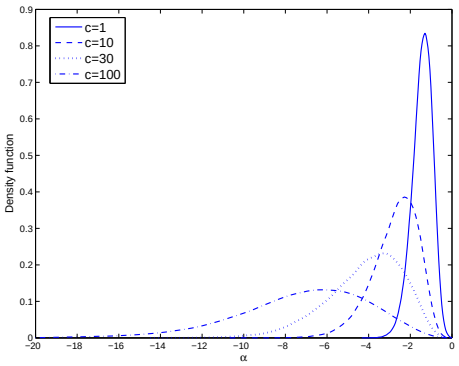

Figure 3: Arthritis data: Posterior density function of the intercept for different values of the hyperparameter c. The prior distribution on α is $N(0, 100)$

3 Estimation of c using predictive criteria

The parameter c is part of the Bayesian model and different values of c indicate alternative prior beliefs and consequently alternative models. Gelfand and Dey (1994) and Gelfand et al. (1992) argue that predictive distributions should be used for model comparison because these are directly comparable and, typically, prediction is a primary purpose for the chosen model. In the typical areas of application we consider in this paper, the key concern is often variable selection (*e.g.* identification of important determinants of a disease status), but good predictive performance tends to be linked to successful variable selection. Gelfand et al. (1992) also argue for a cross-validation viewpoint which, in our case, is implemented through the log predictive score, defined by

$$S(c) = -\frac{1}{n}\sum_{i=1}^{n} \ln \pi(y_i|\boldsymbol{y}_{-i}, c)$$

where $\pi(y_i|\boldsymbol{y}_{-i}, c)$ is the cross-validation density as defined in the Introduction. In a pairwise model comparison this results in the log pseudo-Bayes factor (Geisser and Eddy, 1979). This leave-one-out cross-validation can be extended to $K-$fold cross-validation by partitioning the sample into K subsets and using

$$S(c) = -\frac{1}{n}\sum_{i=1}^{n} \ln \pi(y_i|\boldsymbol{y}_{-\kappa(i)}, c) \tag{4}$$

where $\kappa : \{1, \ldots, n\} \rightarrow \{1, \ldots, K\}$ is an indexing function that indicates the kth $(k = 1, \ldots, K)$ partition to which observation i is allocated and $-\kappa(i)$ represents

the set $\{1, \ldots, n\}$ with the kth element of the partition removed. The random-fold cross-validation of Gneiting and Raftery (2007) can also be considered. The value of c that minimizes $S(c)$ will be our preferred choice for c. Since cross-validation is employed to determine c, the resulting variable selection should be able to successfully partition not just the sample but also the population into the appropriate groups. The underlying idea is that good out-of-sample predictive performance is indicative of good variable selection.

Fernández et al. (2001) also used a log predictive score similar to (4) to evaluate different choices for the g-prior hyperparameter in Bayesian linear regression. Bernardo and Smith (1994) also make use of cross-validation densities to approximate expected utilities in decision problems where a set of alternative models are to be compared.

Alternative proper score functions for binary variables could replace the logarithmic score function in (4). Gneiting and Raftery (2007) list a number of proper score functions. These are the quadratic or Brier predictive score, the spherical predictive score and the continuous ranked probability score which, in the case of a binary variable, is proportional to the quadratic predictive score.

The cross-validation density $\pi(y_i|\boldsymbol{y}_{-\kappa(i)}, c)$ is the main component of all predictive scores. This density for the ith individual is given by

$$\pi(y_i|\boldsymbol{y}_{-\kappa(i)}, c) = \sum_{\boldsymbol{\gamma}} \int_{\boldsymbol{\Theta}_\gamma} \pi(y_i|\boldsymbol{\theta}_\gamma, \boldsymbol{\gamma})\ \pi(\boldsymbol{\theta}_\gamma, \boldsymbol{\gamma}|\boldsymbol{y}_{-\kappa(i)}, c)\ d\boldsymbol{\theta}_\gamma = \mathbb{E}[\pi(y_i|\boldsymbol{\theta}_\gamma, \boldsymbol{\gamma})]\,, \qquad (5)$$

where the expectation is taken with respect to the joint posterior distribution $\pi(\boldsymbol{\theta}_\gamma, \boldsymbol{\gamma}|\boldsymbol{y}_{-\kappa(i)}, c)$ and does not have a closed analytic expression. However, it can be estimated by Markov chain Monte Carlo methodology. The MCMC estimate of (5) is given by

$$\begin{aligned} \hat{\pi}(y_i|\boldsymbol{y}_{-\kappa(i)}, c) &= \frac{1}{T}\sum_{j=1}^{T} \pi(y_i|\boldsymbol{\theta}_{\gamma^{(j)}}, \boldsymbol{\gamma}^{(j)}) \\ &= \frac{1}{T}\sum_{j=1}^{T} \Phi(\tilde{\mathbf{x}}_{\gamma i}\boldsymbol{\theta}_{\gamma^{(j)}})^{y_i}\ (1 - \Phi(\tilde{\mathbf{x}}_{\gamma i}\boldsymbol{\theta}_{\gamma^{(j)}}))^{(1-y_i)} \qquad (6) \end{aligned}$$

where $(\boldsymbol{\theta}_{\gamma^{(1)}}, \boldsymbol{\gamma}^{(1)}), \ldots, (\boldsymbol{\theta}_{\gamma^{(T)}}, \boldsymbol{\gamma}^{(T)})$ is an MCMC sample with stationary distribution $\pi(\boldsymbol{\theta}_\gamma, \boldsymbol{\gamma}|\boldsymbol{y}_{-\kappa(i)}, c)$. The $1 \times (p_\gamma + 1)$-dimensional vector $\tilde{\mathbf{x}}_{\gamma i}$ has 1 as first element and the others are the covariate measurements of the relevant coefficients for the i-th individual.

The $K-$fold log predictive score is estimated at $l = 12$ values of c equally spaced in the logarithmic scale with lower value 0.1 and upper value 1000. This covers values of c inducing a lot of regularisation as well as values inducing very little and significantly extends the guideline range of Sha et al. (2004) for these data. The MCMC estimate of the log predictive score is given by replacing $\pi(y_i|\boldsymbol{y}_{-\kappa(i)}, c)$ in (4) by $\hat{\pi}(y_i|\boldsymbol{y}_{-\kappa(i)}, c)$.

For each data partition in the sum in (4) and each value of c we generated $500,000$ drawings after a burn-in period of $100,000$ and thinned these to every 5^{th} draw, leading to an MCMC sample size T of $80,000$. The right-hand panels of Figure 4 displays both the MCMC estimates and a smooth estimated curve for $S(c)$ for the Arthritis and Colon Tumour datasets. We used $K = n$, that is $\kappa(i) = i$, for the Arthritis dataset and $K = 9$ for the Colon Tumour dataset (using a randomly chosen partition, with 7 observations in each set but one, which has 6 observations). Results for $K = n$ are very similar for the latter data, but execution time is then multiplied by more than n/K ($62/9 = 6.89$ in our case). Cubic smoothing splines were applied to the MCMC estimates of $S(c)$, leading to a roughly convex estimated curve for both datasets, indicating the existence of a value of c that minimizes the log predictive score. This value of c is around 1 for the Arthritis dataset, and is less clear-cut for the Colon Tumour dataset since any value of c in the interval $(15, 145)$ ($\log(c)$ in the interval $(2.7, 5)$) results in quite similar estimates of $S(c)$. In both cases, Bayesian variable selection for the extremes of c (and thus of regularisation) is associated with poorer predictive performance.

The other panels of Figure 4 display the MCMC estimates and a smooth estimated curve of the discussed alternative predictive scores. The estimated curves of all predictive scores are very similar in shape to the ones with the log predictive score and have the same minimizer. Thus, the optimal c is very robust to the choice of predictive score, and we will focus on the log predictive score in the sequel.

The direct MCMC methodology needs an MCMC sample with stationary distribution $\pi(\boldsymbol{\theta}_\gamma, \boldsymbol{\gamma}|\boldsymbol{y}_{-\kappa(i)}, c)$ for all data partitions and c to estimate $\pi(y_i|\boldsymbol{y}_{-\kappa(i)}, c)$ using (6). Therefore, it needs Kl MCMC chains to estimate the log predictive score at l points. Table 2 reports the CPU time in seconds needed by the MCMC methodology (using code in Matlab 7.4.0 on a dual core PC with a 2.2GHz CPU and 3.24GB of RAM) to estimate the log predictive scores of Figure 4. It is obviously a computationally expensive task to employ the direct MCMC methodology. This motivates us to employ importance sampling methods to estimate $\pi(y_i|\boldsymbol{y}_{-\kappa(i)}, c)$. The practical

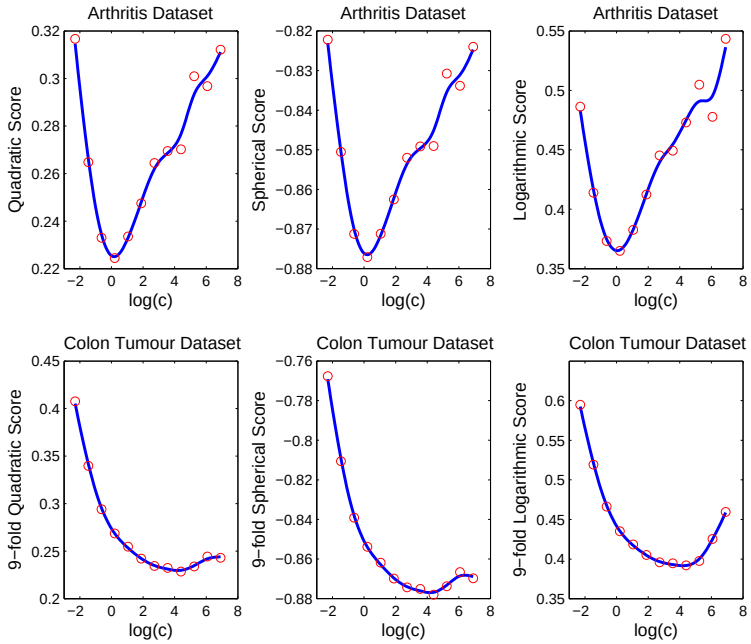

Figure 4: MCMC estimates and smooth estimated curves of different predictive score functions for the Arthritis and Colon Tumour datasets

advantage of these methods is that the same sample (generated from the importance density) can be used repeatedly to estimate $\pi(y_i|\boldsymbol{y}_{-\kappa(i)}, c)$ for different i and c. Ideally, the importance samplers should have similar accuracy in estimating $S(c)$ but need much less CPU time. The right-hand panels of Figure 4 will be used to compare and evaluate the accuracy of the different importance sampling methods introduced in the following section.

Dataset	CPU
Arthritis	290,960
Colon Tumour	122,870

Table 2: The CPU time in seconds needed by the MCMC methodology to estimate the log predictive scores of the Arthritis and Colon Tumour datasets

4 Computational approach

The predictive densities needed to calculate $S(c)$ will be estimated using importance sampling. This method approximates the integral

$$\mathbb{E}_f[h(X)] = \int_x h(x)\, f(x)\, dx \tag{7}$$

by

$$\sum_{j=1}^{T} w^{(j)} h(x_j) \Big/ \sum_{j=1}^{T} w^{(j)}, \tag{8}$$

where a sample $x_1, \ldots, x_T$ is generated from a given distribution g and the importance weight $w^{(j)} \propto f(x_j)/g(x_j)$. The (possibly unnormalized) densities f and g are called the target and importance density respectively. More details on importance sampling can be found in *e.g.* Liu (2001) and Robert and Casella (2004).

The success of the method depends on the accuracy of the approximation which is controlled by the difference between the importance and target densities and can be measured by the effective sample size. If T independent samples are generated from the importance density, then the effective sample size is

$$\text{ESS} = \frac{T}{1 + \text{cv}^2},$$

where cv^2 denotes the coefficient of variation of the importance weights (Liu, 2001). This is interpreted in the sense that the weighted samples are worth ESS independent and identically drawn samples from the target density. In other words, the variance of the importance weights needs to be small to avoid a few drawings dominating the estimate in (8). The ESS will be used as a measure of the efficiency of the importance samplers introduced in the following subsections.

4.1 Importance Samplers Using All Observations

Gelfand et al. (1992) and Gelfand and Dey (1994) suggest using the posterior distribution of the model parameters given all the data as the importance density to estimate cross-validation densities. In our context, we can consider the posterior distribution $\pi(\boldsymbol{\theta}_\gamma, \boldsymbol{\gamma}|\boldsymbol{y}, c_0)$ as an importance density to estimate $\pi(y_i|\boldsymbol{y}_{-\kappa(i)}, c)$, given by (5), for all i and different values of c. A default value of c_0 is required such that the importance density $\pi(\boldsymbol{\theta}_\gamma, \boldsymbol{\gamma}|\boldsymbol{y}, c_0)$ results in accurate estimate of $S(c)$ in the

relevant range of c, $[0.1, 1000]$. As this idea implies large potential computational gains, it is the one we investigate first.

The ESS of $\pi(\boldsymbol{\theta}_\gamma, \boldsymbol{\gamma}|\boldsymbol{y}_{-\kappa(i)}, c)$ with the importance density $\pi(\boldsymbol{\theta}_\gamma, \boldsymbol{\gamma}|\boldsymbol{y}, c_0)$ was calculated for all i and c. The mean ESS at each c is the average over all observations and shows the efficiency of the sampler in estimating the log predictive score at c. For both the Arthritis and Colon Tumour datasets, the mean ESS is high when c is close to c_0 and low for the other values of c. This indicates that the importance density $\pi(\boldsymbol{\theta}_\gamma, \boldsymbol{\gamma}|\boldsymbol{y}, c_0)$ is quite different from $\pi(\boldsymbol{\theta}_\gamma, \boldsymbol{\gamma}|\boldsymbol{y}_{-\kappa(i)}, c)$ when c_0 is not close to c, resulting in estimates of $S(c)$ with high variance. Therefore we only use the importance density $\pi(\boldsymbol{\theta}_\gamma, \boldsymbol{\gamma}|\boldsymbol{y}, c_0)$ when $\pi(\boldsymbol{\theta}_\gamma, \boldsymbol{\gamma}|\boldsymbol{y}_{-\kappa(i)}, c_0)$ is the target density and $\pi(y_i|\boldsymbol{y}_{-\kappa(i)}, c_0)$ is the quantity to be estimated. However, the mean ESS decreases with c_0, indicating that the difference between $\pi(\boldsymbol{\theta}_\gamma, \boldsymbol{\gamma}|\boldsymbol{y}, c_0)$ and $\pi(\boldsymbol{\theta}_\gamma, \boldsymbol{\gamma}|\boldsymbol{y}_{-\kappa(i)}, c_0)$ increases with c_0. Therefore, the observations of the kth data part play a more important role in determining the posterior distribution $\pi(\boldsymbol{\theta}_\gamma, \boldsymbol{\gamma}|\boldsymbol{y}, c_0)$ when c_0 is large. This is a consequence of the models perfect fit to the data for large values of c_0 discussed already in Section 3.

Figure 5 displays the resulting importance estimates of the log predictive score $S(c)$ for the Arthritis and Colon Tumour datasets. In comparison with Figure 4 these log predictive scores are underestimated for large c. The mean of the fitted auxiliary variables displayed in Figure 2 are on the tails of the standard normal distribution for these values of c resulting in an overestimation of $\pi(y_i|\boldsymbol{\theta}_\gamma, \boldsymbol{\gamma})$ and consequently in an overestimation of $\pi(y_i|\boldsymbol{y}_{-i}, c)$. Thus, the perfect fit to the data for large values of c is causing an underestimation of $S(c)$. This effect is also pronounced for the Colon Tumour dataset where the kth element of the partition (the prediction subsample) represents a larger proportion of the data.

4.2 Multiple Importance Samplers

The results of the previous section lead us to concentrate on using the importance densities $\pi(\boldsymbol{\theta}_\gamma, \boldsymbol{\gamma}|\boldsymbol{y}_{-\kappa(i)}, c_0)$ and $\pi(\boldsymbol{\theta}_\gamma, \boldsymbol{\gamma}|\boldsymbol{y}_{-\kappa(i)})$ with the target density $\pi(\boldsymbol{\theta}_\gamma, \boldsymbol{\gamma}|\boldsymbol{y}_{-\kappa(i)}, c)$ for different values of c, where $i = 1, \ldots, n$. We should note that the same importance densities are used for all observations belonging to the same data partition. These importance samplers result in more accurate estimates of the log predictive score. However, they come at greater cost than the methods in the previous subsections, since K MCMC chains, one for each data partition, are needed to estimate the log predictive score $S(c)$. This number of chains is still l times smaller than the

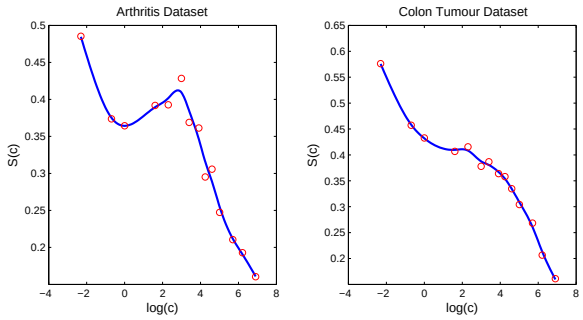

Figure 5: Importance estimates of the log predictive score $S(c)$ for the Arthritis and Colon Tumour datasets

direct MCMC methodology of Section 3, which needs Kl MCMC chains to estimate $S(c)$ at l points. To get a better idea of the accuracy of the procedures, we will replicate each sampler five times throughout this section.

4.2.1 Standard Importance Sampler

First, the $K-$fold standard importance sampler uses the importance density $\pi(\boldsymbol{\theta}_\gamma, \boldsymbol{\gamma}|\boldsymbol{y}_{-\kappa(i)}, c_0)$ to estimate $\pi(y_i|\boldsymbol{y}_{-\kappa(i)}, c)$ for different values of c, where $i = 1, \ldots, n$. We try to find a default value of c_0 that leads to an accurate estimate of $S(c)$ in the relevant range of c, $[0.1, 1000]$. The importance weight is given by

$$w \propto \frac{\pi(\boldsymbol{\theta}_\gamma, \boldsymbol{\gamma}|\boldsymbol{y}_{-\kappa(i)}, c)}{\pi(\boldsymbol{\theta}_\gamma, \boldsymbol{\gamma}|\boldsymbol{y}_{-\kappa(i)}, c_0)} \propto \left(\frac{c_0}{c}\right)^{p_\gamma/2} \exp\left\{-\frac{1}{2}\boldsymbol{\beta}_\gamma{}'\boldsymbol{\beta}_\gamma\left(\frac{1}{c} - \frac{1}{c_0}\right)\right\}. \tag{9}$$

In the case that $c = c_0$, the importance and MCMC estimates of $\pi(y_i|\boldsymbol{y}_{-\kappa(i)}, c)$ are the same and they are given by (6). However, it is a difficult task to suggest a default value of c_0 because other choices of c_0 can lead to rather different results, and without the benchmark of the direct MCMC results, we would not know which value of c_0 to choose. Next, we will introduce importance samplers which do not require choosing a value for c_0.

4.2.2 Mixture Importance Sampler

The difficulty of finding an appropriate value for c_0 leads us to consider methods that estimate $\pi(y_i|\boldsymbol{y}_{-\kappa(i)}, c)$ using importance sampling distributions which are not restricted to a single value c_0. A potentially more efficient method for estimating the score at each value of c uses nonlinear regression methods to combine estimates

at a range of values of c. This generalizes our default MCMC approach by using $\pi(\boldsymbol{\theta}_\gamma, \boldsymbol{\gamma}|\boldsymbol{y}_{-\kappa(i)}, c_m)$ to estimate $\pi(y_i|\boldsymbol{y}_{-\kappa(i)}, c_l)$ when $m \neq l$. We define a positive, increasing sequences of values $c_1, c_2, \ldots, c_M$ and generate an MCMC sample from the importance density $\pi(\boldsymbol{\theta}_\gamma, \boldsymbol{\gamma}|\boldsymbol{y}_{-\kappa(i)}, c_m)$ for each i and c_m. Usually, we would choose $c_1, \ldots, c_M$ to be equally spaced in the logarithmic scale. Since the values of c_m are ordered and increasing, the last value of the MCMC sample from $\pi(\boldsymbol{\theta}_\gamma, \boldsymbol{\gamma}|\boldsymbol{y}_{-\kappa(i)}, c_m)$ could be used as the initial value of the MCMC chain with stationary distribution $\pi(\boldsymbol{\theta}_\gamma, \boldsymbol{\gamma}|\boldsymbol{y}_{-\kappa(i)}, c_{m+1})$. Therefore the MCMC samplers do not need a long burn-in period. The importance estimate $\hat{\pi}_{c_m}(y_i|\boldsymbol{y}_{-\kappa(i)}, c)$ of $\pi(y_i|\boldsymbol{y}_{-\kappa(i)}, c)$ is computed for each one of the M importance densities $\pi(\boldsymbol{\theta}_\gamma, \boldsymbol{\gamma}|\boldsymbol{y}_{-\kappa(i)}, c_m)$ and the M importance estimates $\hat{\pi}_{c_m}(y_i|\boldsymbol{y}_{-\kappa(i)}, c)$ are weighted according to the distance between c and c_m using a Gaussian kernel $K_\lambda(c, c_m) = \phi\left(\frac{|\log(c)-\log(c_m)|}{\lambda}\right)$ with window size $\lambda = 0.5$. The kernel-weighted estimate of $\pi(y_i|\boldsymbol{y}_{-\kappa(i)}, c)$ is

$$\hat{\pi}(y_i|\boldsymbol{y}_{-\kappa(i)}, c) = \sum_{m=1}^{M} K_\lambda(c, c_m)\, \hat{\pi}_{c_m}(y_i|\boldsymbol{y}_{-\kappa(i)}, c) \Big/ \sum_{m=1}^{M} K_\lambda(c, c_m).$$

In the special case that the predetermined values $c_1, \ldots, c_M$ are the 12 equally spaced points stated in Section 3, there are two main differences between the mixture importance sampler and the direct MCMC methodology. Firstly, the mixture importance sampler involves shorter MCMC runs with smaller burn-in. Secondly, the mixture importance sampler makes use of the information contained in different MCMC chains. More specifically, the direct MCMC methodology uses the sample of a single MCMC chain to estimate $\pi(y_i|\boldsymbol{y}_{-\kappa(i)}, c_m)$ at the value c_m whereas the mixture importance sampler combines the sample of M different MCMC chains to estimate $\pi(y_i|\boldsymbol{y}_{-\kappa(i)}, c_m)$. In comparison with the $K-$fold standard importance sampler, the mixture importance sampler involves shorter MCMC runs with smaller burn-in and a mixing over c_0 values. This mixing over c_0 could result in more accurate estimates of $S(c)$ for all c in the studied range and could provide robustness to the specification of c_0.

4.2.3 Auxiliary Importance Sampler

The expression (9) in Section 4.2.1 suggests that the importance density $\pi(\boldsymbol{\theta}_\gamma, \boldsymbol{\gamma}|\boldsymbol{y}_{-\kappa(i)}, c_0)$ can be quite different from $\pi(\boldsymbol{\theta}_\gamma, \boldsymbol{\gamma}|\boldsymbol{y}_{-\kappa(i)}, c)$ when c is not close to c_0. Therefore, we specify an auxiliary distribution on c to produce an importance density that marginalizes over the parameter c. The auxiliary distribution results in sampling

from the importance density $\pi(\boldsymbol{\theta}_\gamma, \boldsymbol{\gamma}|\boldsymbol{y}_{-\kappa(i)})$ used to estimate $\pi(y_i|\boldsymbol{y}_{-\kappa(i)}, c)$ for different values of c, where $i = 1, \ldots, n$.

The importance weight is given by

$$w \propto \frac{\pi(\boldsymbol{\theta}_\gamma, \boldsymbol{\gamma}|\boldsymbol{y}_{-\kappa(i)}, c)}{\pi(\boldsymbol{\theta}_\gamma, \boldsymbol{\gamma}|\boldsymbol{y}_{-\kappa(i)})} \propto \frac{\pi(\boldsymbol{y}_{-\kappa(i)}|\boldsymbol{\theta}_\gamma, \boldsymbol{\gamma})\ \pi(\boldsymbol{\beta}_\gamma|\boldsymbol{\gamma}, c)\ \pi(\alpha)\ \pi(\boldsymbol{\gamma})}{\pi(\boldsymbol{y}_{-\kappa(i)}|\boldsymbol{\theta}_\gamma, \boldsymbol{\gamma})\ \pi(\boldsymbol{\beta}_\gamma|\boldsymbol{\gamma})\ \pi(\alpha)\ \pi(\boldsymbol{\gamma})} = \frac{\pi(\boldsymbol{\beta}_\gamma|\boldsymbol{\gamma}, c)}{\pi(\boldsymbol{\beta}_\gamma|\boldsymbol{\gamma})}. \quad (10)$$

We adopt the Inverse Gamma distribution with shape parameter a, scale parameter b and density function

$$\pi(c) = \frac{b^a}{\Gamma(a)}\, c^{-(a+1)}\ \exp\left\{-\frac{b}{c}\right\}, \qquad c > 0 \qquad \text{and} \qquad a, b > 0$$

as the auxiliary distribution on c. The conditional distribution of the regression coefficients $\pi(\boldsymbol{\beta}_\gamma|\boldsymbol{\gamma})$ is then given by

$$\pi(\boldsymbol{\beta}_\gamma|\boldsymbol{\gamma}) = \frac{\Gamma(\frac{p_\gamma}{2} + a)\ b^a}{(2\pi)^{p_\gamma/2}\ \Gamma(a)} \left(\frac{\boldsymbol{\beta}_\gamma{}'\boldsymbol{\beta}_\gamma}{2} + b\right)^{-(\frac{p_\gamma}{2}+a)}$$

and the full conditional distribution of c is given by

$$c|\boldsymbol{\beta}_\gamma, \boldsymbol{\gamma}, \boldsymbol{y}_{-\kappa(i)} \sim \text{IG}\left(p_\gamma/2 + a, \boldsymbol{\beta}_\gamma{}'\boldsymbol{\beta}_\gamma/2 + b\right).$$

We have experimented with other auxiliary distributions, but we found the Inverse Gamma specification described above to be our preferred choice. The Truncated Inverse Gamma leads to similar results, but is computationally slightly less efficient and requires the choice of a truncation point. The Gamma distribution leads to less accurate results, and the Gamma-Inverse Gamma distribution achieves similar accuracy but at the cost of substantially higher computational demands. Finally, if we choose the parameters of the Inverse Gamma in such a way that the tails are thinner and we try to concentrate the mass on the region of interest for c, we find less accurate results that are comparable to those obtained with a Gamma auxiliary distribution.

4.2.4 Comparison

The $K-$fold log predictive score $S(c)$ is estimated using the standard importance sampler at the 12 equally spaced points stated in Section 3, for $c_0 = 1,\ 10,\ 50,\ 100,\ 150$. We use $500,000$ iterations after a burn-in period of $100,000$ and record every 5^{th} draw, resulting in an MCMC sample size T of $80,000$. The average estimated ESS

is high when c is close to c_0 and low for the other values of c. This indicates that the importance density $\pi(\boldsymbol{\theta}_\gamma, \boldsymbol{\gamma}|\boldsymbol{y}_{-\kappa(i)}, c_0)$ is quite different from the target density $\pi(\boldsymbol{\theta}_\gamma, \boldsymbol{\gamma}|\boldsymbol{y}_{-\kappa(i)}, c)$ when c is not close to c_0 and this may results in estimates of $S(c)$ with large variances. However, there are values of c_0 which result in quite accurate estimates of the log predictive score. Figure 6 displays the Arthritis log predictive

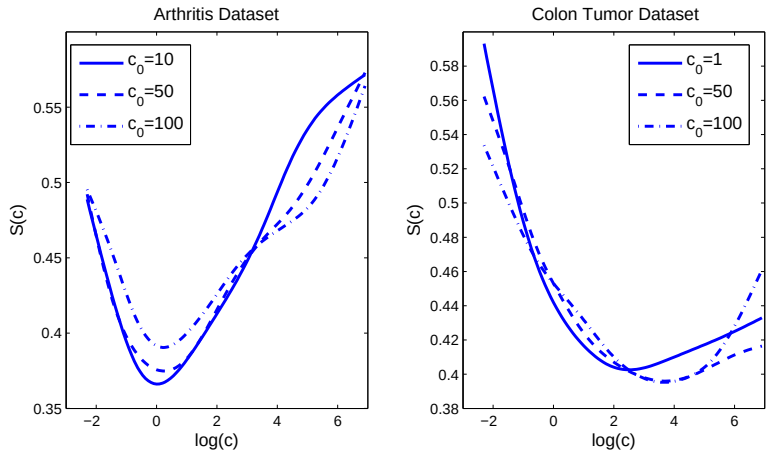

Figure 6: $K-$fold standard importance estimates of the Arthritis and Colon Tumour log predictive scores for selected values of c_0, averaged over 5 replications

score estimates for $c_0 = 10, 50, 100$ (left-hand panel) and the Colon Tumour log predictive score estimates for $c_0 = 1, 50, 100$ (right-hand panel), averaged over the 5 runs. These values of c_0 were specifically selected so that the plots are quite similar to Figure 4 and do not underestimate the log predictive scores for large values of c.

Table 3 presents the average (over the 5 replications) CPU time in seconds of each $K-$fold standard importance sampler, the average sum of squared differences between the importance and MCMC estimates of $S(c)$ and the number of times (out of 5 replications) that the importance minimizer of $S(c)$ is the same (*i.e.* selecting the same of the 12 equally spaced points in the log scale in [0.1,1000]) as the MCMC minimizer. The sum of squared differences (SS) is evaluated at the 12 equally spaced points used for $\log(c)$ and is a measure of the accuracy of the importance samplers.

Thus, some $K-$fold standard importance samplers estimate the log predictive score with virtually the same accuracy as the MCMC methodology. Moreover, the log predictive minimizers of these samplers can be quite close to the MCMC log predictive minimizers. These results suggest the default values $c_0 = 50, 100$. Finally, the required CPU time is more than ten time smaller indicating a ten-fold

Arthritis

c_0	CPU	SS	SMin
1	25,612	0.03	5
10	25,675	0.03	5
50	25,588	0.01	4
100	25,566	0.02	5
150	26,057	0.03	2

Colon Tumour

c_0	CPU	SS	SMin
1	11,723	0.006	1
10	11,586	0.014	0
50	11,612	0.007	4
100	11,564	0.013	5
150	11,621	0.011	4

Table 3: The average CPU time in seconds of the standard importance samplers $\pi(\boldsymbol{\theta}_\gamma, \boldsymbol{\gamma}|\boldsymbol{y}_{-\kappa(i)}, c_0)$ for some c_0 values, the average sum of squares between the importance and MCMC estimates of $S(c)$ and the number of times (out of 5 replications) the importance minimizer of $S(c)$ is the same as that with MCMC

improvement over the MCMC methodology.

The results are potentially sensitive to the choice of c_0 and this leads us to develop a smoothed estimator through the mixed importance sampler. We use $M = 20$ with c_m chosen equally spaced in the log scale from 0.1 to 1000 and generate an MCMC sample from the importance density $\pi(\boldsymbol{\theta}_\gamma, \boldsymbol{\gamma}|\boldsymbol{y}_{-\kappa(i)}, c_m)$ for each i and c_m. The log predictive score $S(c)$ is estimated at the 12 equally spaced points stated in Section 3. Three mixture importance samplers have been used with different run lengths,

Arthritis

Sampler	Burn-in	Sample	CPU	SS	SMin
1	50,000	30,000	173,880	0.003	5
2	20,000	16,000	86,025	0.006	4
3	20,000	6000	42,739	0.013	5

Colon Tumour

Sampler	CPU	SS	SMin
1	77,331	0.009	4
2	38,547	0.014	4
3	19,191	0.04	3

Table 4: The specifications of the MCMC samplers involved in each mixture importance sampler with the average CPU time in seconds of each mixture importance sampler, the average sum of squares between the importance and MCMC estimates of $S(c)$ and the number of times the importance minimizer of $S(c)$ equals that with MCMC

described in Table 4. In each case the chain was thinned every fifth value. Figure 7 displays the importance estimates of the Arthritis and Colon Tumour log predictive

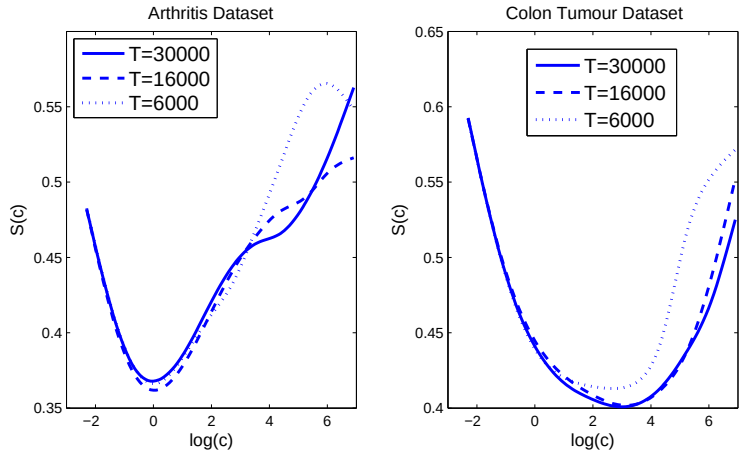

Figure 7: Importance estimates of the Arthritis and Colon Tumour log predictive scores for each mixture importance sampler, averaged over 5 replications. The sample size of the MCMC samplers involved in each mixture importance sampler is denoted by T

scores for each mixture importance sampler, averaged over 5 replications. These log predictive scores are quite similar to the MCMC log predictive score depicted in Figure 4. Table 4 presents the average CPU time of each mixture importance sampler, the average sum of squares between the importance and MCMC estimates of $S(c)$ and the number of coinciding minimizers.

We conclude that the mixture importance samplers estimate the log predictive score with quite similar accuracy as the direct MCMC methodology and lead to very similar minimizers. The CPU times of the second and third sampler are a factor 3 and almost 6.5 smaller than for the direct MCMC method. The third mixture importance sampler estimates the log predictive score with quite similar accuracy as the $K-$fold standard importance samplers with an "optimal" value of c_0 but is computationally less efficient. However, it does not require the difficult task of choosing c_0 and is thus a more "automatic" procedure.

The auxiliary importance sampler offers an alternative method to combine different values of c in the importance sampling distribution. An MCMC sample with stationary distribution $\pi(\boldsymbol{\theta}_\gamma, \boldsymbol{\gamma}|\boldsymbol{y}_{-\kappa(i)})$ was generated. The number of iterations was $500,000$, the burn-in period $100,000$ and the thinning of 5 resulted in an MCMC sample of size $T = 80,000$. Different Inverse Gamma auxiliary distributions on c have been used with shape parameter $a = 0.001$ and scale parameters $b = 1, 0.1, 0.02$. These parameters yield heavy tailed density functions and are not specifically cho-

sen to concentrate the mass on the range of c over which the log predictive score is estimated. The Arthritis and Colon Tumour log predictive scores are estimated at the values of c stated in Section 3, for each Inverse Gamma auxiliary distribution.

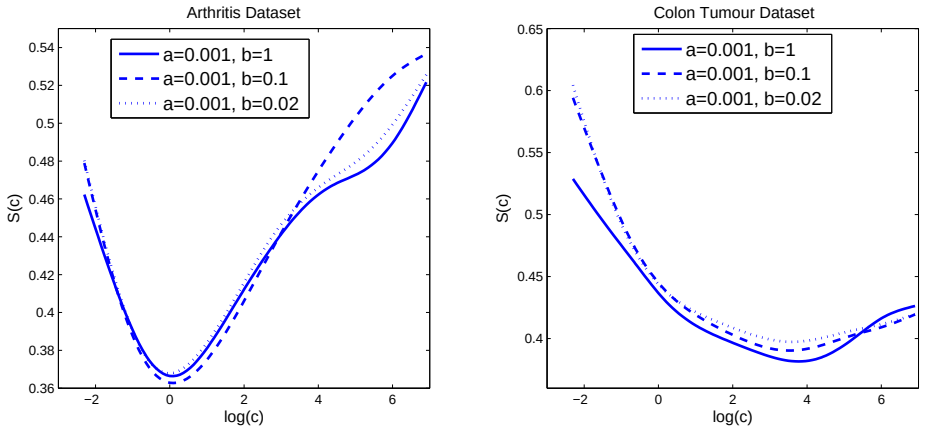

Figure 8: Auxiliary importance estimates of the Arthritis and Colon Tumour log predictive scores for each Inverse Gamma auxiliary distribution on c, averaged over 5 replications

Figure 8 displays the auxiliary importance estimates of the Arthritis and Colon Tumour log predictive scores for three typical Inverse Gamma auxiliary distributions on c, averaged over 5 replications. These log predictive scores are quite similar to the direct MCMC results depicted in Figure 4.

Arthritis

IG$(a = 0.001, b)$	CPU	SS	SMin
$b = 1$	28,494	0.01	5
$b = 0.1$	28,846	0.008	5
$b = 0.02$	28,971	0.004	5

Colon Tumour

IG$(a = 0.001, b)$	CPU	SS	SMin
$b = 1$	12,923	0.006	4
$b = 0.1$	13,012	0.012	4
$b = 0.02$	12,991	0.007	4

Table 5: The average CPU time in seconds of each Inverse Gamma auxiliary importance sampler, the average sum of squares between the importance and MCMC estimates of $S(c)$ and the number of times the importance minimizer of $S(c)$ is the same as that with MCMC

Table 5 presents the average CPU time for each Inverse Gamma auxiliary importance sampler, the average sum of squares between the importance and MCMC

estimates of $S(c)$ and the indicator of the same minimizer. The results show that the Inverse Gamma auxiliary importance samplers estimate the log predictive score with similar accuracy as the direct MCMC method. Values for the sum of squared differences (SS) are very small throughout. Moreover, log predictive minimizers are very similar to those from direct MCMC. The CPU time of these samplers is about ten times smaller than with the MCMC methodology and considerably less than with the mixture importance sampler, indicating a substantial computational gain.

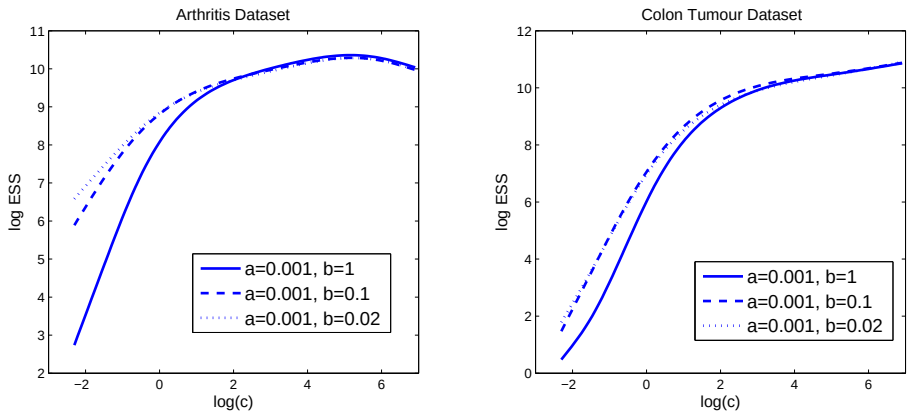

Figure 9: The average log mean ESS of the auxiliary importance samplers at some values of c for the Arthritis and Colon Tumour data, with Inverse Gamma auxiliary distribution on c

Figure 9 shows the average (over 5 replications) log mean (over i) ESS of the Inverse Gamma auxiliary importance samplers at each c for the Arthritis and Colon Tumour datasets. Clearly, mean ESS is an increasing function of c, quite in contrast to the standard importance samplers. Also, we can see that the Inverse Gamma auxiliary distributions with scale parameters $b = 0.1$ and 0.02 result in reasonable high mean ESS over the entire range of c.

5 Conclusions

The "ridge" hyperparameter c crucially affects Bayesian variable selection in probit regression with $p >> n$. In particular, it controls the amount of shrinkage of the regression coefficients and when there is less regularisation (large c) the best models fit the data perfectly. This results in variable selection that discriminates

perfectly within-sample but may not discriminate between the groups out-of-sample. Therefore, we propose to use a predictive criterion like the log predictive score to determine the value of c. In our examples the log predictive score is roughly convex and the value of c that minimizes the log predictive score is the preferred choice for c. Alternative proper score functions lead to very similar minimizers. Since cross-validation densities are employed to determine c the resulting Bayesian variable selection has better out-of-sample predictive properties. The latter is typically linked to successful variable selection, which is our main concern in the type of applications considered here.

In this paper we have focused on the accurate and efficient estimation of the log predictive score and thus the identification of the log predictive score minimizer. The cross-validation density $\pi(y_i|\boldsymbol{y}_{-\kappa(i)}, c)$ is the main component of all predictive scores, but it does not have a closed analytical expression. Therefore, we employ importance sampling methods that use the same sample (generated from the importance density) repeatedly to estimate $\pi(y_i|\boldsymbol{y}_{-\kappa(i)}, c)$ for different i and c. Importance samplers that condition on the entire sample result in inaccurate estimates of the log predictive score. This is mainly a consequence of the perfect fit to the data for large values of c which results in an overestimation of $\pi(y_i|\boldsymbol{\theta}_\gamma, \boldsymbol{\gamma})$. Thus, we propose to use $K-$fold importance samplers with importance densities $\pi(\boldsymbol{\theta}_\gamma, \boldsymbol{\gamma}|\boldsymbol{y}_{-\kappa(i)}, c_0)$ and $\pi(\boldsymbol{\theta}_\gamma, \boldsymbol{\gamma}|\boldsymbol{y}_{-\kappa(i)})$ to estimate $\pi(y_i|\boldsymbol{y}_{-\kappa(i)}, c)$ for different values of c.

The $K-$fold standard importance sampler can result in quite accurate estimates of the Arthritis and Colon Tumour log predictive scores for some values of c_0. The CPU time for this sampler is almost ten time smaller than that required for the direct MCMC methodology. A potential guideline for choosing an appropriate value of c_0 suggests the values $c_0 = 50, 100$, however a mis-specified choice of c_0 can lead to misleading estimates of $S(c)$. Thus, we introduce the $K-$fold mixture and auxiliary importance samplers, which avoid choosing a particular value for c_0.

The $K-$fold mixture importance sampler involves shorter run MCMC chains and mixes over c_0 values, resulting in a six-fold improvement in CPU over the direct MCMC methodology. The $K-$fold auxiliary importance samplers provide quite accurate estimates of the Arthritis and Colon Tumour log predictive scores with a ten-fold computational improvement over the MCMC approach. The preferred choice for the auxiliary distribution is an Inverted Gamma with small values for both parameters.

Thus, we suggest employing the $K-$fold mixture and Inverse Gamma auxiliary

importance samplers to estimate the log predictive score and find the best value for c. The parameters of the Inverse Gamma auxiliary distributions on c are chosen to yield heavy tail density functions and there is no need for further user input. The mixture importance sampler requires predetermined values $c_1, \ldots, c_M$ and we recommend choosing them to be equally spaced in the logarithmic scale and to cover the relevant range of c with $M = 20$.

The procedures described should also work well in other cross-validation contexts, such as random-fold cross-validation (Gneiting and Raftery, 2007). We also successfully used both procedures on a much larger dataset regarding prostate cancer, described in Singh et al. (2002), which has $n = 136$ observations with $p = 10150$ potential covariates. Here the CPU demand of the direct MCMC was of the order of 5.5 days (with $K = 12$), which was reduced to 0.5 days by using the auxiliary importance sampler, representing an 11-fold decrease in computational effort. The improvements in computational efficiency would be even more pronounced if the log predictive score is estimated at a larger number of points l.

Alternatively, we could address the uncertainty in c by a fully Bayesian specification that places a prior on c. In the context of linear splines, the formal Bayesian approach for the ridge prior is studied by Denison et al. (2002) and for linear regression with a g-prior it is studied by Celeux et al. (2006), Bottolo and Richardson (2007), Liang et al. (2008) and Cui and George (2008). This approach is believed to increase robustness to the specification of c. Implementing a fully Bayesian procedure in generalized linear regression with $p >> n$ is the topic of ongoing research.

References

Alon, U., N. Barkai, and D. A. Notterman (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences of the United States of America 96*, 6745–6750.

Bernardo, J. and A. Smith (1994). *Bayesian Theory*. Wiley: Chichester.

Bottolo, L. and S. Richardson (2007). Fully Bayesian variable selection using g-priors. Technical Report.

Brown, P. and M. Vanucci (1998). Multivariate Bayesian variable selection and prediction. *Journal of the Royal Statistical Society 60*(3), 627–641.

Celeux, G., J.-M. Marin, and C. Robert (2006). Sélection bayésienne de variables en régression linéaire. *Journal de la Société Française de Statistique 147*, 59–79.

Cui, W. and E. I. George (2008). Empirical Bayes vs. Fully Bayes variable selection. *Journal of Statistical Planning and Inference 138*, 888–900.

Denison, D. G., C. C. Holmes, B. K. Mallick, and A. F. Smith (2002). *Bayesian Methods for Nonlinear Classification and Regression*. John Wiley and Sons.

Fernández, C., E. Ley, and M. F. Steel (2001). Benchmark priors for Bayesian model averaging. *Journal of Econometrics 100*, 381–427.

Geisser, S. and W. Eddy (1979). A predictive approach to model selection. *Journal of American Statistical Association 74*, 153–160.

Gelfand, A. E. and D. K. Dey (1994). Bayesian model choice: Asymptotics and exact calculations. *Journal of the Royal Statistical Society, Series B 56*, 501–514.

Gelfand, A. E., D. K. Dey, and H. Chang (1992). Model determination using predictive distributions with implementation via sampling-based methods. *Bayesian Statistics 4*, 147–167.

Gneiting, T. and A. E. Raftery (2007). Strictly proper scoring rules, prediction and estimation. *Journal of the American Statistical Association 102*, 359–378.

Good, I. (1952). Rational decisions. *Journal of the Royal Statistical Society, Series B 14*(1), 107–114.

Hoeting, J. A., D. Madigan, A. Raftery, and C. T. Volinsky (1999). Bayesian Model Averaging: A tutorial. *Statistical Science 14*(4), 382–417.

Holmes, C. C. and L. Held (2006). Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Analysis 1*(1), 145–168.

Ley, E. and M. F. Steel (2009). On the effect of prior assumptions in Bayesian Model Averaging with applications to growth regression. *Journal of Applied Econometrics 24*, 651–674.

Liang, F., R. Paulo, G. Molina, M. A. Clyde, and J. Berger (2008). Mixture of $g-$priors for Bayesian variable selection. *Journal of the American Statistical Association 103*, 410–423.

Liu, J. S. (2001). *Monte Carlo Strategies in Scientific Computing*. Springer-Verlag, New York.

Robert, C. and G. Casella (2004). *Monte Carlo Statistical Methods* (Second ed.). Springer, New York.

Sha, N., M. Vanucci, P. Brown, M. Trower, and G. Amphlett (2003). Gene selection in arthritis classification with large-scale microarray expression profiles. *Comparative and Functional Genomics 4*, 171–181.

Sha, N., M. Vanucci, M. G. Tadesse, P. J. Brown, I. Dragoni, N. Davies, T. C. Roberts, A. Contestabile, M. Salmon, C. Buckley, and F. Falciani (2004). Bayesian variable selection in multinomial probit models to identify molecular signatures of disease stage. *Biometrics 60*, 812–819.

Singh, D., P. G. Febbo, K. Ross, D. Jackson, J. Manola, C. Ladd, P. Tamayo, A. Renshaw, A. V. D'Amico, J. P. Richie, E. S. Lander, M. Loda, P. Kantoff, T. Golub, and W. Sellers (2002). Gene expression correlates of clinical prostate cancer behavior. *Cancer cell 1*, 203–209.