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

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

January 5, 2011

**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 scale parameter $c$ in the standard variable selection prior as the minimizer of the log predictive score. Naive evaluation of the log predictive score requires substantial computational effort, and we investigate computationally cheaper methods using importance sampling. 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 can take the values 0 or 1. For example, we may want to find genes that discriminate between two

*Department of Statistics, University of Warwick, Coventry, CV4 7AL, U.K. and † School 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

disease states using samples taken from patients in the first disease state ($y_i = 1$) or the second one ($y_i = 0$). Typically, the number of measured gene expressions (covariates), say $p$, will be much larger than the number of samples, say $n$. A popular approach to this problem is variable selection in a probit regression model (Sha et al., 2004; Lee et al., 2003)[1]. Usually, it is assumed that the response $\boldsymbol{y}$ can be modelled in terms of a (small) subset of the $p$ covariates. The $2^p$ possible subset choices define different models which are indexed by the vector $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)$ where $\gamma_j = 1$ if the $j$-th predictor is included or $\gamma_j = 0$ if it is excluded from the model. The number of variables included in a model is denoted by $p_{\boldsymbol{\gamma}} = \sum_{j=1}^{p} \gamma_j$. Let $\mathbf{x}_{\boldsymbol{\gamma}i}$ be a $1 \times p_{\boldsymbol{\gamma}}$ vector whose $j$-th entry is the measurement of the $j-$th included covariate (after centring) for the $i-$th individual and let $\mathbf{X}_{\boldsymbol{\gamma}} = (\mathbf{x}'_{\boldsymbol{\gamma}1}, \ldots, \mathbf{x}'_{\boldsymbol{\gamma}n})'$ be the $n \times p_{\boldsymbol{\gamma}}$ design matrix of model $\boldsymbol{\gamma}$. Under model $\boldsymbol{\gamma}$, it is assumed that

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

where $\Phi$ is the cumulative distribution function of a standard normal random variable, $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)'$ is a vector of linear predictors, $\mathbf{1}$ represents an $n \times 1$-dimensional vector of ones, $\alpha$ is the intercept and $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ is a $p_{\boldsymbol{\gamma}} \times 1$-dimensional vector of regression coefficients. We will assume that $p \gg n$ and denote the model parameters by $\boldsymbol{\theta}_{\boldsymbol{\gamma}} = (\alpha, \boldsymbol{\beta}'_{\boldsymbol{\gamma}})' \in \boldsymbol{\Theta}_{\boldsymbol{\gamma}}$. In this framework, identification of discriminating genes reduces to finding a suitable model $\boldsymbol{\gamma}$.

In this paper, a Bayesian framework is adopted to deal with the uncertainty regarding the inclusion of covariates. The prior is assumed to have a product form $\pi(\alpha, \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}) = \pi(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\boldsymbol{\gamma})\pi(\alpha)\pi(\boldsymbol{\gamma})$. The intercept $\alpha$ represents the overall mean of the linear predictors since the covariates have been centred and is regarded as a common parameter to all models. Thus, a non-informative improper prior could be used for $\alpha$, as *e.g.* in Fernández et al. (2001). However, we will follow Sha et al. (2004) and Brown and Vannucci (1998) by assuming that $\alpha \sim N(0, h)$. The prior distribution for the regression coefficients $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ is the so-called ridge prior

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

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 $gn(\mathbf{X}_{\boldsymbol{\gamma}}{}' \mathbf{X}_{\boldsymbol{\gamma}})^{-1}$

[1] Alternatively, a logistic regression approach is described by Zhou et al. (2004).

could be used, as in Liang et al. (2008) or the compromise choice $(c^{-1}\mathbf{I}_{p_\gamma}+(gn)^{-1}\mathbf{X}_{\boldsymbol{\gamma}}{}'\mathbf{X}_{\boldsymbol{\gamma}})^{-1}$ which is defined when $p_{\boldsymbol{\gamma}} \geq n$ (unlike the $g$-prior). Finally, we assume that each regressor is included independently with probability $q$, which implies that

$$\pi(\boldsymbol{\gamma}) = q^{p_\gamma}(1-q)^{p-p_\gamma}$$

and $p_{\boldsymbol{\gamma}}$ is binomially distributed, $p_{\boldsymbol{\gamma}} \sim \text{Bin}(p, q)$. Alternatively, Dobra (2009) suggests defining a maximum model size $p_{max}$ and only consider models for which $p_\gamma < p_{\max}$, with a uniform prior placed across all such models.

The choice of the hyperparameters $q$ and $c$ is critical for posterior inference on the model space since $q$ plays the main role in inducing a model size penalty and $c$ regularises the regression coefficients. The hyperparameter $q$ has a natural interpretation as the prior proportion of variables included in the model. Uncertainty about $q$ could be incorporated by choosing a hyperprior for $q$, as discussed by Scott and Berger (2006), which allows the prior to adapt more easily to model size. However, $c$ is harder to choose.

In this work we focus on an empirical Bayes choice of the hyperparameter $c$ using cross-validation. George and Foster (2000) discuss the application of empirical Bayes methods for estimating $c$ and $q$ by maximizing the marginal likelihood. In variable selection for gene expression data, Strimenopoulou and Brown (2008) describe an empirical Bayes method for maximum *a posteriori* estimation. Fully Bayesian analysis would place a prior distribution on $c$ and so allow the inclusion of any uncertainty about $c$ in predictions. However, there is often little prior information about $c$ and vague priors are routinely used. Cui and George (2008) find that empirical Bayes approaches provide an adaptive choice for the $g-$prior hyperparameter in Bayesian linear regression and outperform fully Bayesian analysis that places a prior on $c$. Our results point in the same direction: prediction using a diffuse proper prior on $c$ are worse than those using an "optimal" choice of $c$, even when the prior has ample mass close to the optimal value.

The optimal value of $c$ is chosen to minimize the log predictive score (Good, 1952), which is used as a measure of predictive performance. The score uses cross-validation which has been extensively used in statistics as discussed by e.g. Hastie et al. (2001) and can be justified in a decision theoretic framework if a prior over the model space cannot be specified (Key et al., 1999). 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 $\pi(y_i|\boldsymbol{y}_{-i}, c)$, where $\boldsymbol{y}_{-i} = (y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n)$, is

the main component of all predictive scores. This cross-validation density does not have a closed analytic expression in our context and therefore we propose various novel importance samplers to estimate it. In comparison to the direct 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 $c$. These proposed importance samplers lead to accurate estimates of the optimal value for $c$ with a very considerable saving in computational effort.

The paper is organized as follows: Section 2 discusses the importance of the choice of the hyperparameter $c$ in Bayesian Model Averaging (BMA) for probit regression with $p \gg n$, while Section 3 describes the cross-validation approach and estimates of the log predictive score for some gene expression datasets from DNA microarray studies. Section 4 introduces the novel importance samplers used here. Section 5 evaluates and compares the accuracy and efficiency of these samplers in estimating the log predictive score, and compares our empirical Bayes approach with a full Bayesian analysis. Finally, Section 6 contains some concluding comments including guidelines for the implementation of these samplers that optimize their efficiency and make them more or less automatic procedures. 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 probit regression with $p \gg n$. A value of $c$ that is too small leads to overshrinkage and bad out-of-sample prediction but a value of $c$ that is too large leads to Lindley's paradox (Shafer, 1982) where the smallest model (the model with no regressors) is favoured regardless of the data. This suggests that there are values of $c$ between these extremes that lead to good out-of-sample prediction. We illustrate the effect of $c$ using a study of rheumatoid arthritis and osteoarthritis sufferers where $p = 755$ gene expression measurements were taken on $n = 31$ patients (Sha et al., 2003). We choose $h = 100$ and $q = 5/755 = 0.0066$, implying that the prior mean number of included variables is 5. The posterior $\pi(\boldsymbol{\theta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}|\boldsymbol{y}, c)$ was sampled using the Metropolis-Hastings algorithm of Holmes and Held (2006) (which will be used

throughout the paper). Five independent chains were run generating an MCMC sample of size $T = 190,000$, which is the MCMC sample left after a burn-in period of $100,000$ and a thinning to every tenth draw. In what follows the MCMC samplers will have the same burn-in and thinning, unless otherwise stated. This run length was sufficient for strong agreement between the results for the five chains.

| 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: IDs of genes in the Arthritis dataset included in the ten models with the highest posterior probability for different values of $c$. Boxed genes are selected for all $c$.

Table 1 reports the genes that appeared in the ten highest posterior probability models for different values of $c$ (the posterior probabilities were calculated from the combination of five independent MCMC replications). Genes 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 $c$, indicating substantial differences in variable selection for different values of $c$. There are also substantial differences in posterior inclusion probabilities for $c = 1$ and $c = 100$, as illustrated in Figure 1. For example, Gene 290 has posterior inclusion probability 0.45 when $c = 1$ but 0.2 for $c = 100$. On the other hand, gene 258 has posterior inclusion probability 0.15 when $c = 1$ but 0.4 for $c = 100$. The scatter-plots show that these differences occur with many genes and indicate substantial 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. The average absolute coefficient size, *i.e.*

$$|\boldsymbol{\beta}_{\boldsymbol{\gamma}}| = \frac{1}{p_{\boldsymbol{\gamma}}} \sum_{j=1}^{p_{\boldsymbol{\gamma}}} |\beta_{\boldsymbol{\gamma},j}|,$$

where $\beta_{\boldsymbol{\gamma},j}$ are the components of the regression coefficient vector $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$, can be used to judge the level of regularisation. The posterior density, graphed in the left panel

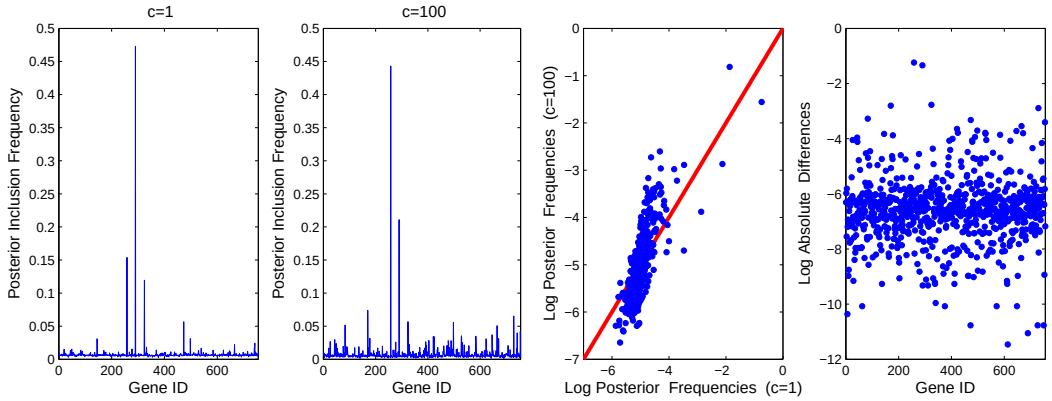


Figure 1: Estimated posterior gene inclusion probabilities, scatter-plot of the logarithms and the log absolute differences of the estimated posterior gene inclusion probabilities of the Arthritis data for different values of c.

of Figure 2, shows probability mass at larger values of $|\boldsymbol{\beta}_{\boldsymbol{\gamma}}|$ increasing with $c$. This suggests that large values of $c$ may lead to poor predictions since large values of $|\boldsymbol{\beta}_{\boldsymbol{\gamma}}|$ are often associated with overfitting. The right panel of Figure 2 displays the

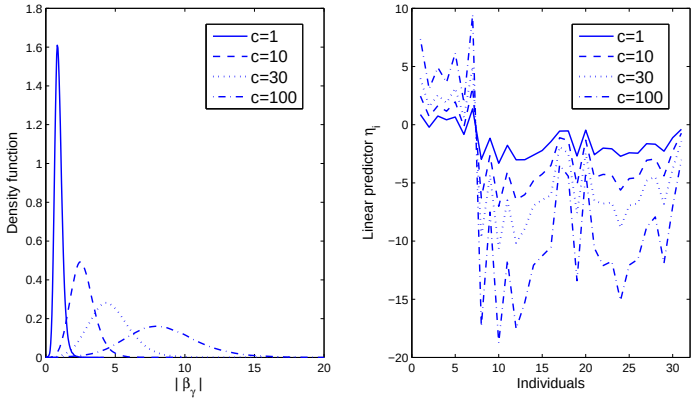


Figure 2: Arthritis data: The left panel displays the posterior density function of $|\boldsymbol{\beta}_{\boldsymbol{\gamma}}|$ for different values of the hyperparameter $c$. The right panel shows the posterior mean of the linear predictor $\eta_i$ for each individual $i$ for different values of $c$.

posterior mean, $\hat{\eta}_i$, of the linear predictor variable $\eta_i = \alpha + \boldsymbol{x}_{\boldsymbol{\gamma} i}\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ 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 $\hat{\eta}_i$ 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

$\Phi(\hat{\eta}_i)$ are very close to 1 when $y_i = 1$ and very close to 0 otherwise. Therefore, the posterior places more mass on models that perfectly discriminate the $n$ observations into the two groups when there is less regularisation (large $c$) on the regression coefficients. However, perfect model fit typically leads to poor predictions and we need to carefully consider the specification of $c$.

The choice of $c$ also strongly affects the posterior distribution of the intercept $\alpha$, shown in Figure 3. The absolute value of the posterior mode and the variance of $\alpha$ clearly increases with $c$. This is a direct consequence of the perfectly fitting models associated with large $c$ since moderate changes to $\alpha$ will leave all $\hat{\eta}_i$ in the tails of the standard normal distribution.

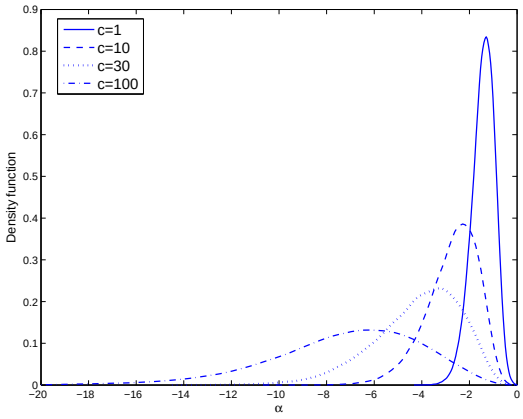


Figure 3: Arthritis data: Posterior density function of the intercept for different values of the hyperparameter $c$. The prior distribution on $\alpha$ is $N(0, 100)$.

These features of the inference are common to many gene expression data sets. For example, we found similar results with the Colon Tumour dataset described by Alon et al. (1999), which contains $n = 62$ observations of tumour and normal colon groups with $p = 1224$ (setting $q = 5/1224 = 0.0041$).

# 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. Fernández et al. (2001) use a log predic-

tive score to evaluate different choices for the $g$-prior hyperparameter in Bayesian linear regression. In the typical areas of application we consider in this paper, the key concern is often variable selection, but good predictive performance tends to be linked to successful variable selection.

In our context, the log predictive score suggested by Gelfand et al. (1992) would be

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

where $\pi(y_i|\boldsymbol{y}_{-i}, c)$ is the cross-validation density mentioned in the Introduction. In a pairwise model comparison this results in the log pseudo-Bayes factor (Geisser and Eddy, 1979). Calculating this leave-one-out cross-validation criterion may be computationally intensive in practice since it involves fitting the model to $n$ different subsets of the data. An alternative is $K-$fold cross-validation where the sample is partitioned into $K$ subsets and the score becomes

$$S(c) = -\frac{1}{n}\sum_{i=1}^{n} \ln \pi(y_i|\boldsymbol{y}_{-\kappa(i)}, c) \tag{2}$$

where $\kappa(i) \in \{1, \ldots, K\}$ represents the partition to which $y_i$ is allocated, and $\boldsymbol{y}_{-\kappa(i)}$ are the observations from the remaining partitions. The random-fold cross-validation of Gneiting and Raftery (2007) (which corresponds to the Bayes factor) could also be considered. The value of $c$ that minimizes $S(c)$ will be our preferred choice for $c$. Other proper score functions for binary variables (Gneiting and Raftery, 2007) could replace the logarithmic score function in (2). In the present paper, we also investigate the use of the quadratic or Brier predictive score and the spherical 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 $i$-th individual is given by

$$\pi(y_i|\boldsymbol{y}_{-\kappa(i)}, c) = \sum_{\boldsymbol{\gamma}} \int_{\boldsymbol{\Theta}_\gamma} \pi(y_i|\boldsymbol{\theta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma})\, \pi(\boldsymbol{\theta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}|\boldsymbol{y}_{-\kappa(i)}, c)\, d\boldsymbol{\theta}_{\boldsymbol{\gamma}} = \mathbb{E}[\pi(y_i|\boldsymbol{\theta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma})]\,, \tag{3}$$

where the expectation is taken with respect to the joint posterior distribution $\pi(\boldsymbol{\theta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}|\boldsymbol{y}_{-\kappa(i)}, c)$. It does not have a closed analytic expression but can be estimated by

$$\hat{\pi}(y_i|\boldsymbol{y}_{-\kappa(i)}, c) = \frac{1}{T}\sum_{j=1}^{T} \Phi(\tilde{\mathbf{x}}_{\boldsymbol{\gamma}i}\boldsymbol{\theta}_{\boldsymbol{\gamma}}^{(j)})^{y_i}\, (1 - \Phi(\tilde{\mathbf{x}}_{\boldsymbol{\gamma}i}\boldsymbol{\theta}_{\boldsymbol{\gamma}}^{(j)}))^{1-y_i}, \tag{4}$$

where $(\boldsymbol{\theta}_{\boldsymbol{\gamma}}^{(1)}, \boldsymbol{\gamma}^{(1)}), \ldots, (\boldsymbol{\theta}_{\boldsymbol{\gamma}}^{(T)}, \boldsymbol{\gamma}^{(T)})$ is an MCMC sample with stationary distribution $\pi(\boldsymbol{\theta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}|\boldsymbol{y}_{-\kappa(i)}, c)$ and $\tilde{\mathbf{x}}_{\boldsymbol{\gamma}i} = (1, \mathbf{x}_{\boldsymbol{\gamma}i})$ is a $1 \times (p_{\boldsymbol{\gamma}}+1)$-dimensional vector. The MCMC estimate of the log predictive score is given by replacing $\pi(y_i|\boldsymbol{y}_{-\kappa(i)}, c)$ in (2) by $\hat{\pi}(y_i|\boldsymbol{y}_{-\kappa(i)}, c)$.

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 for the two datasets. 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. Applying their guidelines leads to a range of (0.1, 2.27) for the Arthritis dataset and (0.1, 2.26) for the Colon Tumour dataset. MCMC samples of size $T = 80,000$ (after thinning to every fifth draw) were generated for each data partition in the sum in (2) and each value of $c$. 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).

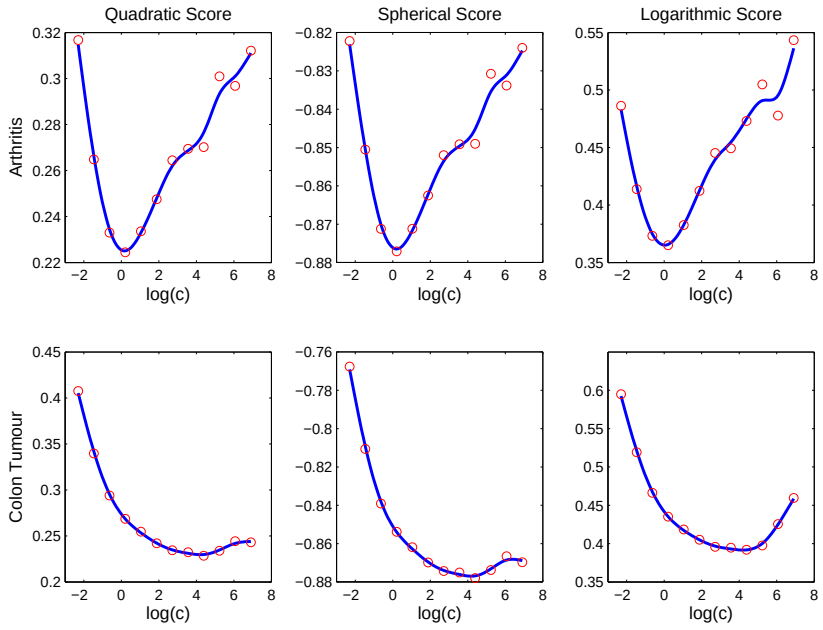


Figure 4: MCMC estimates and smooth estimated curves of different predictive score functions for the Arthritis and Colon Tumour datasets.

The right-hand panels of Figure 4 display both the MCMC estimates and a smooth estimated curve for $S(c)$ (estimated using cubic smoothing splines). In both

datasets $S(c)$ is roughly convex and so a unique minimizer can be determined. This value of $c$ is around 1 for the Arthritis dataset, but 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 guideline range for $c$ suggested by Sha et al. (2004) includes the optimal value of $c$ in the case of the Arthritis dataset, but the optimal value of $c$ is well outside this range for the Colon Tumour dataset.

The other panels of Figure 4 display the MCMC estimates and a smooth estimated curve of 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.

This direct MCMC methodology needs $Kl$ MCMC runs for $K$ data partitions to estimate the log predictive score at $l$ points. Table 2 reports the CPU time in minutes needed (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 use the direct MCMC methodology. This motivates us to employ importance sampling methods to estimate $\pi(y_i|\boldsymbol{y}_{-\kappa(i)}, c)$ using fewer MCMC runs. 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 | 4849 |
| Colon Tumour | 2048 |

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

# 4 Computational approaches

The predictive densities needed to calculate $S(c)$ will be estimated using importance sampling (Liu, 2001; Robert and Casella, 2004). In general, this method approxi-

mates the integral

$$\mathbb{E}_f[h(X)] = \int_{\mathcal{X}} h(x)\, f(x)\, dx \Big/ \int_{\mathcal{X}} f(x)\, dx$$

by

$$\sum_{j=1}^{T} w^{(j)} h\left(x^{(j)}\right) \Big/ \sum_{j=1}^{T} w^{(j)}, \tag{5}$$

where a sample $x^{(1)}, \ldots, x^{(T)}$ is generated from a given distribution $g$ and the importance weight is $w^{(j)} = f(x^{(j)})/g(x^{(j)})$. The (possibly unnormalized) densities $f$ and $g$ are called the target and importance density respectively.

The accuracy of the approximation 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 $\text{cv}^2$ 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 (5). 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, this involves choosing a value $c_0$ and using $\pi(\boldsymbol{\theta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}|\boldsymbol{y}, c_0)$ as an importance density (which can be sampled using MCMC) to estimate $\pi(y_i|\boldsymbol{y}_{-\kappa(i)}, c)$, given by (3), for all $i$ and all values of $c$. As this idea implies large potential computational gains, it is the one we investigated first by calculating the ESS for all data partitions and values of $c$. Figure 5 plots the mean ESS over all observations at each $c$ 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 $\pi(\boldsymbol{\theta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}|\boldsymbol{y}, c_0)$ is quite different from $\pi(\boldsymbol{\theta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}|\boldsymbol{y}_{-\kappa(i)}, c)$ when $c_0$ is not close to $c$, resulting in estimates of $S(c)$

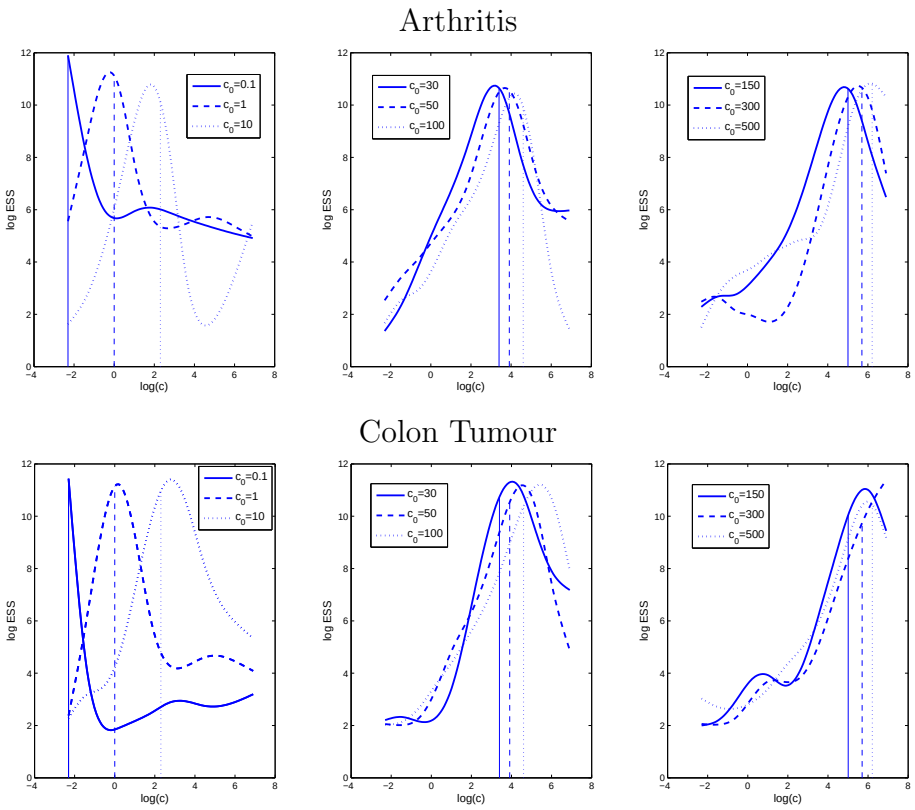


Figure 5: The log mean ESS of the importance densities $\pi(\boldsymbol{\theta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}|\boldsymbol{y}, c_0)$ at some values of $c$ for the Arthritis and Colon Tumour dataset. The vertical lines indicate the $\log c_0$ values.

with high variance. Therefore we only use the importance density $\pi(\boldsymbol{\theta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}|\boldsymbol{y}, c_0)$ when $\pi(\boldsymbol{\theta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}|\boldsymbol{y}_{-\kappa(i)}, c_0)$ is the target density and $\pi(y_i|\boldsymbol{y}_{-\kappa(i)}, c_0)$ is the quantity to be estimated.

Figure 6 displays the resulting importance estimates of the log predictive score $S(c)$ for the Arthritis and Colon Tumour datasets. In comparison with Figure 4 the log predictive scores are underestimated for large $c$ which suggests that $\pi(y_i|\boldsymbol{y}_{-\kappa(i)}, c)$ is overestimated. This is perhaps surprising given the unbiasedness of importance sampling estimates but is an example of "pseudo-bias" (Ventura, 2002). The large number of variables combined with the potential for overfitting (especially for large $c$) means that there can be substantial differences between $\pi(\boldsymbol{\theta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}|\boldsymbol{y}_{-\kappa(i)}, c_0)$ and $\pi(\boldsymbol{\theta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}|\boldsymbol{y}, c_0)$. The importance weights adjust for this difference but some models with substantially more mass under $\pi(\boldsymbol{\theta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}|\boldsymbol{y}_{-\kappa(i)}, c_0)$ than $\pi(\boldsymbol{\theta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}|\boldsymbol{y}, c_0)$ may not

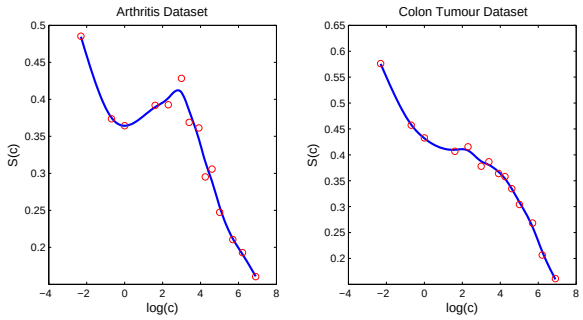


Figure 6: Importance estimates of the log predictive score $S(c)$ for the Arthritis and Colon Tumour datasets.

be sampled and so are excluded from (5). This leads to bad estimates of both numerator and denominator. It seems reasonable to assume that models sampled under $\pi(\boldsymbol{\theta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}|\boldsymbol{y}, c_0)$ will tend to predict $y_i$ better than those sampled under $\pi(\boldsymbol{\theta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}|\boldsymbol{y}_{-\kappa(i)}, c_0)$ leading to the overestimation.

## 4.2 $K$-fold Importance Samplers

An alternative approach to the one taken in the previous subsection uses the posterior conditioned on the correct subset of the data, $\pi(\boldsymbol{\theta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}|\boldsymbol{y}_{-\kappa(i)}, c_0)$, as the importance density when the target density is $\pi(\boldsymbol{\theta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}|\boldsymbol{y}_{-\kappa(i)}, c)$. This approach will be termed a $K-$fold standard importance sampler. The estimates of the log predictive score $S(c)$ are improved but at the cost of longer computing times since $K$ MCMC chains, one for each data partition, must be run. However, the number of chains is still $l$ times smaller than the direct MCMC methodology of Section 3. But choosing a value of $c_0$ is difficult and restricts the range of $c$ for which the log predictive score $S(c)$ can be well estimated. One solution is Deterministic Mixture Sampling (Owen and Zhou, 2000) which combines estimates using different values of $c_0$. In general, suppose that we have $M$ importance sampling densities $g_1, g_2, \ldots, g_M$ (where $g_m$ is a probability density function) and we have samples of $x_m^{(1)}, x_m^{(2)}, \ldots, x_m^{(T)}$ from $g_m$. A new estimator of $\mathbb{E}_f[h(X)]$ is constructed by using the mixture $\tilde{g}(x) = \frac{1}{M}\sum_{m=1}^{M} g_m(x)$ as the importance sampling density leading to the estimate

$$\sum_{m=1}^{M}\sum_{j=1}^{T} w_m^{(j)} h\left(x_m^{(j)}\right) \Bigg/ \sum_{m=1}^{M}\sum_{j=1}^{T} w_m^{(j)}, \tag{6}$$

where the importance weights are $w_m^{(j)} = f\left(x_m^{(j)}\right) \Big/ \tilde{g}\left(x_m^{(j)}\right)$. Our problem is slightly different since we only know $g_m$ up to proportionality and (6) requires the normalizing constant for each $m$. Geyer (1994) describes a version of this estimator for MCMC output and introduces a method for approximating the normalizing constants. We propose two novel importance sampling methods which avoid estimation of normalizing constants.

### 4.2.1 The Auxiliary Importance Sampler

A similar method to Deterministic Mixture Sampling introduces a probability distribution, $\pi_A(c_0)$, for $c_0$ and updates its value in the MCMC sampler. This distribution is not a prior but leads to a heavier-tailed importance density

$$\pi_A(\boldsymbol{\theta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}|\boldsymbol{y}_{-\kappa(i)}) \propto \pi(\boldsymbol{y}_{-\kappa(i)}|\boldsymbol{\theta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma})\pi_A(\boldsymbol{\theta}_{\boldsymbol{\gamma}}|\boldsymbol{\gamma})\pi(\boldsymbol{\gamma})$$

where $\pi_A(\boldsymbol{\theta}_{\boldsymbol{\gamma}}|\boldsymbol{\gamma}) = \pi(\alpha) \int \pi(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\boldsymbol{\gamma}, c_0)\pi_A(c_0)\, dc_0$. We refer to $\pi_A(c_0)$ as an auxiliary distribution and the resulting sampling method as an Auxiliary Importance Sampler. In practice, it is more straightforward to sample $T$ values from

$$\pi_A(\boldsymbol{\theta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}, c_0|\boldsymbol{y}_{-\kappa(i)}) \propto \pi(\boldsymbol{y}_{-\kappa(i)}|\boldsymbol{\theta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma})\pi(\boldsymbol{\theta}_{\boldsymbol{\gamma}}|\boldsymbol{\gamma}, c_0)\pi(\boldsymbol{\gamma})\pi_A(c_0)$$

using MCMC and estimate $\pi(y_i|\boldsymbol{y}_{-\kappa(i)}, c)$ by

$$\hat{\pi}(y_i|\boldsymbol{y}_{-\kappa(i)}, c) = \frac{\sum_{j=1}^{T} w^{(j)}\pi(y_i|\boldsymbol{\theta}_{\boldsymbol{\gamma}}^{(j)}, \boldsymbol{\gamma}^{(j)})}{\sum_{j=1}^{T} w^{(j)}}, \tag{7}$$

where the importance weight for the $j$-th sample is given by

$$w^{(j)} = \frac{\pi\left(\boldsymbol{\beta}_{\boldsymbol{\gamma}}^{(j)} \middle| \boldsymbol{\gamma}^{(j)}, c\right)}{\pi_A\left(\boldsymbol{\beta}_{\boldsymbol{\gamma}}^{(j)} \middle| \boldsymbol{\gamma}^{(j)}\right)},$$

which can easily be calculated if $\pi_A(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\boldsymbol{\gamma})$ has an analytic form. In comparison with the Deterministic Mixture Sampler, this method concentrates more sampling effort on values of $c_0$ which have larger marginal pseudo-posterior density $\pi_A(c_0|\boldsymbol{y}_{-\kappa(i)})$ and so reduces the variance of the estimates of $S(c)$ at those values.

The auxiliary distribution $\pi_A(c_0)$ is chosen to be an Inverse Gamma distribution with shape parameter $a$, scale parameter $b$, denoted by $\mathrm{IG}(a, b)$, with density function

$$\pi_A(c_0) = \frac{b^a}{\Gamma(a)}\, c_0^{-(a+1)}\, \exp\left\{-\frac{b}{c_0}\right\}, \qquad c_0 > 0 \text{ and } a, b > 0.$$

The distribution $\pi_A(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\boldsymbol{\gamma})$ is then a multivariate Student $t$ distribution with density

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

and the full conditional distribution of $c_0$ is given by

$$c_0|\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}, \boldsymbol{y}_{-\kappa(i)} \sim \mathrm{IG}\left(p_{\boldsymbol{\gamma}}/2+a, \boldsymbol{\beta}_{\boldsymbol{\gamma}}{}'\boldsymbol{\beta}_{\boldsymbol{\gamma}}/2+b\right).$$

We have experimented with other auxiliary distributions, but we found the Inverse Gamma specification described above provides the best performance.

### 4.2.2 A Multiple Importance Sampler

The previous method avoids the need to approximate normalizing constants and concentrates sampling effort on promising values of $c_0$. However, the variance of the estimates of $S(c)$ will be increased at values of $c$ which are not supported by the pseudo-posterior. Alternatively, we can use a Multiple Importance Sampler (Veach and Guibas, 1995; Owen and Zhou, 2000) which, again, avoids calculating normalizing constants. We define a positive, increasing sequence $c_1, c_2, \ldots, c_M$ and let $\hat{\pi}_{c_k}(y_i|\boldsymbol{y}_{-\kappa(i)}, c)$ be the importance sampling estimate of $\pi(y_i|\boldsymbol{y}_{-\kappa(i)}, c)$ using the importance sampling density $\pi(y_i|\boldsymbol{y}_{-\kappa(i)}, c_k)$ which leads to the estimator

$$\hat{\pi}_{c_k}(y_i|\boldsymbol{y}_{-\kappa(i)}, c) = \frac{\sum_{j=1}^{T} w_k^{(j)} \pi(y_i|\boldsymbol{\theta}_{\boldsymbol{\gamma}}^{(j)}, \boldsymbol{\gamma}^{(j)})}{\sum_{j=1}^{T} w_k^{(j)}}$$

where the weights are

$$w_k{}^{(j)} = \frac{\pi\left(\boldsymbol{\beta}_{\boldsymbol{\gamma}}^{(j)} \middle| \boldsymbol{\gamma}^{(j)}, c\right)}{\pi\left(\boldsymbol{\beta}_{\boldsymbol{\gamma}}^{(j)} \middle| \boldsymbol{\gamma}^{(j)}, c_k\right)}.$$

Since the values of $c_m$ are ordered and increasing, the last value of the MCMC sample from $\pi(\boldsymbol{\theta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}|\boldsymbol{y}_{-\kappa(i)}, c_m)$ could be used as the initial value of the MCMC chain with stationary distribution $\pi(\boldsymbol{\theta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}|\boldsymbol{y}_{-\kappa(i)}, c_{m+1})$. Therefore the MCMC samplers do not need a long burn-in period. The estimator for each data partition uses a kernel weighted average to combine the estimates at the values $c_1, c_2, \ldots, c_M$ which has the form

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

where $d_m = \log(c) - \log(c_m)$ and $K_\lambda(x)$ is a kernel with window size parameter $\lambda$ for which $K_\lambda(0) = 1$ and $K_\lambda(x)$ is monotonically decreasing away from 0. For example, we adopt a Gaussian kernel $K_\lambda(x) = \exp\{-x^2/(2\lambda)\}$ in our examples.

The variance of $\hat{\pi}_{c_k}(y_i|\boldsymbol{y}_{-\kappa(i)}, c)$ is proportional to the reciprocal of the ESS and tends to be smaller for values of $c_k$ closer to $c$. Any combined estimator needs to take this effect into account[2]. In our estimator we can choose $c_1, \ldots, c_M$, $K_\lambda(x)$ and $\lambda$ to downweigh estimates $\hat{\pi}_{c_k}(y_i|\boldsymbol{y}_{-\kappa(i)}, c)$ which tend to have larger variances. The kernel weights should be proportional to the reciprocal of the variance of each estimate and this suggests making $K_\lambda(d_m)$ roughly proportional to ESS. In our examples, we have found that placing $c_1, \ldots, c_M$ to be equally spaced on the logarithmic scale and setting $\lambda$ to be the difference between $\log(c_{m+1})$ and $\log(c_m)$ is a good proxy.

In the special case that $c_1, \ldots, c_M$ are the 12 equally spaced points stated in Section 3, there are two main differences between the multiple importance sampler and the direct MCMC methodology. Firstly, the multiple importance sampler involves shorter MCMC runs with smaller burn-in. Secondly, the multiple importance sampler uses $M$ different MCMC chains to estimate $\pi(y_i|\boldsymbol{y}_{-\kappa(i)}, c_m)$ whereas the direct MCMC methodology only uses a single chain. In comparison with the $K-$fold standard importance sampler, the multiple 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 $c_0$. Finally, Table 3 summarizes the proposed importance samplers discussed in this section.

| Sampler | Density | Weight | Estimate of $\pi(y_i\|\boldsymbol{y}_{-\kappa(i)}, c)$ |
|---|---|---|---|
| $K$-fold Standard | $\pi(\boldsymbol{\theta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}\|\boldsymbol{y}_{-\kappa(i)}, c_0)$ | $w = \frac{\pi(\boldsymbol{\beta}_{\boldsymbol{\gamma}}\|\boldsymbol{\gamma},c)}{\pi(\boldsymbol{\beta}_{\boldsymbol{\gamma}}\|\boldsymbol{\gamma},c_0)}$ | $\frac{\sum_{j=1}^T w^{(j)}\pi(y_i\|\boldsymbol{\theta}_{\boldsymbol{\gamma}}^{(j)},\boldsymbol{\gamma}^{(j)})}{\sum_{j=1}^T w^{(j)}}$ |
| $K$-fold Auxiliary | $\pi(\boldsymbol{\theta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}\|\boldsymbol{y}_{-\kappa(i)})$ | $w = \frac{\pi(\boldsymbol{\beta}_{\boldsymbol{\gamma}}\|\boldsymbol{\gamma},c)}{\pi(\boldsymbol{\beta}_{\boldsymbol{\gamma}}\|\boldsymbol{\gamma})}$ | $\frac{\sum_{j=1}^T w^{(j)}\pi(y_i\|\boldsymbol{\theta}_{\boldsymbol{\gamma}}^{(j)},\boldsymbol{\gamma}^{(j)})}{\sum_{j=1}^T w^{(j)}}$ |
| $K$-fold Multiple | $\pi(\boldsymbol{\theta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}\|\boldsymbol{y}_{-\kappa(i)}, c_m)$, $m = 1, \ldots, M$ | $w_m = \frac{\pi(\boldsymbol{\beta}_{\boldsymbol{\gamma}}\|\boldsymbol{\gamma},c)}{\pi(\boldsymbol{\beta}_{\boldsymbol{\gamma}}\|\boldsymbol{\gamma},c_m)}$ | $\frac{\sum_{m=1}^M K_\lambda(d_m)\,\hat{\pi}_{c_m}(y_i\|\boldsymbol{y}_{-\kappa(i)},c)}{\sum_{m=1}^M K_\lambda(d_m)}$, $d_m = \log(c) - \log(c_m)$ $\hat{\pi}_{c_m}(y_i\|\boldsymbol{y}_{-\kappa(i)}, c) = \frac{\sum_{j=1}^T w_m^{(j)}\pi(y_i\|\boldsymbol{\theta}_{\boldsymbol{\gamma}}^{(j)},\boldsymbol{\gamma}^{(j)})}{\sum_{j=1}^T w_m^{(j)}}$ |

Table 3: The importance density, weight and estimate of $\pi(y_i|\boldsymbol{y}_{-\kappa(i)}, c)$ for each importance sampler.

[2]Owen and Zhou (2000) discuss how this happens in the Deterministic Mixture Sampler.

# 5 Results

## 5.1 Comparison of $K$-Fold Importance Samplers

The $K-$fold log predictive score $S(c)$ is estimated at the 12 equally spaced points stated in Section 3 using the $K$-fold standard importance sampler, the multiple importance sampler and the auxiliary importance sampler. In each case, the different samplers are run on each partition of the data. The direct MCMC output will be used as a "gold standard". Comparing the results to the MCMC runs in Section 3 allows us to measure the accuracy of the importance samplers. We will use the following measures: the mean squared error of the importance estimates of $S(c)$ evaluated at the 12 equally spaced points (MSE) 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 direct MCMC minimizer (SMin).

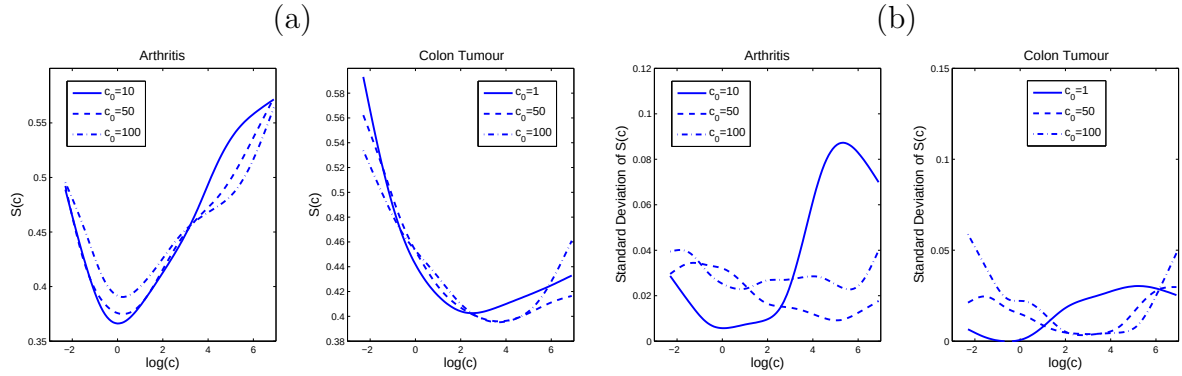


Figure 7: $K-$fold standard importance estimates of the Arthritis and Colon Tumour log predictive scores for selected values of $c_0$: (a) estimates averaged over 5 replications and (b) the standard deviation over 5 replications.

The $K-$fold standard importance sampler was implemented for $c_0 = 1,\ 10,\ 50,\ 100,\ 150$. We generated an MCMC sample of size $T = 80,000$ (after thinning to every fifth draw) with stationary distribution $\pi(\boldsymbol{\theta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}|\boldsymbol{y}_{-\kappa(i)}, c_0)$. 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 $\pi(\boldsymbol{\theta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}|\boldsymbol{y}_{-\kappa(i)}, c_0)$ is quite different from the target density $\pi(\boldsymbol{\theta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}|\boldsymbol{y}_{-\kappa(i)}, c)$ when $c$ is not close to $c_0$ and this may result in estimates of $S(c)$ with large variances. Figure 7 displays the sample mean and standard deviation of

the Arthritis log predictive 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) over the 5 replications. These values of $c_0$ were specifically selected so that the importance estimates are similar to those in Figure 4 and these results are the best that we can hope to get with the $K-$fold standard importance sampler.

| | Arthritis | | | Colon Tumour | | |
|---|---|---|---|---|---|---|
| $c_0$ | CPU | MSE | SMin | CPU | MSE | SMin |
| 1 | 427 | 0.03 | 5 | 195 | 0.006 | 1 |
| 10 | 428 | 0.03 | 5 | 193 | 0.014 | 0 |
| 50 | 426 | 0.01 | 4 | 194 | 0.007 | 4 |
| 100 | 426 | 0.02 | 5 | 193 | 0.013 | 5 |
| 150 | 434 | 0.03 | 2 | 194 | 0.011 | 4 |

Table 4: The average CPU time in minutes of the standard importance samplers $\pi(\boldsymbol{\theta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}|\boldsymbol{y}_{-\kappa(i)}, c_0)$ for some $c_0$ values, the mean squared error of the importance 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 direct MCMC.

Table 4 presents the average (over the 5 replications) CPU time in minutes, MSE and SMin of each $K-$fold standard importance sampler. Some $K-$fold standard importance samplers estimate the log predictive score and the minimizer with virtually the same accuracy as the direct MCMC methodology. However, the required CPU time is more than ten times smaller than the direct MCMC methodology. Unfortunately, the large differences in performance for the two data sets for the same $c_0$ suggests that finding a default value of $c_0$ for use with other data sets would be virtually impossible and motivates our development of the multiple and auxiliary importance samplers.

The multiple importance sampler was implemented with $M = 20$, $\lambda = 0.5$ and $c_m$ chosen equally spaced on the log scale from 0.1 to 1000. This implies that $\lambda$ is roughly the difference between $\log(c_{m+1})$ and $\log(c_m)$. Three multiple importance samplers have been used with different run lengths, described in Table 5. In each case the chain had been thinned to every fifth value. The sample mean of the estimated log predictive scores, over 5 replications, are shown in Figure 8(a) and give quite similar results to the MCMC log predictive scores depicted in Figure 4. The standard deviation for all three multiple importance samplers are smaller for smaller values of $c$ and larger for larger values of $c$ compared to the standard

| | | | Arthritis | | | Colon Tumour | | |
|---|---|---|---|---|---|---|---|---|
| Sampler | Burn-in | Sample | CPU | MSE | SMin | CPU | MSE | SMin |
| 1 | 50,000 | 30,000 | 2898 | 0.003 | 5 | 1289 | 0.009 | 4 |
| 2 | 20,000 | 16,000 | 1434 | 0.006 | 4 | 642 | 0.014 | 4 |
| 3 | 20,000 | 6000 | 712 | 0.013 | 5 | 320 | 0.04 | 3 |

Table 5: The specifications of three MCMC samplers involved in each multiple importance sampler with the average CPU time in minutes, MSE and SMin.

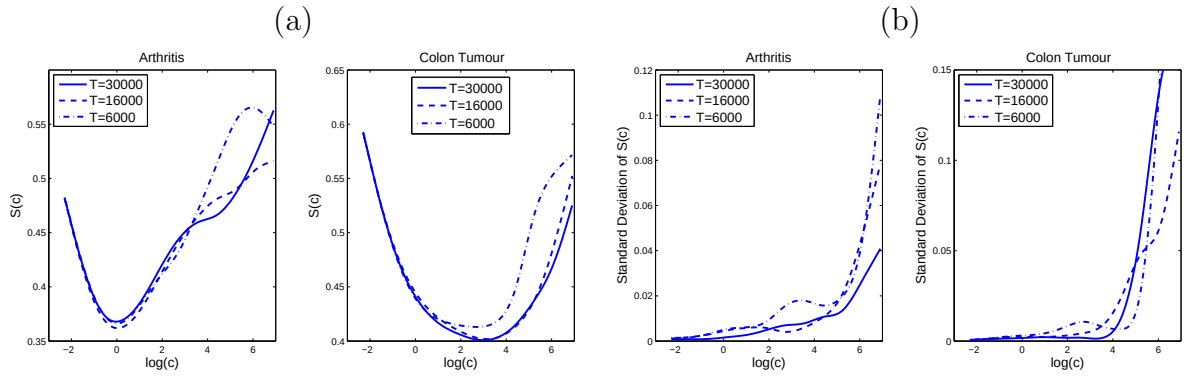


Figure 8: Importance estimates of the log predictive scores for the Arthritis and Colon Tumour using each multiple importance sampler: (a) estimates averaged over 5 replications and (b) the standard deviation over 5 replications. The sample size of the MCMC samplers involved in each multiple importance sampler is denoted by $T$.

deviations shown in Figure 7. This is due to the choice of $c_1, \ldots, c_M$ which are concentrated on those smaller values of $c$. This leads to more consistent results for SMin than using the $K-$fold standard importance sampler at the expense of longer run times (Table 5).

The previous estimates use a Gaussian kernel for $K_\lambda(x)$ and pre-specified window size $\lambda$. We also looked at the Gaussian, Epanechnikov and Tri-Cube kernels for $\lambda$ equal to 0.2, 0.4, 0.6, 0.8 and 1. The results showed that the mean squared errors decreased as $\lambda$ increased within each kernel and was smaller with the Gaussian kernel than with the other two kernels for all $\lambda$. However, the differences were small suggesting the results are fairly robust to the choice of $K_\lambda(x)$.

We conclude that the multiple importance samplers estimate the log predictive score with similar accuracy to 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 first multiple importance sampler estimates the log predictive score with similar accuracy to the basic $K-$fold standard importance sampler with an "optimal" value of $c_0$ but avoids finding this optimal value (leading to a more "automatic" procedure). However, the method comes with a higher computational cost.

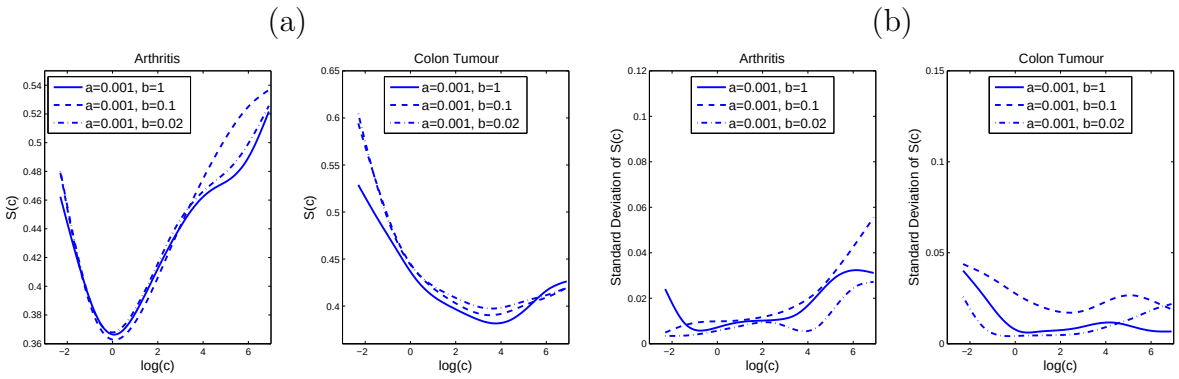


Figure 9: Auxiliary importance estimates of the Arthritis and Colon Tumour log predictive scores for different Inverse Gamma auxiliary distributions on $c_0$: (a) estimates averaged over 5 replications and (b) the standard deviation over 5 replications.

The auxiliary importance sampler offers an alternative method to combine different values of $c$ in the importance sampling distribution. An MCMC sample of size $T = 80,000$ (after thinning to every fifth draw) with stationary distribution $\pi_A(\boldsymbol{\theta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}|\boldsymbol{y}_{-\kappa(i)})$ was generated. 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 chosen to concentrate the mass on the range of $c$ over which the log predictive score is estimated. 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. 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.

The average log predictive scores over 5 replications, shown in Figure 9(a) for three typical Inverse Gamma auxiliary distributions on $c$, are quite similar to the direct MCMC results depicted in Figure 4. The Mean Squared Errors (shown in

Table 6) are small and the methods provide very similar minimizers of the log predictive score. Figure 9(b) indicates that replicates are closer than the other importance samplers for all cases and values of $c$. The CPU time of these samplers

| | Arthritis | | | Colon Tumour | | |
|---|---|---|---|---|---|---|
| IG$(a, b)$ | CPU | MSE | SMin | CPU | MSE | SMin |
| $a = 0.001, b = 1$ | 475 | 0.01 | 5 | 215 | 0.006 | 4 |
| $a = 0.001, b = 0.1$ | 481 | 0.008 | 5 | 217 | 0.012 | 4 |
| $a = 0.001, b = 0.02$ | 483 | 0.004 | 5 | 217 | 0.007 | 4 |

Table 6: The average CPU time in minutes, MSE and SMin for different Inverse Gamma auxiliary importance samplers.

is about ten times smaller than with the MCMC methodology and considerably less than with the multiple importance sampler, indicating a substantial computational gain.

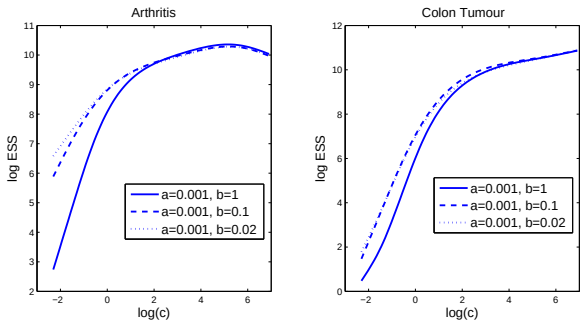


Figure 10: The average log mean ESS of the Inverse Gamma auxiliary importance samplers as a function of $c$ for the Arthritis and Colon Tumour data.

Figure 10 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. Mean ESS is a (mostly) increasing function of $c$, quite in contrast to the standard $K$-fold importance sampler. 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 for values of $c$ around the optimal value.

## 5.2 Comparison to Fully Bayesian Inference

An alternative to the empirical Bayes approach taken in this paper is fully Bayesian inference where a prior is placed 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), Liang et al. (2008) and Cui and George (2008). This approach is a natural way to account for uncertainty in the estimation of $c$ and is often believed to increase robustness to the specification of $c$. However, results can be sensitive to the choice of prior on $c$ which can be particular acute when there is little information in the data and the prior is chosen to be diffuse (to represent a lack of prior knowledge).

The approaches were compared using out-of-sample[3] prediction accuracy on the Arthritis dataset and a larger dataset regarding prostate cancer. This data comprises $n = 136$ observations, divided into prostate tumour and nontumour groups, with $p = 10150$ gene expression measurements (Singh et al., 2002). The datasets were partitioned into $K = 12$ subsets, where 10 subsets formed the training set and two subsets formed the test set. The empirical Bayes method was applied using 10-fold cross-validation on the training set with the auxiliary importance sampler. The minimizers of the log predictive score were $c = 1.23$ for the Arthritis dataset and $c = 15.2$ for the Prostate data. Fitting the Bayesian probit regression model with these fixed values of $c$ leads to log predictive scores for the test dataset of 0.21 for the Arthritis data and 0.08 for the Prostate data.

| IG$(a, b)$ | Arthritis | Prostate |
|---|---|---|
| $a = 0.001, b = 1$ | 0.30 | 0.12 |
| $a = 0.001, b = 0.1$ | 0.28 | 0.10 |
| $a = 0.001, b = 0.02$ | 0.29 | 0.13 |

Table 7: MCMC estimates of the out-of-sample log predictive scores for the Arthritis and Prostate datasets and three representative Inverse Gamma priors on $c$.

We implemented a fully Bayesian approach with an Inverse Gamma prior distribution for $c$ which is the standard, conditionally conjugate prior and is made diffuse

[3]We compare out-of-sample predictions in this case, since the empirical Bayes approach specifically selected $c$ to optimize cross-validation prediction. Using the latter criterion to compare empirical Bayes and fully Bayes procedures leads to similar conclusions.

by choosing a small shape parameter (in fact, the same values for the hyperparameters are chosen as for the auxiliary distribution in the previous subsection). The methods were compared using out-of-sample log predictive scores. MCMC samples of size $T = 190,000$ after thinning to every tenth draw were generated for the training dataset using the algorithm of Holmes and Held (2006). Table 7 shows that the empirical Bayes approach had much smaller out-of-sample log predictive scores than the fully Bayesian method indicating better predictions.

An alternative empirical Bayes approach is described by Strimenopoulou and Brown (2008) who minimize minus log-likelihood error (equation (4) of their paper) on the training data. They suggest estimating the regression coefficients using a maximum *a posteriori* approach. The log predictive scores for the test data were 0.19 for the Arthritis dataset and 0.09 for the Prostate dataset. These values are quite similar to those found by the empirical Bayes approach proposed in the article.

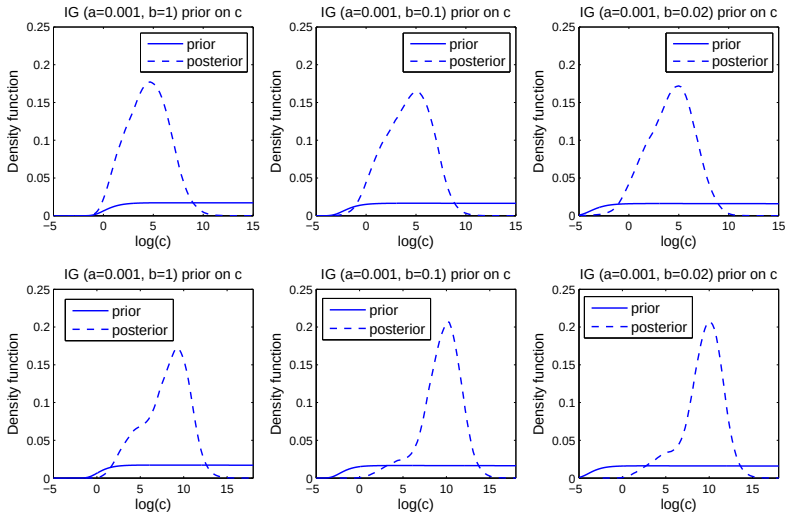


Figure 11: Prior and posterior density functions of $\log c$ for the Arthritis data set (top row) and Prostate dataset (bottom row).

Figure 11 shows the posterior distribution of $\log(c)$ for both data sets under the three priors. These posteriors place some mass at the value of $c$ that minimizes $S(c)$ but this tends to be in the left-hand tail of the distribution and far from the mode. Furthermore, they place substantial mass on large values of $\log(c)$, which are associated with poor prediction performance due to overfitting. This explains the poor performance of fully Bayesian inference in these examples. The choice of prior

distribution for $c$ which encourages good prediction is an area that we are currently investigating. This choice becomes increasingly important when $p$ is much larger than $n$ and the prior has a substantial effect on the inference.

# 6 Conclusions

The "ridge" hyperparameter $c$ crucially affects Bayesian variable selection in probit regression with $p \gg n$. 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. Interestingly, the guideline range for choosing $c$ proposed in Sha et al. (2004) covers our preferred value in one of the datasets we examine here, but remains very far from this optimal value in the other[4].

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 $\pi(y_i|\boldsymbol{\theta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma})$. Thus, we propose to use $K-$fold importance samplers with importance densities $\pi(\boldsymbol{\theta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}|\boldsymbol{y}_{-\kappa(i)}, c_0)$ and $\pi(\boldsymbol{\theta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}|\boldsymbol{y}_{-\kappa(i)})$ to estimate $\pi(y_i|\boldsymbol{y}_{-\kappa(i)}, c)$ for different values of $c$.

[4]For the prostate data mentioned in Subsection 5.2, the Sha et al. (2004) guidelines lead to the range (0.004, 0.1) for $c$, again not covering the minimizer of the log predictive score, which is in the interval (1.6, 20).

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 multiple and auxiliary importance samplers, which avoid choosing a particular value for $c_0$.

The $K-$fold multiple importance sampler involves shorter run MCMC chains and mixes over $c_0$ values, resulting in a six-fold improvement in CPU time 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 multiple 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 tailed density functions and there is no need for further user input. The multiple 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 the much larger prostate cancer dataset, where $n = 136$ and $p = 10150$. Here the demand in CPU time 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 $l$. The proposed methods could be extended to choosing the hyperparameter vector $(q, c)$. In the case of the auxiliary importance sampler, a Beta$(a, b)$ distribution could be a reasonable auxiliary distribution on $q$.

Fully Bayesian inference where a prior is placed on $c$ could be used but default, diffuse priors tend to produce poor predictions which suggests that an empirical Bayes approach, such as ours, will be useful in the context of regression with many

more regressors than observations.

## Acknowledgements

We gratefully acknowledge insightful comments by two anonymous referees.

# References

Alon, U., N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine (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.

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

Celeux, G., J.-M. Marin, and C. P. 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. T., C. C. Holmes, B. K. Mallick, and A. F. M. Smith (2002). *Bayesian Methods for Nonlinear Classification and Regression*. John Wiley and Sons.

Dobra, A. (2009). Variable selection and dependency networks for genomewide data. *Biostatistics 10*, 621–639.

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

Geisser, S. and W. F. 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.

George, E. I. and D. P. Foster (2000). Calibration and empirical Bayes variable selection. *Biometrika 87*(4), 731–747.

Geyer, C. J. (1994). Estimating normalizing constants and reweighting mixtures in MCMC. Technical Report 568, University of Minnesota, School of Statistics.

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

Hastie, T., R. Tibshirani, and J. H. Friedman (2001). *The Elements of Statistical Learning*. Springer series in statistics, New York.

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

Key, J., L. Pericchi, and A. F. M. Smith (1999). Bayesian model choice: what and why? In J. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith (Eds.), *Bayesian Statistics Volume 6*, pp. 343–370. Oxford: Oxford University Press.

Lee, K. E., N. Sha, E. R. Dougherty, M. Vannucci, and B. Mallick (2003). Gene selection: A Bayesian variable selection approach. *Bioinformatics 19*, 90–97.

Liang, F., R. Paulo, G. Molina, M. A. Clyde, and J. O. 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.

Owen, A. and Y. Zhou (2000). Safe and effective importance sampling. *Journal of the American Statistical Association 95*, 135–143.

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

Scott, J. G. and J. O. Berger (2006). An exploration of aspects of Bayesian multiple testing. *Journal of Statistical Planning and Inference 136*, 2144–2162.

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

Sha, N., M. Vannucci, 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.

Shafer, G. (1982). Lindley's paradox. *Journal of the American Statistical Association 77*, 325–351.

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

Strimenopoulou, F. and P. J. Brown (2008). Empirical Bayes logistic regression. *Statistical Applications in Genetics and Molecular Biology 7*, Article 9.

Veach, E. and L. Guibas (1995). Optimally combining sampling techniques for Monte Carlo rendering. In *SIGGRAPH '95 Conference Proceedings*, pp. 419–428. Reading, MA: Addision-Wesley.

Ventura, V. (2002). Non-parametric bootstrap recycling. *Statistics and Computing 12*, 261–273.

Zhou, X., K.-Y. Liu, and S. T. C. Wong (2004). Cancer classification and prediction using logistic regression with Bayesian gene selection. *Journal of Biomedical Informatics 37*(4), 249–259.