Transdimensional Sampling Algorithms for Bayesian Variable Selection in Classification Problems with Many More Variables Than Observations

Demetris Lamnisos, Jim E. Griffin and Mark F. J. Steel *

Abstract

Model search in probit regression is often conducted by simultaneously exploring the model and parameter space, using a reversible jump MCMC sampler. Standard samplers often have low model acceptance probabilities when there are many more regressors than observations. Implementing recent suggestions in the literature leads to much higher acceptance rates. However, high acceptance rates are often associated with poor mixing of chains. Thus, we design a more general model proposal that allows us to propose models "further" from our current model. This proposal can be tuned to achieve a suitable acceptance rate for good mixing. The effectiveness of this proposal is linked to the form of the marginalisation scheme when updating the model and we propose a new efficient implementation of the automatic generic transdimensional algorithm of Green (2003). We also implement other previously proposed samplers and compare the efficiency of all methods on some gene expression data sets. Finally, the results of these applications lead us to propose guidelines for choosing between samplers.

Key Words: Data augmentation, Gene expression data, Probit model, Reversible jump sampler, Transdimensional Markov chain.

Supplementary Materials

1. Data Sets

`arthritis.mat` The Arthritis data set consists of rheumatoid arthritis and osteoarthritis groups. The vector TARGET takes values 0 or 1 and indicates class membership. The matrix X is the centred design matrix containing the gene expression levels.

`colon_tumor.mat` The Colon Tumour data set contains tumour and normal colon groups. The vector TARGET and the matrix X are as for the `arthritis.mat`.

*Demetris Lamnisos is PhD Student, Department of Statistics, University of Warwick, Coventry, U.K., CV4 7AL (Email: D.S.Lamnisos@warwick.ac.uk). Jim E. Griffin is Lecturer, Institute of Mathematics, Statistics and Actuarial Science, University of Kent, Canterbury, U.K., CT2 7NF (Email: J.E.Griffin-28@kent.ac.uk) and Mark F.J. Steel is Professor, Department of Statistics, University of Warwick, Coventry, U.K. (Email: M.F.Steel@stats.warwick.ac.uk).

`leukemia.mat` The Leukemia data set consists of samples from patients with either acute lymphoblastic leukemia (ALL) or acute myeloid leukemia (AML). Golub *et al* (1999) noted that the global expression profiles also reflect two ALL subtypes (B-cell) and (T-cell). Hence, this data set can be divided into either two or three classes. The vector TARGET takes values 0, 1 or 2 and indicates membership to AML, ALL-T cell and ALL-B cell respectively. The matrix X is as for the `arthritis.mat`.

`prostate.mat` The Prostate data set consists of prostate tumour and nontumour groups. The vector TARGET and the matrix X are as for the `arthritis.mat`.

2. Computer Code

`Holmes_Held.m` This MATLAB's file implements the Holmes and Held algorithm. The user is responsible for setting the response variable TARGET and the design matrix X of the dataset, the prior on the intercept α PRIOR_INTRCP which has two options PROPER and IMPROPER. In the case of PROPER the user needs to set the prior variance h of the univariate normal prior $N(0, h)$. The other user's input are the prior variance of the regression coefficients C, the model proposals parameters N and P and the prior mean of the model size W. The last input variables are the number of MCMC iterations NUM_ITER, the burn-in period NUM_BURN and the thinning of the chain NUM_THIN. It is optional for the user to set the number of genes (the number of columns of the design matrix X). These genes are pre-selected using the ratio of between-groups to within-groups sum of squares of Dutoid *et al* (2002). The output is the posterior gene inclusion probabilities PROB_INCLUSION, their effective sample sizes ESS, the posterior sample of regression coefficients BETA, the posterior sample of inclusion variables L and the between model acceptance rate ACCEPTANCE.

`AG_LA.m` This MATLAB's file implements the Automatic generic sampler with Laplace approximation. The directions to use this program are as those of the `Holmes_Held.m`.

`AG_IWLS.m` This MATLAB's file implements the Automatic generic sampler with Iterated Weighted Least Squares approximation. The directions to use this program are as those of the `Holmes_Held.m`.

`Zeroth_Order.m` This MATLAB's file implements the Zeroth Order method. The directions to use this program are as those of the `Holmes_Held.m`.

`First_Order.m` This MATLAB's file implements the First Order method. The directions to use this program are as those of the `Holmes_Held.m`.

`Second_Order.m` This MATLAB's file implements the Second Order method. The directions to use this program are as those of the `Holmes_Held.m`.

`Cond_Max.m` This MATLAB's file implements the Conditional Maximization method. The directions to use this program are as those of the `Holmes_Held.m`.

`Run_code.m` This MATLAB's file contains examples and directions on how to run the above programs with input variables those described in the paper. It also contains two examples of processing the output.

3. Starting value of the Bayesian IWLS method

The initial value $\boldsymbol{\mu}_\gamma^{(0)}$ of the Bayesian IWLS method is defined by $\boldsymbol{\mu}_\gamma^{(0)} = (\tilde{\boldsymbol{X}}_\gamma' \tilde{\boldsymbol{X}}_\gamma)^{-1} \tilde{\boldsymbol{X}}_\gamma' \boldsymbol{z}$ which is the least square estimates of the following linear model

$$\begin{aligned} \boldsymbol{z} &= \tilde{\boldsymbol{X}}_\gamma \boldsymbol{\theta}_\gamma + \varepsilon \\ \varepsilon &\sim \mathrm{N}_n(\mathbf{0}, \mathrm{I}_n). \end{aligned}$$

The response vector $\boldsymbol{z}$ is the mean of the auxiliary variable introduced by Albert and Chip (1993) and is roughly estimated by running few iterations of the Holmes and Held algorithm (Section 3.1.1).