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

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

March 30, 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 parameter 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.

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

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. The cell array VALID_SET contains the cross-validation sets.

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

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

*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

2. Computer Code

`Standard.m` This MATLAB's file implements the standard importance sampler. The user is responsible for setting the response variable TARGET and the design matrix X of the dataset, the cross-validation set VALID_SET and the default value C_0. The other user's input are the lower and upper bounds of c CONSTR_C, the number of different values of c and 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 last input parameters are the prior mean of the model size W, the model proposals parameters N and P, 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 log prediction of the cross-validation set LOG_PREDICTIVE and the effective sample size ESS of the importance sampler at the vector of different values of c VAR_COEF.

`Mixture.m` This MATLAB's file implements the mixture importance sampler. The user is responsible for setting the response variable TARGET and the design matrix X of the dataset, the cross-validation set VALID_SET, the lower and upper bounds of c CONSTR_C and the number of different values of both c and c_0 K. The other user's input PRIOR_INTRCP, h, W, N, P, NUM_ITER, NUM_BURN, NUM_THIN and the optional argument are as for the `Standard.m`. The output is the log prediction of the cross-validation set LOG_PREDICTIVE at the vector of different values of c VAR_COEF. The last output is the vector of distinct values of c_0 DINST_C0.

`Auxiliary.m` This MATLAB's file implements the auxiliary importance sampler. The user is responsible for setting the response variable TARGET and the design matrix X of the dataset, the cross-validation set VALID_SET and the auxiliary distribution on c AUX_VARIABLE which has two options 'IGAMMA' and 'GAMMA-IGAMMA'. The parameters of the auxiliary distribution are specified in the vector DELTA. The other user's input CONSTR_C, K, PRIOR_INTRCP, h, W, N, P, NUM_ITER, NUM_BURN, NUM_THIN and the optional argument are as for the `Standard.m`. The output is the log prediction of the cross-validation set LOG_PREDICTIVE and the effective sample size ESS of the importance sampler at the vector of different values of c VAR_COEF.

`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 examples of processing the output.