Matlab codes BVSGS_I, BVSGS_G and BVSGS_GI

Written by V.A. Delouille

September 1998

As part of

Bayesian Variable Selection with Related Predictors using the Gibbs Sampler

Thesis submitted to the University of Kent at Canterbury in the subject of Statistics for a degree of Master of Science (M.Sc.)

1 Introduction

Programs contained in the directories BVSGS_I, BVSGS_G and BVSGS_GI all perform Bayesian variable selection using Gibbs sampler and a selection prior (that is, if X_j is said to be excluded from a model, the corresponding regression coefficient is put equal to zero with probability 1). The particular setting is the multivariate Gaussian regression model (see Brown *et al.* [1]). The reason for having divided the programs into 3 parts lies in the fact that each part is related to a specific type of regression model. The particular use of each directory is reviewed below.

2 BVSGS_I directory

The programs in this directory perform Bayesian variable selection on models which may contain, in addition to some main effects, some two-way interaction and quadratic terms. Suppose we have at our disposal a data set with n observations on pm main effects, $X_1, \ldots, X_{pm}$, but the response Y is thought to be also affected by some second order terms X_iX_j. Then the first step consists in processing the data as follows.

1. Center the main effects $X_1, \ldots, X_{pm}$, and the response variables Y.

2. Generate the quadratic and interaction terms in the order

$$X_1^2, X_1X_2, X_1X_3, \ldots, X_1X_{pm},$$
$$X_2^2, X_2X_3, \ldots, X_2X_{pm}, \ldots, X_{pm-1}X_{pm}, X_{pm}^2,$$

   and center them. The program *processing.m* might be helpful in this.

The centered variables $X_1, \ldots, X_{pm}, X_1^2, \ldots, X_{pm}^2$ then consitute the model matrix X. The next step is to specify values of hyperparameters in prior distributions. Specifically, a latent vector γ serves to identify each submodel via $(\gamma_j = 1) \Leftrightarrow (X_j$ is included), (see George and McCullogh [3] for a review of Bayesian variable selection); and the programs in this directory enable the user to include in the prior distribution of γ some implicit assumptions about the form of the selected models. More precisely, The user has the choice between two types of assumptions.

1. If it is believed that an interaction $X_i X_j$ should be included in a model only when both parents are also selected, then structuring the prior of γ through the *strong heredity principle*(a naming convention adopted by Chipman [2]) will ensure that the selected model is of the desired form.

2. If it is rather believed that an interaction should be forbidden to enter a model only when both parents are absent from this model, then one should use the *weak heredity principle* to structure the prior distribution on γ.

Each type of prior requires hyperparameters to be specified. This is explained in the *readme_bvsgsi* file of the BVSGS_I directory. For a study on how to account for relationships between predictors in a Bayesian variable selection process, the reader is referred to Chipman [2].

3 BVSGS_G directory

The programs in this directory perform Bayesian variable selection for models containing independent effects and some groups of variables. By 'group of variables', it is meant a set of variables such that one wants them to be either all included, or all excluded from the selected models. Suppose for example that the model contains a set of, say, l dummy variables that code the $l+1$ levels of a categorical variable. If the interest lies in testing the effect of this categorical variable as a whole, one should treat the l dummy variables as a 'group'. Only one component of the vector γ will serve to identify if a group of variables is active. In the model matrix X, the independent effects should be provided first, followed by the different groups of variables. This is explained in more detail in the *readme_bvsgsg* file of this directory.

4 BVSGS_GI directory

The programs in BVSGS_GI should be used for models containing both interactions and groups of variables. More precisely, the list of all explanatory variables contains the following terms.

- Independent main effects $X_1, \ldots, X_{pm}$.

- Quadratic and two-way interactions of these main effects.
- Groups of variable $D_1, \ldots, D_r$ (in the group D_i there are, say, l_i variables, $i = 1, \ldots, r$).
- Two-way interaction between the independent main effects and the grouped variables. These interactions are also considered as 'groups' of variables, in the sense given in section 3. That is, for example, the terms $X_1 D_{1_1}, \ldots, X_1 D_{1_{l_1}}$ will either be all included or all excluded from a selected model.

Hence, the list of predictors that can possibly be elected consists of

$$X_1, \ldots, X_{pm}, X_1^2, X_1X_2, \ldots, X_{pm-1}X_{pm}, X_{pm},$$
$$D_1, \ldots, D_r, D_1X_1, \ldots, D_1X_{pm}, \ldots, D_rX_1, \ldots, D_rX_{pm}.$$

The variables in the matrix X should be arranged in columns as in the following sequence.

1. Main effects and second-order terms of independent variables, in the order described in section 2.
2. Main effects of grouped variables.
3. Interactions between grouped variables and independent terms in the order

$$X_1D_{1_1}, \ldots, X_1D_{1_{l_1}}, \ldots, X_{pm}D_{1_1}, \ldots, X_{pm}D_{1_{l_1}}, X_1D_{2_1}, \ldots, X_1D_{2_{l_2}}, \ldots$$

The programs *processing.m* and *process_gi.m* generate the interactions terms for, respectively, independent and grouped variables. As in section 2, implicit assumptions on the form of a model will be translated into a particular specification of the prior distribution on γ. Again, the user has the choice between the strong heredity and the weak heredity principle to stipulate which type of model is allowed to be selected.

References

[1] P. J. Brown, M. Vannucci, and T Fearn. Multivariate bayesian variable selection and prediction. *Journal Royal Statistical Society. Series B*, 60:627–641, 1998.

[2] H. Chipman. Bayesian variable selection with related predictors. *Canadian Journal of Statistics*, 24:17–36, 1996.

[3] E.I George and R.E McCulloch. Approaches for bayesian variable selection. *Statistical Sinica*, 7:339–373, 1997.