# Matlab Code for Bayesian Variable Selection

Marina Vannucci
Texas A&M University, USA

*The ISBA Bulletin, September 2000, Vol.7 n.3*

This software provides a set of Matlab functions that perform Bayesian variable selection in a multivariate regression setting. There are different sets of functions currently available, implementing different approaches and models for the variable selection problem: *bvgs.tar, bvsme.tar, bvssa.tar* (written by Marina Vannucci) and *bvsgs_i.tar, bvsgs_g.tar, bvsgs_gi.tar* (written by Veronique Delouille and Marina Vannucci).

Consider the multivariate regression model with $p$ regressors, $q$ responses and $n$ observations, where $p$ can largely exceed $n$. Bayesian variable selection approaches use a latent vector with $p$ binary entries to identify the different submodels. The marginal posterior distribution of the binary vector is derived and, in high dimensions, Markov chain Monte Carlo algorithms are used to sample from this posterior distribution. Also, prediction can be done by computing a weighted average of the predictive distribution for the different models, or at least for a restricted set of them in the case of high dimensions, i.e. of a large $p$. The weights of the average are determined as the posterior probabilities of the single models.

Functions in *bvsgs.tar* perform Bayesian variable selection as described in Brown, Vannucci, Fearn (1998a), using conjugate priors on the parameters of the regression model and independent Bernoulli priors on the $p$ entries of the selecting binary vector. The posterior distribution is searched via a Gibbs sampler that moves from a model to another by generating componentwise from the full conditional distributions. The input of the main function *bvsgs_sp.m* requires the data and all the hyperparameter specifications, it asks for possible permutations of the data and for the Gibbs parameters (initial number of variables included, number of iterations). It returns as

output the list of all visited models and their relative posterior probabilities, the normalized ordered relative probabilities of distinct visited models and the marginal probabilities of inclusion of the single variables. Distinct visited models and their corresponding posterior probabilities can then serve as input to the function *pbvs_sp.m* that performs Bayesian model averaging prediction on a set of future data using a specified number of the most probable visited models.

Functions in *bvsme.tar* still use conjugate priors on the parameters of the regression model but with a g-prior for the regression coefficients and a more flexible Beta-Binomial prior on the selecting binary vector. Also, the posterior distribution is searched via a Metropolis algorithm that moves through a sequence of models generating a new candidate by randomly modifying the current model via deletion, addition and swapping moves. This set of functions was used to produce the results of Brown, Vannucci, Fearn (1998b). The input of the main function *bvsme_gp.m* requires the data and the hyperparameter specifications, it asks for possible permuting of the data and for the Metropolis parameters (initial number of variables included, probabilities of deletion/addition and swap moves, total number of iterations) and returns a similar output to that of the function *bvsgs_sp.m*.

Both models implemented in *bvsgs.tar* and *bvsme.tar* require the user to pre-process the data by centering the training data and subtracting the training means from the future data.

Functions in *bvsgs_i.tar*, *bvsgs_g.tar* and *bvsgs_gi.tar* implement Bayesian variable selection using Gibbs sampler in a multivariate regression with related predictors and with a prior on the latent binary vector suitably modified to allow different combinations of predictor terms. Specifically, *bvsgs_i.tar* allows for main effects and two-way interaction and quadratic terms, *bvsgs_g.tar* for main effects and "grouped" variables (sets of variables to be included or excluded as a group), and *bvsgs_gi.tar* for main effects and both interactions and grouped variables. The code automatically generates interaction terms for independent and grouped variables and centers responses and explanatory variables. The prior on the latent binary vector allows for two different forms of inclusion of interaction terms, one if both variables of the interaction are also included, the other if at least one of the two variables is included.

Functions in *bvssa.tar* implement an alternative approach to the variable selection problem that uses Bayesian decision theory as described in Brown, Fearn, Vannucci (1999), attaching costs to the inclusion of the single vari-

ables. A non-conjugate proper prior distribution is used for the regression parameters and a measure of predictive performance is computed for the different models. A simulated annealing algorithm is used to optimise the expected utility in the case of many regressors. The input of the main function *bvs_sa.m* requires the data and parameter specifications, it asks for the annealing parameters (initial temperature, updating temperature parameter, probabilities of moves and stopping parameter) and returns as output the list of visited models and their prediction costs. The model with minimum cost can then be used to do least squares and Bayes predictions via *pbvs_sa.m*.

All sets of functions here described have been used in calibration problems for the prediction of the chemical composition of a sample from its near-infrared spectrum comprising as many as 700 frequencies. Training samples available usually had few observations. In applications, the number of explanatory variables has been as high as 350, though *bvsme.tar*, with the g-prior and a Metropolis search, could have handled the total up to 700. Currently under study is a fast updating algorithm that will lead to revised code (to be released) capable of handling much larger datasets. Code and related documentation can be freely downloaded from

HTTP://STAT.TAMU.EDU/ MVANNUCCI/WEBPAGES/CODES.HTML

and should be distributed for non-commercial purposes only. All code require Matlab 5 by MathWorks.

## References:


BROWN, P. J., FEARN, T. and VANNUCCI, M. (1999). The choice of variables in multivariate regression: a Bayesian non-conjugate decision theory approach. *Biometrika*, **86(3)**, 635-648.

BROWN, P.J., VANNUCCI, M. and FEARN, T. (1998a). Multivariate Bayesian variable selection and prediction. *Journal of the Royal Statistical Society,* B, **60(3)**, 627-641.

BROWN, P.J., VANNUCCI, M. and FEARN, T. (1998b). Bayesian wavelength selection in multicomponent analysis. *Journal of Chemometrics*, **12**, 173-182.