# BAYESIAN VARIABLE SELECTION FOR PROBIT MIXED MODELS

Meïli Baragatti[*,a,b]

[a] *Institut de Mathématiques de Luminy (IML), CNRS Marseille, case 907, Campus de Luminy, 13288 Marseille Cedex 9, France*
[b] *Ipsogen SA, Luminy Biotech Entreprises, Case 923, Campus de Luminy, 13288 Marseille Cedex 9, France*

**Abstract**

In computational biology, gene expression datasets are characterized by very few individual samples compared to a large number of measurments per sample. Thus, it is appealing to merge these datasets in order to increase the number of observations and diversify the data, allowing a more reliable selection of genes relevant to the biological problem. This necessitates the introduction of the dataset as a random effect. Extending previous work of Lee et al. (2003), a method is proposed to select relevant variables among tens of thousands in a probit mixed regression model, considered as part of a larger hierarchical Bayesian model. Latent variables are used to identify subsets of selected variables and the collapsing technique of Liu (1994) is combined with a Metropolis-within-Gibbs algorithm (Robert and Casella, 2004). The method is applied to a merged dataset made of three individual gene expression datasets, in which tens of thousands of measurements are available for each of several hundred human breast cancer samples. Even for this large dataset comprised of around 20000 predictors, the method is shown to be efficient and feasible. As a demonstration, it is used to select the most important genes that characterize the estrogen receptor status of the cancer patients.

*Key words:* Bayesian selection, selection of covariables, random effects, probit mixed regression model, collapsing technique, Metropolis-within-Gibbs algorithm

Selection of variables is a common problem in many scientific fields, and particularly in bioinformatics. Gene expression profiling analyses are notorious for generating a very large number of predictors compared to the number of observations. Microarray technology is important for finding genes that are implicated in biological processes including development, disease, and response to

*Corresponding author
*Email addresses:* baragatt@iml.univ-mrs.fr (Meïli Baragatti ), baragattimeili@hotmail.com (Meïli Baragatti )

treatment, and it plays an important part in the current tendancy towards personalized medicine. Identified genes can be used to classify future observations, influencing the treatment of patients. However, these experiments are expensive, and datasets have often less than 100 specimens. The goal, therefore, is to advance a method allowing variable selection from merged microarray datasets, each of which presenting it's own individual experimental bias.

Several approaches have been proposed to identify differentially expressed genes in different classes. One of the simplest is the multiple testing method of Dudoit, with corrected levels of p-values. It was limited by the fact that it did not consider the interactions between genes, and that it selected the most differentially expressed genes as opposed to the most relevant classifiers. As a consequence, model-based approaches have been developed to select variables. A well-known example is SVM (Support Vector Machine) with a recursive feature elimination of the genes (Guyon et al. (2002)). George and McCulloch (1993) and Chipman et al. (2001) developed Bayesian variable selection with the use of Gibbs sampling for linear models; a review of this type of selection is provided by O'Hara and Sillanpää. (2009). Tadesse et al. (2005) proposed a Bayesian variable selection in a model-based clustering approach, using a multivariate Gaussian mixture model. Binary responses are often encountered in biostatistics studies, therefore probit or logistic models are implied. Bayesian variable selection methods have been proposed by Lee et al. (2003), Sha et al. (2004), Zhou et al. (2004b), and Zhou et al. (2004c) for probit regression, and by Zhou et al. (2004a), Chen and Dey (2003) and Tüchler (2008) for logistic regression. Extension to multi-category data has been done for the probit model; see Albert and Chib (1993).

The motivation behind the variable selection method developed is to take the design of the study into account by using random effects in a mixed model. It is particularly suited to a merged microarray dataset design, and many such datasets are freely available from the NCBI GEO website (Edgar et al., 2002). The increased size of a merged dataset facilitates it's resplitting into training and validation sets. In addition, a merged set comprises more data diversity than an individual set, masking bias due to any one particular dataset. Among all the methods previously proposed for data selection, only that of Tüchler (2008) considered mixed models. However, her approach was specific for logistic models, and the method was applied to datasets with only few dozens of predictors, whereas the aim of the developed method is to select a few predictors among tens of thousands. While the coefficients of logistic regression are more easily interpreted than those obtained through probit regression, we are more concerned with developing an efficient tool for variable selection and classification. Furthermore, probit models are more computationally advantageous: Gaussian latent variables can be introduced that render the conditional distribution of the model's parameters equivalent to those under classical Bayesian linear regression models. These models are then easy to use in a Bayesian framework and in Gibbs-samplers if conjugate priors are well chosen.

The method developed extends the approach of George and McCulloch (1993) and Lee et al. (2003). A probit mixed regression model is considered as part of a larger hierarchical Bayesian model, and latent variables are used to identify subsets of selected variables by a combination of the collapsing technique of Liu (1994) and the Metropolis-within-Gibbs algorithm (Robert and Casella, 2004). The promising subsets are those with higher posterior probability; that is, they appear more frequently in the Gibbs-sampler. The selected variables can then be used to fit a final probit mixed model and to classify future observations.

To apply the method, Affymetrix microarray data is used, so predictors (genes) will be referred to as "probesets", according to that technology. An Affymetrix U133plus2 microarray profiles all of the genes in the human genome, many of them more than once, using over 54000 gene-specific "probesets". Our Bayesian variable selection method for probit mixed models is developed to select a few important probesets, among tens of thousands, which are indicative of the activity of the estrogen receptor gene in breast cancer. The severity of this common and deadly disease is directly related to estrogen receptor (ER) status, which is traditionally measured biochemically.
Three different breast cancer datasets are used, all with clinically defined ER status: one private dataset from the Institut Paoli Calmettes (Marseille, France), consisting of 151 samples, and two datasets freely available from the NCBI GEO public website (Edgar et al., 2002): accession numbers GSE2109 (310 samples) and GSE5460 (124 samples). One microarray experiment is done per patient, and ten of thousands of probesets are measured per experiment. The dataset will be introduced as a random effect in the model, thus accounting for the different experimental conditions implicit in each set. The three merged datasets are split into training and validation sets, and the relevance of the selected probesets is checked by fitting a probit mixed model on the training set and predicting the ER status for the patients from the validation set and other independant sets avalaible from the NCBI GEO website. The stability and the sensitivity of the algorithm are also checked by using the stability index of Kuncheva (2007) and the relative weighted consistency measure of Somol and Novovicova (2008).

The talk will be organized as follows. The probit mixed model with latent variables will be described and the full conditional distributions necessary for the Gibbs sampling algorithm will be given. Then the algorithm will be outlined and a way to construct a classification rule using the selected probesets will be proposed. Next some experimental results on real datasets, on the relevance of selected probesets, and on the sensitivity and the stability of the method will be provided. Finally the method will be discussed.

**References**

Albert, J., Chib, S., 1993. Bayesian analysis of binary and polychotomous response data. Journal of the American Statistical Association 88 (422), 669–

679.

Chen, M., Dey, D., 2003. Variable selection for multivariate logistic regression models. Journal of Statistical Planning and Inference 111, 37–55.

Chipman, H., George, E., McCulloch, R., 2001. The practical implementation of bayesian model selection. In: Model selection - IMS Lecture Notes. P. LAHIRI. Institute of Mathematical Statistics.

Edgar, R., Domrachev, M., Lash, A., 2002. Gene expression omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Research 30 (1), 207–210.

George, E., McCulloch, R., 1993. Variable selection via gibbs sampling. Journal of the American Statistical Association 88 (423), 881–889.

Guyon, I., Weston, J., Barnhill, S., Vapnik, V., 2002. Gene selection for cancer classification using support vector machines. Machine Learning (46), 389–422.

Kuncheva, L., February 2007. A stability index for feature selection. In: Proceedings of the 25th IASTED International Multi-Conference, Artificial Intelligence and Applications. Innsbruck, Austria.

Lee, K., Sha, N., Dougherty, E., Vannucci, M., Mallick, B., 2003. Gene selection: a bayesian variable selection approach. Bioinformatics 19 (1), 90–97.

Liu, J., 1994. The collapsed gibbs sampler in bayesian computations with application to a gene regulation problem. Journal of the American Statistical Association 89 (427), 958–966.

O'Hara, R., Sillanpää., M., 2009. A review of bayesian variable selection methods: What, how and which. Bayesian Analysis 4 (1), 85–118.

Robert, C., Casella, G., 2004. Monte Carlo statistical methods, second edition. Springer.

Sha, N., Vannucci, M., Tadesse, M., Brown, P., Dragoni, I., Davies, N., Roberts, T., Contestabile, A., Salmon, M., Buckley, C., Falciani, F., 2004. Bayesian variable selection in multinomial probit models to identify molecular signatures of disease stage. Biometrics 60, 812–819.

Somol, P., Novovicova, J., 2008. Evaluating the stability of feature selectors that optimize feature subset cardinality. In: da Vitora Lobo et al., N. (Ed.), Lecture Notes in Computer Science, vol 5342. Springer-Verlag Berlin Heidelberg, pp. 956–966.

Tadesse, M., Sha, N., Vannucci, M., 2005. Bayesian variable selection in clustering high-dimensional data. Journal of the American Statistical Association 100, 602–617.

Tüchler, R., 2008. Bayesian variable selection for logistic models using auxiliary mixture sampling. Journal of Computational and Graphical Statistics 17 (1), 76–94.

Zhou, X., Liu, K., Wong, S., 2004a. Cancer classification and prediction using logistic regression with bayesian gene selection. Journal of Biomedical Informatics 37, 249–259.

Zhou, X., Wang, X., Dougherty, E., 2004b. A bayesian approach to non linear probit gene selection and classification. Journal of the Franklin Institute (341), 137–156.

Zhou, X., Wang, X., Dougherty, E., 2004c. Gene prediction using multinomial probit regression with bayesian gene selection. EURASIP Journal on Applied Signal Processing (1), 115–124.