Bayesian Analysis (2011) **6**, Number 2, pp. 209–230

Bayesian Variable Selection for Probit Mixed Models Applied to Gene Selection

Meli Baragatti*

Abstract. In computational biology, gene expression datasets are characterized by very few individual samples compared to a large number of measurements 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. Besides, the increased size of a merged dataset facilitates its re-splitting into training and validation sets. This necessitates the introduction of the dataset as a random effect. In this context, extending a 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 grouping (or blocking) 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 an illustration, it is used to select the most important genes that characterize the estrogen receptor status of patients with breast cancer.

Keywords: Bayesian variable selection, random effects, probit mixed regression model, grouping technique (or blocking technique), Metropolis-within-Gibbs algorithm

1 Introduction

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 or high throughput sequencing technologies are important for finding genes that are implicated in biological processes including development, disease, and response to treatment, and it plays an important role in the current tendency towards personalized medicine. Identified genes or sequences can be used to classify future observations, influencing the treatment of patients. However, these experiments are expensive, and datasets have often no more than 100 specimens. The goal, therefore, is to advance a method allowing variable selection from merged microarray datasets, each of them presenting its own individual experimental bias.

*Institut de Mathmatiques de Luminy (IML), CNRS Marseille and Ipsogen SA Luminy Biotech Entreprises, Marseille, France, mailto:baragattimeili@hotmail.com

 DOI:10.1214/11-BA607

Several 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. Recently Bottolo and Richardson (2010) proposed an algorithm based upon Evolutionary Monte Carlo. 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), Zhou et al. (2004c) and Yang and Song (2010) 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 in Albert and Chib (1993).

The motivation behind the variable selection method developed in this paper 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 may provide improved power, and facilitates its re-splitting into training and validation sets. In addition a merged set comprises more data diversity than an individual set, hence we can avoid bias due to a particular dataset as explained by various authors, see Cheng et al. (2010) and references therein. Among all the methods previously proposed for variable selection, 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 predictors, whereas the aim of this paper is to select a few predictors among tens of thousands in a Bayesian framework. Recently Frühwirth-Schnatter and Wagner (2010) considered variable selection for random effects, but in this paper we are more interested by variable selection for the fixed effects, assuming that random effects are present.

The approach developed in this paper extends the approach of George and McCulloch (1993) and Lee et al. (2003). George and McCulloch (1993) introduced latent variables to identify subsets of selected variables in a linear model. Then Lee et al. (2003) used these latent variables in a probit regression model, which is considered as part of a larger hierarchical Bayesian model. Our method extends the model used by Lee et al. (2003) by adding random effects. We are then confronted with several difficulties. One concerns the simulation of conditional distributions, since full conditional distributions cannot be directly simulated. A solution is to use the grouping (or blocking) technique of Liu (1994), and to combine Gibbs sampler and Metropolis-Hastings algorithms. Therefore the algorithm developed is a combination of the grouping method of Liu and the Metropolis-within-Gibbs algorithm (Robert and Casella 2004). A computational difficulty due to the large number of genes was also overcome by imposing a fixed number of selected genes at each iteration of the algorithm. As a consequence

the influence of the value chosen for *the variable selection coefficient* of our model is reduced. That represents an advantage, since the value of this coefficient can impact the results of other methods, see for instance Bottolo and Richardson (2010) who proposed to put a hyperprior distribution on this coefficient.

In this paper, Affymetrix microarray data are 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 were used, all with clinically defined ER status. One microarray experiment was done per patient, and ten of thousands of probesets were measured per experiment. The dataset is introduced as a random effect in the model, thus accounting for the different experimental conditions implicit in each set. The three merged datasets were split into training and validation sets, and the relevance of the selected probesets was 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 independent sets available from the NCBI GEO website. The stability and the sensitivity of the algorithm were also checked by using the relative weighted consistency measure of Somol and Novovicova (2008).

The remainder of the paper is organized as follows. Section 2 describes the probit mixed model with latent variables. Section 3 gives the full conditional distributions necessary for the Gibbs sampling algorithm, outlines the algorithm and proposes a way to construct a classification rule using the selected probesets. Section 4 provides some experimental results on real datasets, on the relevance of selected probesets, and on the sensitivity and the stability of the method. Finally Section 5 discusses the method.

2 Probit mixed model for gene selection

2.1 The hierarchical model

Suppose that n binary events are observed, denoted by the Y_i, $i = 1, \ldots, n$. The set of potential regressors is of size p, with $p \gg n$. The goal is to select a subset of regressors related to the events $Y_1, \ldots, Y_n$. The following probit mixed model is considered,

$$P(Y_i = 1 \mid U, \beta) = p_i = \Phi(X_i'\beta + Z_i'U),$$

where Φ stands for the standard Gaussian cumulative distribution function, and X_i and Z_i for the fixed and random effect regressors associated with the i^{th} observation. The parameter β corresponds to the fixed-effect coefficients and the parameter U to the

random-effect coefficients. X and Z are design matrices associated with the fixed and random effects.
Assuming that we have K random effects, $U = (U_1', \ldots, U_K')'$. Each U_l is of size q_l, and $\sum_{l=1}^{K} q_l = q$. The size of β is p.

Following Albert and Chib (1993) and Lee et al. (2003), a vector of latent variables L is introduced. We write $L = (L_1, \ldots, L_n)$ and we assume that $L \mid U, \beta \sim \mathcal{N}_n(X\beta + ZU, I_n)$ with I_n the identity matrix. We then have

$$Y_i = \begin{cases} 1 & \text{if } L_i > 0 \\ 0 & \text{if } L_i < 0, \end{cases} \tag{1}$$

To perform variable selection, a vector γ of p indicator variables is introduced:

$$\gamma_j = \begin{cases} 1 & \text{if } \beta_j \neq 0, \quad \text{variable } j \text{ selected} \\ 0 & \text{if } \beta_j = 0, \quad \text{variable } j \text{ not selected.} \end{cases}$$

Given γ, β_γ is the vector of all nonzero elements of β, and X_γ is the matrix X with only the columns corresponding to the elements of γ that are equal to 1.

2.2 Prior distributions

To complete the hierarchical model, some prior assumptions have to be made on $U \mid D$, $\beta_\gamma \mid \gamma$, γ and D, where D is a covariance matrix of dimension q.

- If the data supports $\gamma_j = 0$ over $\gamma_j = 1$, then the j^{th} variable will not be needed in the model and we can let $\beta_j = 0$. We then focus on the prior distribution of the non null vector β_γ. Like Lee et al. (2003), we take the following conventional prior:

$$\beta_\gamma \mid \gamma \sim \mathcal{N}_d(0, c(X_\gamma' X_\gamma)^{-1}), \qquad \text{with} \qquad d = \sum_{j=1}^{p} \gamma_j, \tag{2}$$

  This prior corresponds to the g-prior of Zellner (1986), and c is a positive scale factor specified by the user. Bottolo and Richardson (2010) called it the variable selection coefficient. Several authors discussed the choice of its value, see Chipman et al. (2001), George and Foster (2000), Clyde and George (2000) and Smith and Kohn (1997) among others. Raftery et al. (1997) used a similar form of prior. In our algorithm the value of c will be fixed, but will not be too influential (see the discussion).

- The γ_j are assumed to be independent Bernoulli variables, with

$$P(\gamma_j = 1) = \pi_j, \qquad 0 \leq \pi_j \leq 1.$$

  We do not want to use prior knowledge to favor any probesets, so we put $\pi_j = \pi$, $\forall j = 1, \ldots p$.

- The vector of coefficients associated with the random effects is assumed to be Gaussian and centered:
$$U \mid D \sim \mathcal{N}_q(0, D).$$

This definition allows three cases to be distinguished:

General case: No structure is assumed for the variance-covariance matrix D, its prior distribution is an Inverse-Wishart $\mathcal{W}^{-1}(\Psi, m)$.

Case of a block-diagonal matrix D: The different random effects are assumed independent. The vectors of coefficients associated with each random effect have Gaussian prior distributions:

$$U_l \mid A_l \sim \mathcal{N}_{q_l}(0, A_l), \quad l = 1, \ldots, K,$$

where the A_l are symmetric design matrices of dimension q_l. D is a block-diagonal matrix denoted by $diag(A_1, \ldots, A_K)$. The prior distributions for each A_l are Inverse-Wishart $\mathcal{W}^{-1}(\Psi, m)$.

Case of a diagonal matrix D: $D = diag(A_1, \ldots, A_K)$ where $A_l = \sigma_l^2 I_{q_l}$, $l = 1, \ldots, K$ and I_{q_l} the identity matrix. The prior distributions for the σ_l^2 are then Inverse Gamma $\mathcal{IG}(a, b)$ (b denotes the scale).

3 Bayesian sampler for variable selection

3.1 The conditional distributions

The posterior distribution of γ is of particular interest since it encapsulates the effectiveness of the different explanatory variables in explaining the variation in the responses Y. The number of possible explanatory variables is on the order of tens of thousands, so the number of possible γ-vectors is extremely large. The idea is to use a Gibbs sampling algorithm to explore this posterior distribution and search for high probability γ values.

In order to use the classical Gibbs sampler, we must be able to simulate from all of the full conditional distributions (simplified by the hierarchical structure): $f(L \mid Y, \beta, U)$, $f(\beta \mid L, U, \gamma)$, $f(U \mid L, \beta, D)$, $f(\gamma \mid L, U, \beta)$ and $f(D \mid U)$.

- Full conditional distribution of L.

$$\begin{aligned} L_i \mid \beta, U, Y_i = 1 &\sim \mathcal{N}(X_i'\beta + Z_i'U, 1) \quad \text{left truncated at 0} \\ L_i \mid \beta, U, Y_i = 0 &\sim \mathcal{N}(X_i'\beta + Z_i'U, 1) \quad \text{right truncated at 0.} \end{aligned} \tag{3}$$

- Full conditional distribution of β.
Given γ, we know which elements of β are not null. So we focus on the generation

of the non null elements of β_γ. Letting $V_\gamma = \frac{c}{1+c}(X'_\gamma X_\gamma)^{-1}$, we have

$$\beta_\gamma \mid L, U, \gamma \quad \sim \quad \mathcal{N}_d(V_\gamma X'_\gamma (L - ZU), V_\gamma) \qquad \text{with} \qquad d = \sum_{i=1}^{p} \gamma_i. \tag{4}$$

- Full conditional distribution of U.
  Defining $W = (Z'Z + D^{-1})^{-1}$, we have

$$U \mid L, \beta, D \sim \mathcal{N}_q(WZ'(L - X\beta), W). \tag{5}$$

- Full conditional distributions of γ.

$$f(\gamma \mid \beta_\gamma, L, U) \propto (2\pi)^{-\frac{d}{2}} \exp\Big[\frac{1}{2}\Big(L'X_\gamma\beta_\gamma + \beta'_\gamma X'_\gamma L - \beta'_\gamma X'_\gamma ZU - U'Z'X_\gamma\beta_\gamma \tag{6}$$
$$- \beta'_\gamma V_\gamma^{-1}\beta_\gamma\Big)\Big] \times \mid c(X'_\gamma X_\gamma)^{-1} \mid^{-\frac{1}{2}} \prod_{j=1}^{p} \pi_j^{\gamma_j}(1-\pi_j)^{1-\gamma_j}.$$

- Full conditional distribution of D.
  General case: The full conditional distribution of D is an Inverse-Wishart:

$$D \mid U \quad \sim \quad \mathcal{W}^{-1}(UU' + \Psi, m + 1). \tag{7}$$

  Case of a block-diagonal matrix D: $D = diag(A_1, \ldots, A_K)$. The full conditional distribution of A_l $(\forall l = 1, \ldots, K)$ is an Inverse-Wishart:

$$A_l \mid U_l \quad \sim \quad \mathcal{W}^{-1}(U_l U'_l + \Psi, m + 1). \tag{8}$$

  Case of a diagonal matrix D: $D = diag(A_1, \ldots, A_K)$, and $\forall l = 1, \ldots, K$, $A_l = \sigma_l^2 I_{q_l}$. The full conditional distribution of σ_l^2 is an Inverse-Gamma:

$$\sigma_l^2 \mid U_l \quad \sim \quad \mathcal{IG}\Big(\frac{q_l}{2} + a, \big(\frac{1}{2}U'_l U_l + b\big)\Big). \tag{9}$$

3.2 Use of the grouping technique

The classical Gibbs sampler cannot be used because the full conditional distribution of γ cannot be directly simulated (see (6)). However, this full conditional distribution can be simulated with a Metropolis-Hastings algorithm, and the complete algorithm would be a Metropolis-within-Gibbs algorithm. Roberts and Rosenthal (2006) have shown the Harris-recurrence of this algorithm, therefore its convergence is guaranteed. But even with a Metropolis-Hastings algorithm, the full conditional distribution of γ is difficult to obtain, since it depends on the actual value of β_γ. Thus the acceptance rate for a candidate γ^* in the Metropolis-Hastings algorithm will depend both on the actual $\gamma^{(t)}$ and $\beta_{\gamma^{(t)}}$, and on the proposed γ^* and β_{γ^*}. The problem is that β_{γ^*} is unknown.
To get around this problem, we combine the Metropolis-within-Gibbs algorithm with

the grouping (or blocking) technique of Liu (1994). The idea is to group the parameters β_γ and γ, so we will be interested in the full conditional distribution of $(\beta_\gamma, \gamma) \mid L, U$. This technique improves the algorithm and facilitates the convergence of the Markov chain, see Liu (1994) and van Dyk and Park (2008). We note that the sampler obtained is then a special case of a Partial Collapsed Gibbs Sampler, see van Dyk and Park (2008).
As we have

$$f(\beta_\gamma, \gamma \mid L, U) \propto f(\gamma \mid L, U) f(\beta_\gamma \mid \gamma, L, U),$$

we remark that simulating from the full conditional distribution $(\beta_\gamma, \gamma) \mid L, U$ is equivalent to simulating γ from its full conditional distribution integrated on β_γ, then simulating β_γ from its full conditional distribution. The "integrated distribution" for γ will not depend anymore on the nuisance parameter β_γ and will be easily simulated by a Metropolis-Hastings algorithm.
In each iteration of the algorithm, we will take care to simulate γ before $\beta\gamma$, to keep the dependence between $\beta\gamma$ and γ, as noted by van Dyk and Park (2008).

We use $f(L \mid \gamma, U)$ and the Bayes Theorem to get the integrated distribution of $\gamma \mid L, U$ (the target distribution):

$$\begin{aligned} f(\gamma \mid L, U) \quad \propto \quad & (1+c)^{-\frac{\sum \gamma_i}{2}} \exp\Big[-\frac{1}{2}\Big\{ (L - ZU)'(L - ZU) \\ & - \frac{c}{1+c}(L - ZU)' X_\gamma (X_\gamma' X_\gamma)^{-1} X_\gamma' (L - ZU) \Big\}\Big] \times \prod_{j=1}^{p} \pi_j^{\gamma_j} (1 - \pi_j)^{1-\gamma_j}. \end{aligned} \tag{10}$$

3.3 The Metropolis-within-Gibbs sampler modified by the grouping technique

A Metropolis-Hastings step to simulate γ

At iteration $(i+1)$ of the Metropolis-Hastings algorithm, a candidate γ^* will be proposed from $\gamma^{(i)}$. We want a symmetric transition kernel, to simplify the acceptance rate of the algorithm. The simplest way to have a symmetric transition kernel is to propose a γ^* which corresponds to $\gamma^{(i)}$ in which r components have been randomly changed (see Chipman et al. (2001) and George and McCulloch (1997)).

Given the target distribution (10), the acceptance rate ρ is then:

$$\begin{aligned} \rho(\gamma^{(i)}, \gamma^*) = min\Bigg\{ & \exp\Big[\frac{c}{2(1+c)} (L - ZU)' \Big(X_{\gamma^*} (X_{\gamma^*}' X_{\gamma^*})^{-1} X_{\gamma^*}' - X_{\gamma^{(i)}} (X_{\gamma^{(i)}}' X_{\gamma^{(i)}})^{-1} X_{\gamma^{(i)}}' \Big) \\ & \times (L - ZU) \Big] \times (1+c)^{\frac{\sum \left(\gamma_j^{(i)} - \gamma_j^*\right)}{2}} \times \Big(\frac{\pi}{1-\pi} \Big)^{\sum_1^p \left(\gamma_j^* - \gamma_j^{(i)}\right)}, 1 \Bigg\}. \end{aligned} \tag{11}$$

To facilitate the computation of the algorithm, the proposed γ^* still corresponds to $\gamma^{(i)}$ for which r components have been changed, but in such a way that the number of components whose values are 1 (and so the number of selected variables) is invariant. In so doing, $r/2$ components among the 1 values, and $r/2$ components among the 0 values are chosen at random and switched. There are several advantages to propose such a γ^*:

- In an iteration of the algorithm, if we have the number of variables selected d higher than the number of observations n, the $X'_\gamma X_\gamma$ matrix would be singular, and the prior distribution of β_γ could not be defined as in (2). An advantage of fixing the number of variables to be selected at each iteration is that this number cannot increase during a run of the algorithm, and if d is chosen lower than n this case of non singularity of $X'_\gamma X_\gamma$ is avoided.
- The acceptance rate is simplified, as we obtain $\sum \left(\gamma_j^{(t)} - \gamma_j^*\right) = 0$.
- The choice of the prior value for the variable selection coefficient c used in the prior distribution of β is less influent (see the discussion).

Remark. In the method of Lee et al. (2003), the γ vector is generated component by component at each iteration, while in our method a Metropolis-Hastings algorithm is used to generate it. There are two advantages to use a Metropolis-Hastings algorithm: it is computationally advantageous for a very large number of variables compared to a generation component by component, and it enables us to easily generate a γ vector with an invariant number of components whose values are 1.

The sampler

The Metropolis-within-Gibbs sampler modified by the grouping technique of Liu generates a sequence:

$$\gamma^{(1)}, \beta_\gamma^{(1)}, D^{(1)}, L^{(1)}, U^{(1)}, \ldots\ldots, \gamma^{(b+m)}, \beta_\gamma^{(b+m)}, D^{(b+m)}, L^{(b+m)}, U^{(b+m)}.$$

The sequence of the $\gamma^{(t)}$, which is of interest for the variable selection problem, is embedded in this "Gibbs sequence". To generate it, at each iteration γ is simulated from its integrated distribution and β_γ, L, U and D are simulated from their full conditional distributions.

Algorithm:
Starting with initial values $\gamma^{(0)}, \beta^{(0)}, D^{(0)}, L^{(0)}, U^{(0)}$. At iteration $t+1$:

1. Simulate $\gamma^{(t+1)}$ from $f(\gamma \mid L^{(t)}, U^{(t)})$ (see 10), using the Metropolis-Hasting (MH) step. The MH step begins with $\gamma^{(t)}$ as an initial value, and k iterations are performed given $L^{(t)}$ and $U^{(t)}$ (k arbitrarily fixed). At iteration $i+1$ of the MH step:
   (a) Generate the γ^* candidate, by randomly switching $r/2$ components among the 1 values, and $r/2$ components among the 0 values.

(b) Take

$$\gamma^{(i+1)} = \begin{cases} \gamma^* & \text{with probability} \quad \rho(\gamma^{(i)}, \gamma^*) \quad \text{see (11)} \\ \gamma^{(i)} & \text{with probability} \quad 1 - \rho(\gamma^{(i)}, \gamma^*) \end{cases}$$

$\gamma^{(t+1)}$ will be the $\gamma^{(k)}$ obtained at the k^{th} iteration of the MH step.

2. Simulate $\beta_\gamma^{(t+1)}$ from $f(\beta_\gamma \mid L^{(t)}, U^{(t)}, \gamma^{(t+1)})$ (see (4)).
3. Simulate $D^{(t+1)}$ from $f(D \mid U^{(t)})$ (see (7), (8) or (9)).
4. Simulate $L^{(t+1)}$ from $f(L \mid Y, \beta^{(t+1)}, U^{(t)})$ (see (3)).
5. $U^{(t+1)}$ from $f(U \mid L^{(t+1)}, \beta^{(t+1)}, D^{(t+1)})$ (see (5)).

We use the fact that $X\beta = X_\gamma \beta_\gamma$ and that β can be obtained from γ and β_γ. The number of iterations is $b + m$, where b corresponds to the burn-in period and m to the observations from the posterior distributions.

The selected probesets

For selection of variables, the sequence $\{\gamma^{(t)} = (\gamma_1^{(t)}, \ldots, \gamma_p^{(t)}), t = b+1, \ldots, b+m\}$ is used. The most relevant variables for the regression model are those which are supported by the data and prior information. Thus they are those corresponding to the γ components with higher posterior probabilities, and can be identified as the γ components that are most often equal to 1.

3.4 Classification and prediction

Once a set of relevant variables have been selected, it can be used to fit a probit mixed model in a classical way and to classify future observations. However, if more variables than necessary to fit a probit mixed model have been selected in the Bayesian selection step, a second selection has to be performed on them in order to build a reliable probit mixed model. This second selection is performed on the training set using standard selection tools like AIC, BIC, Bayes factors,.... The final probit mixed model can be tested on the validation set. Moreover, the variables selected in the Bayesian selection step can be used in other classification methods, such as Support Vector Machines (but random effects are not taken into account).

4 Experimental results

4.1 Application to the ER status of patients with breast cancer

Description of the datasets

Three different datasets were used: 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). Each dataset was treated for background noise and normalized with respect to a reference distribution by the RMA procedure (Irizarry et al. 2003). Each dataset was split into a training set and a validation set having the same proportions of ER positive and ER negative observations. Then the three training datasets were merged on one side (497 patients) and the three validation sets were merged on the other side (88 patients).
For each patient, more than 54000 probesets were available. Two filters were applied on all of these probesets. Only the probesets sufficiently expressed so that they can be differentiated from noise and which could not be considered as invariant were kept, resulting in 19384 probesets. The goal was to select only a few probesets which are related to the ER status of the patients, by taking into account the different experimental conditions between the different merged datasets.

In this illustration, there are thousands of fixed regressors corresponding to the expression measurements of probesets, and only one random effect, which corresponds to the different datasets. X_{ij} corresponds to the measurement of the expression level of the j^{th} probeset for the i^{th} patient, and $Z_{il} = 1$ if the i^{th} patient is from the l^{th} dataset, 0 otherwise.

Prior settings for the algorithm

- Following the recommendations of Smith and Kohn (1997), a value of $c = 50$ was chosen for the variable selection coefficient used in the prior distribution of β.
- Thirty probesets were selected at each iteration of the Gibbs sampler, when γ is generated; $r = 10$ of them were changed at each iteration of the Metropolis-Hastings step (5 zeros and 5 ones).
- The random effect corresponds to the dataset, and the three datasets are considered independent: they were generated in different countries, by different teams, using different equipments and different patients. Therefore the variance-covariance matrix of the random effect D was a diagonal matrix 3×3 with $A_1 = \sigma_1^2 I_3$. Gelman (2006) noted that an inverse-gamma prior should not be too non-informative, otherwise serious problems can arise. Given our data, we knew that σ_1^2 is probably not too high, and a $\mathcal{IG}(2, 3)$ seemed reasonable for the prior distribution of σ_1^2.
- For the Metropolis-within-Gibbs sampler modified by the grouping technique, 60000 iterations were computed, among which 30000 were burn-in iterations. For the Metropolis-Hastings step in this sampler, 500 iterations were computed, and the simulated γ was the one corresponding to the 500^{th} iteration.

Results and predictions

We performed a first selection of variables by selecting the top-rank probesets, those which have been selected the most often. A boxplot can help, see Figure 1. Forty

probesets were selected at least once from the 30000 post-burn-in iterations of the simulated Markov Chain for γ. Twenty three probesets were selected in the 30000 iterations, and thirty were selected at least from 20000 iterations. There is a gap between probesets selected in more than 20000 iterations and others, so the first selection is made of these probesets selected at least in 20000 iterations.
A second selection was performed on the thirty probesets from the first selection, to build a reliable probit mixed model. This second selection was performed on the training set, using a stepwise selection (with AIC and BIC criteria) and the classification performance of the model on the validation set. Five probesets were kept: Affymetrix symbols `228241_at`, `205862_at`, `202376_at`, `216222_s_at` and `1568760_at`. See Table 1 for the associated gene symbols and coefficients. The estimated random effects of this final model were reasonable: -0.284 for the dataset from the Institut Paoli Calmettes, 0.199 for the GSE2109 dataset and 0.087 for the GSE5460 dataset.

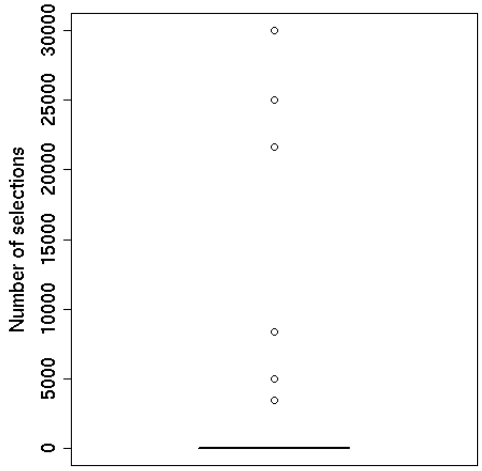

Figure 1: Boxplot of the number of selections of a probeset after the burn-in period, for the real datasets example. Forty probesets were selected at least once, all of the other probesets were never selected. A point represents a probeset (or several probesets if they have been selected the same number of times).

Using this 5-probeset model, two methods were used to predict the ER status of the patients in the validation set:

1. Using knowledge of the dataset to which each patient belonged and using the estimated random effects coefficients.
2. The estimated random effects coefficients are not used in order to mimic a real-life scenario of an experiment for a patient coming from an unknown dataset.

The patients were predicted positive if their probability to be positive was higher than 0.5 and negative if it was lower than 0.5. The two methods gave us the same

Probeset	gene	Coefficient	Pvalue
Intercept		-9.12074	1.92e-05
228241_at	AGR3	0.45046	1.12e-15
205862_at	GREB1	0.77639	4.18e-08
202376_at	SERPINA3	0.37965	0.000149
216222_s_at	MYO10	-0.63551	0.004967
1568760_at	MYH11	0.42742	0.050219

Table 1: Probesets selected in the final model and associated coefficients.

predictions, which were very good: a specificity of 1 and a sensitivity of 0.98 (1 wrong predictions among 88), see Figure 2.

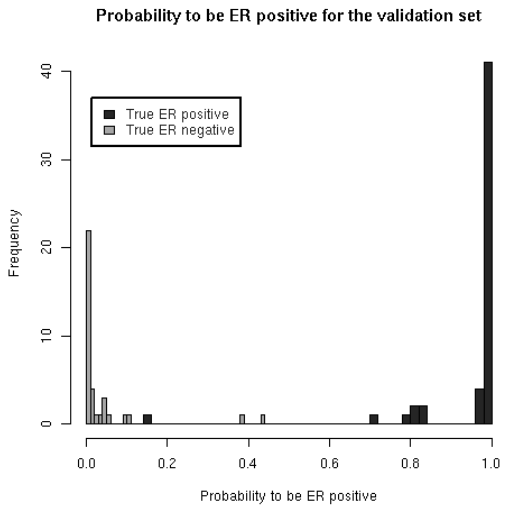

Figure 2: Histogram of probabilities to be ER positive given by the final model, for patients from the validation set.

Remark. In biomedical studies, when continuous variables are often reclassified as binary, it is common to define an "undetermined zone" of probabilities for which no prediction are given. Indeed, it is sometimes better than giving a wrong prediction, because these predictions imply treatments. Defining an "undetermined zone" between 10% and 90% probability of being positive, false predictions were eliminated, and 10 were considered undetermined (11.4%) (estimated random effects coefficients not used).

As a final test of our model, two more independent datasets were brought in from the NCBI GEO website: the GEO series GSE6532 and the GEO series GSE12763. The

random effects associated with these datasets were entirely unknown, simulating an even more realistic case of prediction for a patient coming from an unknown dataset. Once again the results were very good : only 1 wrong prediction among 29 for the GSE12763 dataset, and no wrong predictions among 86 for the GSE6532 dataset.

4.2 Sensitivity and stability studies

The sensitivity and the stability of the algorithm were assessed by using the relative weighted consistency measure of Somol and Novovicova (2008), denoted by CW_{rel}. It is a measure evaluating how much subsets of selected variables for several runs overlap, and it shows the relative amount of randomness inherent in the concrete variable selection process. It takes values between 0 and 1, where 0 represents the outcome of completely random occurrence of variables in the selected subsets and 1 indicates the most stable variable selection outcome possible.
Stability is defined as sensitivity to variations in the training set. Referring to our breast cancer data set, 4000 probesets were randomly chosen from among the 19384 originally available. Since the aim here was only to check the sensitivity and stability of the method, these 4000 were not chosen in relation to the ER status.
Several runs of the algorithm were performed, and are reported in Table 2. Concerning the stability, the algorithm was run on three different training sets of 497 microarrays (among 585), using the same prior values for the hyperparameters. Concerning the sensitivity, the algorithm was run on the same training set with different values of c, different prior distributions for σ_1^2, different numbers of probesets to be selected at each iteration of the algorithm and different numbers of iterations. For the prior distributions for σ_1^2, we chose a $\mathcal{IG}(2,3)$ which seemed reasonable given our data, a $\mathcal{IG}(2,5)$ to have a prior favoring higher values compared to the first one, a $\mathcal{IG}(3,1)$ to favor lower values, and a $\mathcal{IG}(1,1)$ to have a non-informative prior without too small parameters to avoid problems, see Gelman (2006).

For each run, a reasonable number of probesets could be easily selected. Indeed, two to ten probesets were selected much more often than the others, see Figure 3 (two to four probesets were selected for most of the runs). Hence there is no need to perform a second selection, as in section 4.1. To compare the results of the different runs, the relative weighted consistency measure of Somol and Novovicova CW_{rel} was used.

Using the results of the 15 runs together, $CW_{rel} = 0.398$. Subsets of selected variables for the different runs overlapped: among the 15 runs the probesets `215552_s_at`, `209603_at` and `209602_s_at` were kept in 12, 12 and 6 runs respectively. Apparently the prior for σ^2 ($CW_{rel} = 0.292$) has more impact than the number of probesets to be selected at each iteration ($CW_{rel} = 0.5$). We note that the number of probesets to be selected at each iteration of the algorithm and the number of iterations do not seem to modify the number of probesets more often selected than the others during the run.

This was satisfying and the method appears relatively stable. First because the ran-

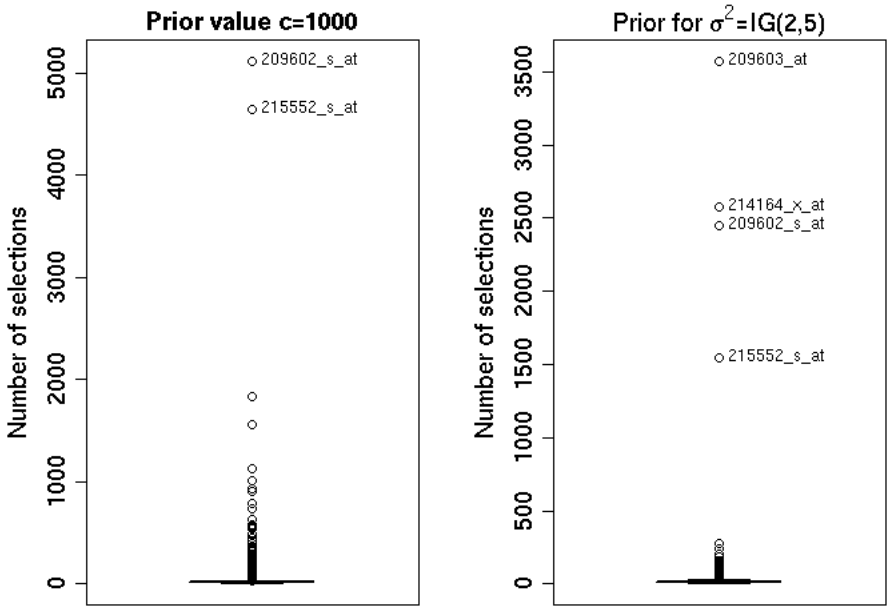

Figure 3: Boxplot of the number of selections of a probeset after the burn-in period, for two runs of the sensitivity analysis. A point represents a probeset (or several probesets if they have been selected the same number of times). The left boxplot corresponds to the run with $c = 1000$: there is a gap between the two probesets selected in more than 4000 iterations and the others, hence we selected these two probesets. The right boxplot corresponds to the run with $\sigma_1^2 \sim \mathcal{IG}(2, 5)$: there is a gap between the four probesets selected in more than 1500 iterations and the others, hence we selected these four probesets.

dom selection of 4000 probesets carries a risk of destabilization of the results, since these 4000 are not necessarily those which are most indicative of ER status. Secondly, several probesets can represent the same gene, and different genes can be implied in the same biological pathway. Thus, it is possible that subsets of probesets are more similar than they appear, and therefore that CW_{rel} is underestimated. For example, the probesets `209603_at` and `209602_s_at` mentioned above both represent the gene GATA3. Finally, these simulations indicate that there is not a parameter whose choice introduces more sensitivity than the others.

4.3 Comparison with other methods

We compared the performance of our method with the performances of other methods which do not take into account random effects: we considered the model of Lee et al.

Simu	Data-set	Value of c	Prior for σ^2	Nb selected at each iter of the algo (nb changed at each iter of the MH)	Nb iter total	burn-in	CW_{rel}
1 2 3	1 2 3	50	$IG(2,3)$	15 (6)	12000	6000	0.25
4 5 6 7	1	10 50 100 1000	$IG(2,3)$	15 (6)	12000	6000	0.375
8 9 10 11	1	50	$IG(2,3)$ $IG(1,1)$ $IG(1,1)$ $IG(2,5)$	15 (6)	12000	6000	0.292
12 13 14	1	50	$IG(2,3)$	15 (6) 5 (2) 30 (10)	12000	6000	0.5
15	1	50	$IG(2,3)$	15 (6)	30000	15000	

Table 2: Parameters of the runs for the stability and sensitivity study and associated relative weighted consistency measure of Somol and Novovicova CW_{rel}. For the Metropolis-Hastings step, always 500 iterations are computed.

(2003) and Support Vector Machine with Recursive Feature Elimination of the variables, with linear or non linear kernels (Guyon et al. 2002). We used simulated data: 200 observations of 1000 variables following a uniform distribution $\mathcal{U}_{[-5,5]}$ are generated. We assumed that 5 variables and a random effect U of size 4 explain a vector of binary variables Y by a probit mixed model:

$$\begin{aligned} p_i &= \Phi(X_i'\beta + Z_i'U), \qquad i = 1, \ldots, 200 \\ Y_i &\sim \mathcal{B}(p_i), \qquad i = 1, \ldots, 200 \end{aligned}$$

We took $\beta_\gamma = (-1, -1, 1, 1, 2)$ and we assumed that 50 observations are coming from each modality of the random effect. Different values of U were used: $U1 = (0, 0, 0, 0)$, $U2 = (-3, -2, 2, 3)$, $U3 = (-5, -3, 3, 5)$, $U4 = (-10, -5, 5, 10)$ and $U5 = (-30, -10, 10, 30)$. This set of 200 observations was splitted into training and validation sets, each of them of size 100, with 25 observations coming from each modality of the random effect. For our method and the method of Lee et al. we took $c = 50$, 5 probesets were selected at each iteration of the Gibbs sampler and $r = 2$ of them were changed at each iteration of the Metropolis-Hastings step (1 zero and 1 one), D was a diagonal matrix 3×3 with $A_1 = \sigma_1^2 I_3$ and a prior $\mathcal{IG}(1, 1)$ was chosen for σ_1^2, 500 iterations were performed for each Metropolis-Hastings step, and a total of 3000 and 5000 iterations were performed for the whole algorithm.
Concerning our method and the method of Lee et al. (2003), the top-ranked variables

(variables selected more often than others, box-plots were used) were used to perform predictions on the validation set. The RFE-SVM method gave us directly sets of "best variables" and associated models, and these models were used to perform the predictions. The results obtained are in Table 3.

Random effect	Our method		Lee et al. (2003)		RFE-SVM	
U	3000 iter	5000 iter	3000 iter	5000 iter	linear	non linear
$U1$	17	26	19	22	25	23
$U2$	19	21	19	19	20	26
$U3$	21	23	24	24	25	26
$U4$	19	19	35	35	29	31
$U5$	14	11	44	44	52	56

Table 3: Number of misclassifications on the validation set, for different methods and different random effects.

When there is no random effect or when the magnitude of the random effect is small, our method is comparable to the one of Lee et al., and the results of these two methods are better than or comparable with those obtained by RFE-SVM. But when the magnitude of the random effect is high, especially for $U4$ and $U5$, it appears that our method outperforms the method of Lee et al. and the RFE-SVM method.

5 Discussion

In this article we have developed an approach for Bayesian gene selection for a probit mixed model, as an extension of previous works by George and McCulloch (1993) and Lee et al. (2003). An important contribution of our method is that it allows selection of variables in a mixed framework, taking into account the design of the data. It is particularly useful for gene selection, as it enables the use of merged datasets in order to introduce more observations and greater diversity. That may provide improved power, and we can avoid bias due to a particular dataset. The increased size of a merged dataset facilitates its re-splitting into training and validation sets, hence we do not need to evaluate the performance of a classification rule by a cross-validation procedure. It is advantageous compared to other methods which do not take into account random effects. Indeed, as these methods can use only one dataset which is usually of small size, they often need to perform leave-one-out-cross-validation, which can be time-consuming (see Lee et al. (2003), Yang and Song (2010), Sha et al. (2004), Zhou et al. (2004b) and Zhou et al. (2004c) for instance). On the contrary, if several datasets are merged then a separated training set can be used and the performance of a classifier can be directly obtained on it. Using simulations to make comparisons with other methods which do not take into account random effects, we showed that the proposed method is comparable to others when the magnitude of the random effects is low, but performs better than the others for classification when the magnitude of the random effects is high. This method should prove widely useful in microarray bioinformatics, since many diverse

datasets are freely available on the Internet. But it can also be used for data obtained from high throughput sequencing technologies, which will probably be used a lot in few years. Indeed, the method can be applied when we have a matrix with $n << p$ and an associated vector of random effects.

In practice, before running an analysis, one must decide how many variables will be selected at each iteration of the Gibbs sampler. We do not consider this to be a drawback, since in order to have a reliable selection, the number of probesets should be limited compared to the size of the training set. Besides, fixing the number of variables selected at each iteration is a computational advantage, as discussed in section 3.3. In particular, the singularity of the $X'_\gamma X_\gamma$ matrix is avoided.
In addition, one must choose a value for the hyperparameter c which is large enough to have a relatively non-informative prior. Only the simulations of $\beta_\gamma \mid L, U, \gamma$ and of $\gamma \mid L, U$ directly depend on c. Concerning $\beta_\gamma \mid L, U, \gamma$, we can see in (4) that the density is proportional to $c/(1+c)$, which is relatively close to 1 if c is large. Concerning the density of $\gamma \mid L, U$ (see (10)), it depends on $c/(1+c)$ and on $(1+c)^{-\frac{\sum \gamma_i}{2}}$. The factor $(1+c)^{-\frac{\sum \gamma_i}{2}}$ does not play a role in the simulation of γ, because the number of variables to be selected at each iteration is fixed: this factor vanishes in the acceptance rate of the Metropolis-Hastings step of the algorithm. Therefore the value chosen should not be too influential, as long as it is large enough. We chose arbitrarily $c = 50$, following Smith and Kohn's (1997) recommendations. However different authors suggested different ranges, see Chipman et al. (2001), George and Foster (2000) and Clyde and George (2000) among others. For example Zhou et al. (2004b) and Zhou et al. (2004c) used $c = 10$, and Lee et al. (2003) used $c = 100$. However, it is possible to include another level in our Bayesian hierarchical model and to put a prior distribution on c. Zellner and Siow (1980) for instance proposed a mixture of g-priors and an inverse-gamma prior on c. Recently Bottolo and Richardson (2010) considered putting a hyperprior on c and using a Metropolis-within-Gibbs with adaptive proposal for updating this coefficient. In our application this coefficient was held fixed for convenience, and good results were obtained. Besides the sensitivity study showed us that the method is not overly sensitive to the value chosen for c, as expected.
More generally, it appeared that the algorithm is fairly stable to variations in the training set, and is robust to prior value of any of the hyperparameters.

Convergence could not be verified because we did not have formal diagnostic tools to prove it, as the parameters vectors used in the proposed algorithm were not associated to the same variables from one iteration to the next. Besides, the different runs could have converged to a local mode of the posterior distribution of γ, and not to a global one. But the results obtained in the stability and sensitivity analyses were satisfactory, as different runs with different starting points and different prior hyperparameters selected broadly the same variables, which means that these different chains had basically the same behavior. From our experience, it appeared that having a total number of iterations equal to three times the size of the set of predictors is sufficient, the results

were not significantly different when more iterations were performed.

The probesets selected by our method to characterize the estrogen receptor status enabled us to fit a model with good predictions. Moreover, three genes among the five used in the model were also selected using a Support Vector Machine method (twenty-four genes were selected by SVM), and another group of three among those five is known to be associated with estrogen receptor pathways and breast cancer: GREB1 (Nagaraja et al. 2006; Townson and O'Connell 2006; Rae et al. 2005), SERPINA3 (Cimino et al. 2008) and MYH11 (Singh and Chaturvedi 2009). Therefore, it seems that the probesets selected by our method are quite biologically relevant.

The algorithm developed is efficient and feasible, even for very large datasets with around 20000 variables. Therefore this approach has a clear advantage over other selection methods which handle less variables or which do not take into account random effects. However, Bayesian variable selection is an active research area, and it would be interesting to combine our method with recent proposals. For instance by studying the performance of the method with other prior distributions for σ^2, like half-Cauchy or folded-noncentral-t distributions, see Gelman (2006). Or by putting a prior distribution on c, like in Bottolo and Richardson (2010). It would also be of interest to consider an alternative prior distribution for β_γ to handle a non-invertible $X'_\gamma X_\gamma$ (when γ is itself singular or when $n < d$), by combining our approach with the concept of ridge regression (work in progress, Baragatti and Pommeret (2011)).

References

Albert, J. and Chib, S. (1993). "Bayesian analysis of binary and polychotomous response data." *Journal of the American Statistical Association*, 88(422): 669–679. 210, 212

Baragatti, M. and Pommeret, D. (2011). "Ridge parameter for g-prior distribution in probit mixed models with collinearity." *arxiv:1102.0470*. 226

Bottolo, L. and Richardson, S. (2010). "Evolutionary Stochastic Search for Bayesian Model Exploration." *Bayesian Analysis*, 5(3): 583–618. 210, 211, 212, 225, 226

Chen, M. and Dey, D. (2003). "Variable selection for multivariate logistic regression models." *Journal of Statistical Planning and Inference*, 111: 37–55. 210

Cheng, W., Tsai, M., Chang, C., Huang, C., Chen, C., Shu, W., Lee, Y., Wang, T., Hong, J., Li, C., and Hsu, I. (2010). "Microarray meta-analysis database (M2DB): a uniformly pre-processed, quality controlled, and manually curated human clinical microarray database." *BMC Bioinformatics*, 11. 210

Chipman, H., George, E., and McCulloch, R. (2001). "The practical implementation of Bayesian model selection." In *Model selection - IMS Lecture Notes*. P. LAHIRI. Institute of Mathematical Statistics. 210, 212, 215, 225

Cimino, D., Fuso, L., Sfiligoi, C., Biglia, N., Ponzone, R., Maggiorotto, F., Russo, G., Cicatiello, L., Weisz, A., Taverna, D., Sismondi, P., and De-Bortoli, M. (2008). "Identification of new genes associated with breast cancer progression by gene expression analysis of predefined sets of neoplastic tissues." *International Journal of Cancer*, 123(6): 1327–1338. 226

Clyde, M. and George, E. (2000). "Flexible empirical Bayes estimation for wavelets." *Journal of the Royal Statistical Society B*, 62(4): 681–698. 212, 225

Edgar, R., Domrachev, M., and Lash, A. (2002). "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository." *Nucleic Acids Research*, 30(1): 207–210. 210, 218

Frühwirth-Schnatter, S. and Wagner, H. (2010). "Bayesian Variable Selection for Random Intercept Modeling of Gaussian and non-Gaussian Data." In Bernardo, J. et al. (eds.), *Bayesian Statistics 9*, Proc. of the 9th Valencia Internat. Meeting, 165–200. Oxford University Press. 210

Gelman, A. (2006). "Prior distributions for variance parameters in hierarchical models (Comment on Article by Browne and Draper)." *Bayesian Analysis*, 1(3): 515–534. 218, 221, 226

George, E. and Foster, D. (2000). "Calibration and empirical Bayes variable selection." *Biometrika*, 87(4): 731–747. 212, 225

George, E. and McCulloch, R. (1993). "Variable selection via Gibbs sampling." *Journal of the American Statistical Association*, 88(423): 881–889. 210, 224

— (1997). "Approaches for Bayesian variable selection." *Statistica Sinica*, 7: 339–373. 215

Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). "Gene selection for cancer classification using support vector machines." *Machine Learning*, (46): 389–422. 210, 223

Irizarry, R., Hobbs, B., Collin, F., Beazer-Barclay, Y., Antonellis, K., Scherf, U., and Speed, T. (2003). "Exploration, normalization, and summaries of high density oligonucleotide array probe level data." *Biostatistics*, 4(2): 249–264. 218

Lee, K., Sha, N., Dougherty, E., Vannucci, M., and Mallick, B. (2003). "Gene selection: a Bayesian variable selection approach." *Bioinformatics*, 19(1): 90–97. 209, 210, 212, 216, 222, 223, 224, 225

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. 209, 210, 215

Nagaraja, G., Othman, M., Fox, B., Alsaber, R., Pellegrino, C., Zeng, Y., Khanna, R., Tamburini, P., Swaroop, A., and Kandpal, R. (2006). "Gene expression signatures and

biomarkers of noninvasive and invasive breast cancer cells: comprehensive profiles by representational difference analysis, microarrays and proteomics." *Oncogene*, 25(16): 2328–2338. 226

O'Hara, R. and Sillanpää., M. (2009). "A review of Bayesian variable selection methods: What, How and Which." *Bayesian Analysis*, 4(1): 85–118. 210

Rae, J., Johnson, M., Scheys, J., Cordero, K., Larios, J., and Lippman, M. (2005). "GREB 1 is a critical regulator of hormone dependent breast cancer growth." *Breast Cancer Research and Treatment*, 92(2): 141–149. 226

Raftery, A., Madigan, D., and Hoeting, J. (1997). "Bayesian model averaging for linear regression models." *Journal of the American Statistical Association*, 92: 179–191. 212

Robert, C. and Casella, G. (2004). *Monte Carlo statistical methods, second edition*. Springer. 209, 210

Roberts, G. and Rosenthal, J. (2006). "Harris recurrence of Metropolis-within-Gibbs and trans-dimensional Markov chains." *The Annals of Applied Probability*, 16(4): 2123–2139. 214

Sha, N., Vannucci, M., Tadesse, M., Brown, P., Dragoni, I., Davies, N., Roberts, T., Contestabile, A., Salmon, M., Buckley, C., and Falciani, F. (2004). "Bayesian variable selection in multinomial probit models to identify molecular signatures of disease stage." *Biometrics*, 60: 812–819. 210, 224

Singh, D. and Chaturvedi, R. (2009). "Recent patents on genes and gene sequences useful for developing breast cancer detection systems." *Recent Patents on DNA and Gene Sequences*, 3(2): 139–147. 226

Smith, M. and Kohn, R. (1997). "Non parametric regression using Bayesian variable selection." *Journal of Econometrics*, 75: 317–344. 212, 218

Somol, P. and 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*, 956–966. Springer-Verlag Berlin Heidelberg. 211, 221

Tadesse, M., Sha, N., and Vannucci, M. (2005). "Bayesian variable selection in clustering high-dimensional data." *Journal of the American Statistical Association*, 100: 602–617. 210

Townson, S. and O'Connell, P. (2006). "Identification of estrogen receptor alpha variants in breast tumors: implications for predicting response to hormonal therapies." *Journal of Surgical Oncology*, 94(4): 271–273. 226

Tüchler, R. (2008). "Bayesian variable selection for logistic models using auxiliary mixture sampling." *Journal of Computational and Graphical Statistics*, 17(1): 76–94. 210

van Dyk, D. and Park, T. (2008). "Partially Collapsed Gibbs Samplers: Theory and Methods." *Journal of the American Statistical Association*, 103: 790–796. 215

Yang, A. and Song, X. (2010). "Bayesian variable selection for disease classification using gene expression data." *Bioinformatics*, 26(2): 215–222. 210, 224

Zellner, A. (1986). *Bayesian Inference and Decision Techniques – Essays in honour of Bruno De Finetti*, chapter On assessing prior distributions and Bayesian regression analysis with g-prior distributions., 233–243. Amsterdam. 212

Zellner, A. and Siow, A. (1980). "Posterior Odds Ratios for Selected Regression Hypotheses." In *Bayesian Statistics: Proceedings of the First International Meeting Held in Valencia*, 585–603. University of Valencia Press. 225

Zhou, X., Liu, K., and Wong, S. (2004a). "Cancer classification and prediction using logistic regression with Bayesian gene selection." *Journal of Biomedical Informatics*, 37: 249–259. 210

Zhou, X., Wang, X., and Dougherty, E. (2004b). "A Bayesian approach to non linear probit gene selection and classification." *Journal of the Franklin Institute*, (341): 137–156. 210, 224, 225

— (2004c). "Gene prediction using multinomial probit regression with Bayesian gene selection." *EURASIP Journal on Applied Signal Processing*, (1): 115–124. 210, 224, 225

Acknowledgments

We would like to thank an Associate Editor and two anonymous reviewers for careful reading of the paper and constructive comments which have led to an improvement of the manuscript. We are grateful to Dr Daniel Birnbaum and Pr Franois Bertucci from the Dpartement d'Oncologie Molculaire of the Institut PAOLI CALMETTES (Marseille, France) for permission to use their data. We also thank Pr Denys Pommeret and Rebecca Tagett for useful discussions and comments.