Applications of the Mode Oriented Stochastic Search (MOSS) Algorithm for Discrete Multi-way Data to Genomewide Studies

December 9, 2009

Adrian Dobra[1], Laurent Briollais[2], Hamdi Jarjanazi[2], Hilmi Ozcelik[2] and Hélène Massam[3]

Abstract

We present a Bayesian variable selection procedure that is applicable to genomewide studies involving a combination of clinical, gene expression and genotype information. We use the Mode Oriented Stochastic Search (MOSS) algorithm of Dobra and Massam (2010) to explore regions of high posterior probability for regression models involving discrete covariates and to perform hierarchical log-linear model search to identify the most relevant associations among the resulting subsets of regressors. We illustrate our methodology with simulated data, expression data and SNP data.

Key words: Bayesian analysis; contingency tables; expression data; log-linear models; model selecton; SNP data; stochastic search; variable selection.

1 Introduction

High-throughput sequencing studies together with clinical and physiological data produce large amounts of biological information that is used for molecular phenotyping of many diseases. Due to the extremely small sample size relative to the total number of possible covariates, it is imperative that the key feature in the development of predictive models based on a combination of gene expression, genotype and phenotype data should be the selection of a small number of predictors. Many variable selection approaches proposed in the literature involve using univariate rankings that individually measure the dependency between each candidate predictor and the response – see, for example, Golub et al. (1999); Nguyen and

[1]Department of Statistics, University of Washington, Seattle
[2]Samuel Lunenfeld Research Institute, Mount Sinai Hospital, Toronto
[3]Department of Mathematics and Statistics, York University, Toronto

Rocke (2002); Dudoit et al. (2002); Tusher et al. (2001). Besides the need to address complex issues related to assessing the statistical significance of a huge number of null hypothesis tests associated with each individual predictor (Benjamini and Hochberg (1995); Efron and Tibshirani (2002); Storey and Tibshirani (2003)), there is no theoretical justification of the implicit claim that the resulting set of predictors can be converted into a good classification model for two reasons: (i) the variable selection criteria are not necessarily related to the actual classification method; (ii) there is no clear way to include other candidate variables if the covariates already selected do not lead to the desired classification peformance.

The alternative is to take into consideration *combinations* of predictors, that leads to an exponential increase of the number of candidate models. Exhaustively enumerating all these models is computationally infeasible for genomewide studies, thus a lot of efforts have been invested in finding stochastic search algorithms that are capable of quickly discovering regions of interest in the model space. In the context of linear regression, the stepwise methods of Furnival and Wilson (1974) can only be used for very small datasets due to their inability to escape local modes created by complex patterns of collinear predictors. A significant step forward were Markov chain Monte Carlo (MCMC) algorithms that explore the models space by sampling from the joint posterior distribution of the candidate models and regression parameters – see, for example, George and McCulloch (1993, 1997); Green (1995); Raftery et al. (1997); Nott and Green (2004). Excellent review papers about Bayesian variable selection for Gaussian linear regression models are Carlin and Chib (1995), Chipman et al. (2001) and Clyde and George (2004). Lee et al. (2003) make use of MCMC techniques in the context of probit regression to develop cancer classifiers based on expression data. Theoretical considerations related to the choice of priors for regression parameters are discussed in Fernández et al. (2003) and Liang et al. (2008).

Of particular interest for us is the analysis of genomewide single-nucleodite polymorphisms (SNPs) data – see, for example, Christensen and Murray (2007). While some relative consensus seems to emerge on the design of such studies, emphasizing the advantage of multistage designs in terms of cost-efficiency, their analysis raises many methodological questions, yet not answered – see Thomas et al. (2005). Hoggart et al. (2008) address the problem of multiple testing in SNP data arising from multiple studies and genotyping platforms. Although Schaid (2004) points out that single-marker analysis exploits a fraction of the information that exists in multilocus genotype data, only a few papers proposed analytic approaches that go beyond this exhaustive single-marker association testing. Marchini et al. (2005); Zhang and Liu (2007) showed the feasibility of genome-scale testing of the joint effect of two markers, but their extension to higher dimensionality problems still needs to be demonstrated. Wang and Abbott (2008) describe a multiple regression approach in which the predictors are the principal components of the sample covariance matrix of the SNP genotype scores. Clark et al. (2007) identify combinations of SNPs and environmental factors associated with disease status by extending the logic trees of Ruczinski et al. (2003) to a Bayesian framework. Swartz et al. (2007) discuss the use of model selection and Bayesian methods in the context of genomewide association studies (GWAS, henceforth) with the aim of selecting the best subsets of markers associated with the outcome of interest. Cordell

and Clayton (2002) give a stepwise logistic regression procedure applicable for small, tightly linked subsets of the genome, while Lunn et al. (2006) extend this approach to a Bayesian framework that accounts for uncertainty in variable selection and missing data. Verzilli et al. (2006) proposed Bayesian graphical models for GWAS discrete data. Model selection and fitting is based on the MCMC approach of Madigan and York (1995). This methodology focused only on decomposable graphical models restricted to SNPs physically close to each other, thereby potentially ignoring the complex nature of association patterns in GWAS. Therefore there is a compelling need for developing appropriate methodologies for GWAS, allowing an efficient evaluation of a large number of multi-marker models, where the markers can be linked or unlinked.

MCMC methods can have a slow convergence rate due to the high model uncertainty resulting from the small number of available samples. This uncertainty is readily apparent when there are many models each with a very small posterior probability and the total posterior probability of these models approches one. Any MCMC approach might fail to discover better models in a reasonable number of iterations because it spends most of its time moving around models that are ultimately not relevant. MCMC algorithms are required only if the model parameters cannot be integrated out. If the marginal likelihood of each model can be evaluated exactly or at least approximated, more efficient stochastic search methods that do not attempt to sample from the posterior distribution over all candidate models can be developed. A very good example is Yeung et al. (2005) who develop a multi-class classification method by introducing a stochastic search algorithm called *iterative Bayesian model averaging*. While this method performs very well in the context of gene selection in microarray studies, it is still based on an univariate ordering of the candidate predictors. Hans et al. (2007) make another step forward and propose the *shotgun stochastic search* (SSS) algorithm that is capable of quickly moving towards high-probable models while evaluating and recording complete neighborhoods around the current most promising models.

The aim of this chapter is to describe the use of the Mode Oriented Stochastic Search (MOSS, henceforth) algorithm of Dobra and Massam (2010) to perform variable selection in regression models involving discrete covariates. We further employ MOSS to determine the most probable hierarchical log-linear models associated with the small subsets of regressors identified at the variable selection step. As such, our proposed methodology represents a unified framework of the variable selection MCMC methods of Lunn et al. (2006) and of the graphical models selection approach of Verzilli et al. (2006). However, our model determination step is not restricted to a subset of decomposable log-linear models as proposed in Verzilli et al. (2006). The use of MOSS makes our methods scale to genomewide data due to its abilility to rapidly reach regions of high posterior probability in the target models space.

Following the ideas in Pittman et al. (2004) and Hu et al. (2009), we transform the observed data in multi-way contingency tables. While there is an inherent loss of information associated with our proposed discretization, we are able to treat covariates coming from various sources (e.g., expression levels, SNPs or clinical data) in a coherent manner. In addition, we reduce the influence of the particular choice of the normalization method employed for the initial pre-processing of the available information.

The structure of this chapter is as follows. In Section 2 we recall MOSS introduced by Dobra and Massam (2010) in the context of hierarchical log-linear models. We show the connections between MOSS and SSS, and describe how to use MOSS to perform small subsets regression selection. In Section 3 we briefly discuss the conjugate priors for log-linear parameters from Massam et al. (2009) and give the formula for the marginal likelihood of a regression involving the response variable and a reduced number of predictors. In Section 4 we show how to transform the observed data into a multi-way contingency table. In Section 5 we describe our Bayesian model averaging algorithm for performing variable selection and for developing efficient multi-class classifiers. Sections 6, 7 and 8 illustrate the use of our proposed methods for simulated datasets, expression data and SNP data, respectively. In Section 9 we make some concluding remarks.

2 MOSS

We denote by (n) the available data. The Bayesian paradigm to model determination involves choosing models m with high posterior probability $\Pr(m|(n))$ selected from a set $\mathcal{M}$ of competing models. We associate with each candidate model $m \in \mathcal{M}$ a neighborhood $\text{nbd}(m) \subset \mathcal{M}$. Any two models $m, m' \in \mathcal{M}$ are connected through at least a path $m = m_1, m_2, \ldots, m_l = m'$ such that $m_j \in \text{nbd}(m_{j-1})$ for $j = 2, \ldots, l$. The neighborhoods are defined with respect to the class of models considered. There are two types of models of interest for the development of our framework:

(i) Regressions. As described in Hans et al. (2007), the neighborhood of a regression m containing k regressors is obtained by: (i) addition moves: individually including in m any variable that is not in m. The resulting neighbors contain $k + 1$ regressors; (ii) deletion moves: individually deleting from m any variable that belongs to m. The resulting neighbors contain $k - 1$ regressors; and (iii) replacement moves: replacing any one variable in m with any one variable that is not in m. The resulting neighbors contain k regressors. Replacement moves are extremely important especially if we restrict the search to models having at most k' regressors, where k' is small (say, at most 5) with respect to the total number of available predictors. Since a variable cannot be added to a regression with k' predictors, a stochastic search algorithm would have to move to a smaller dimensional model before being able to consider a new covariate for inclusion in the current model. Here we assumed that no interactions among two or more predictors are considered for inclusion in a regression model. Since the search involves only main effects, each term can be included or removed independently. The removal of a main effect involved in an interaction requires the deletion of the corresponding interaction term, hence other types of moves would have to be developed – see Hans et al. (2007).

(ii) Hierarchical log-linear models. The neighborhood of a hierarchical model m consists of those hierarchical models obtained from m by adding one of its dual generators (i.e., minimal interaction terms not present in the model) or deleting one of its generators (i.e., maximal interaction terms present in the model). For details see Edwards and Havranek

(1985) and Dellaportas and Forster (1999).

The MC^3 algorithm proposed by Madigan and York (1995) constructs an irreducible Markov chain m_t, $t = 1, 2, \ldots$ with state space $\mathcal{M}$ and equilibrium distribution $\{\Pr(m|(n)) : m \in \mathcal{M}\}$. If the chain is in state m_t at time t, a candidate model m' is drawn from a uniform distribution on $\text{nbd}(m_t)$. The chain moves in state m' at time $t+1$, i.e. $m_{t+1} = m'$ with probability

$$\min\left\{1, \frac{\Pr(m_{t+1}|(n))/\#\text{nbd}(m_{t+1})}{\Pr(m_t|(n))/\#\text{nbd}(m_t)}\right\}, \tag{1}$$

where $\#\text{nbd}(m)$ denotes the number of neighbors of m. Otherwise the chain does not move, i.e. we set $m_{t+1} = m_t$.

Jones et al. (2005) and Hans et al. (2007) build on the MC^3 approach to construct the shotgun stochastic search (SSS) algorithm specifically designed to explore regions of high posterior probability in the candidate model space. SSS originates from earlier ideas proposed in Dobra et al. (2004). It visits the space by evaluating the posterior probability $\Pr(m'|(n))$ of each model m' in the neighborhood of the current model m and by choosing a new current model by sampling from these candidates with probabilities proportional with $\Pr(m'|(n))$. The efficiency of SSS comes from the fact that it focuses on rapidly moving towards models that maximize $\Pr(m|(n))$, $m \in \mathcal{M}$. As opposed to SSS, MC^3 might spend most of its time exploring models with low posterior probability if there are many such models and their total posterior probability dominates the space. This is precisely what happens when the sample size remains fixed while the number of candidate models increases rapidly which leads to a high degree of model uncertainty.

Dobra and Massam (2010) define the mode oriented stochastic search (MOSS, henceforth) algorithm that identifies models in

$$\mathcal{M}(c) = \left\{m \in \mathcal{M} : \Pr(m|(n)) \geq c \cdot \max_{m' \in \mathcal{M}} \Pr(m'|(n))\right\}, \tag{2}$$

where $c \in (0, 1)$. As proposed in Madigan and Raftery (1994), models with a low posterior probability compared to the highest posterior probability model are discarded. This choice drastically reduces the size of the target space from $\mathcal{M}$ to $\mathcal{M}(c)$. For suitable choices of c, $\mathcal{M}(c)$ can be exhaustively enumerated.

MOSS makes use of a current list $\mathcal{S}$ of models that is updated during the search. Define the subset $\mathcal{S}(c)$ of $\mathcal{S}$ in the same way we defined $\mathcal{M}(c)$ based on $\mathcal{M}$. Define $\mathcal{S}(c')$ with $0 < c' < c$ so that $\mathcal{S}(c) \subset \mathcal{S}(c')$. Let q be the probability of pruning the models in $\mathcal{S} \setminus \mathcal{S}(c)$. A model m is called *explored* if all its neighbors $m' \in \text{nbd}(m)$ have been visited. A model in $\mathcal{S}$ can be explored or unexplored. MOSS proceeds as follows:

`PROCEDURE MOSS(`c`,`c'`,`q`)`

(a) Initialize the starting list of models $\mathcal{S}$. For each model $m \in \mathcal{S}$, calculate and record its posterior probability $\Pr(m|(n))$. Mark m as unexplored.

(b) Let $\mathcal{L}$ be the set of unexplored models in $\mathcal{S}$. Sample a model $m \in \mathcal{L}$ according to probabilities proportional with $\Pr(m|(n))$ normalized within $\mathcal{L}$. Mark m as explored.

(c) For each $m' \in \text{nbd}(m)$, check if m' is currently in $\mathcal{S}$. If it is not, evaluate and record its posterior probability $\Pr(m'|(n))$. If $m' \in \mathcal{S}(c')$, include m' in $\mathcal{S}$ and mark m' as unexplored. If m' is the model with the highest posterior probability in $\mathcal{S}$, eliminate from $\mathcal{S}$ the models in $\mathcal{S} \setminus \mathcal{S}(c')$.

(d) With probability q, eliminate from $\mathcal{S}$ the models in $\mathcal{S} \setminus \mathcal{S}(c)$.

(e) If all the models in $\mathcal{S}$ are explored, eliminate from $\mathcal{S}$ the models in $\mathcal{S} \setminus \mathcal{S}(c)$ and `STOP`. Otherwise go back to step (b).

`END`.

At step (c) the models $m' \in \text{nbd}(m)$ can be considered in any possible order. As opposed to MC^3 or SSS, MOSS edobra623652nds by itself without having to specify a maximum number of iterations to run. In a MOSS search, the model explored at each iteration is selected from the most promising models identified so far. In MC^3 or SSS, this model is selected from the neighbors of the model evaluated at the previous iteration. This feature allows MOSS to move faster towards regions of high posterior probability in the models space.

Choosing c in the intervals $(0, 0.01]$, $(0.01, 0.1]$, $(0.1, 1/3.2]$, $(1/3.2, 1]$ means eliminating models having decisive, strong, substantial or "not worth more than a bare mention" evidence against them with respect to the highest posterior probability model – see Kass and Raftery (1995). We recommend using a value for c' as close to zero as possible. The role of the parameter c' is to limit the number of models that are included in $\mathcal{S}$ to a manageable number. We also suggest running the algorithm with several choices of c, c' and q to determine the sensitivity of the set of models selected.

3 Conjugate Priors for Hierarchical Log-linear Models

Massam et al. (2009) developed and studied the conjugate prior as defined by Diaconis and Ylvisaker (1979) (henceforth abbreviated the DY conjugate prior) for the log-linear parameters for the general class a discrete hierarchical log-linear models. We outline the notation and most relevant results from Massam et al. (2009). We also give a formula for the marginal likelihood of a regression model induced by these conjugate priors.

3.1 Model Parameterization

Let V be the set of criteria defining the contingency table. Denote the power set of V by $\mathcal{E}$ and take $\mathcal{E}_\ominus = \mathcal{E} \setminus \{\emptyset\}$. Let $X = (X_\gamma,\ \gamma \in V)$ such that X_γ takes its values (or levels) in the finite set I_γ of dimension $|I_\gamma|$. When a fixed number of individuals are classified according to

the $|V|$ criteria, the data is collected in a contingency table with cells indexed by combination of levels for the $|V|$ variables. We adopt the notation of Lauritzen (1996) and denote a cell by $i = (i_\gamma,\ \gamma \in V) \in \mathcal{I} = \times_{\gamma \in V} \mathcal{I}_\gamma$. The count in cell i is denoted $n(i)$ and the probability of an individual falling in cell i is denoted $p(i)$.

For $E \subset V$, cells in the E-marginal table are denoted $i_E \in \mathcal{I}_E = \times_{\gamma \in E} \mathcal{I}_\gamma$. The marginal counts are denoted $n(i_E)$. For $N = \sum_{i \in \mathcal{I}} n(i),\ (n) = (n(i),\ i \in \mathcal{I})$ follows a multinomial $\mathcal{M}(N, p(i),\ i \in \mathcal{I})$ distribution with probability density function

$$P\Big((n)\Big) = \left(\begin{array}{c} N \\ (n) \end{array} \right) \prod_{i \in \mathcal{I}} p(i)^{n(i)} \ . \tag{3}$$

Let i^* be a fixed but arbitrary cell that we take to be the cell indexed by the "lowest levels" of each factor. We denote these lowest levels by 0. Therefore i^* can be thought to be the cell $i^* = (0, 0, \ldots, 0)$. We define the log-linear parameters to be

$$\theta(i_E) = \sum_{F \subseteq E} (-1)^{|E \setminus F|} \log p(i_F, i^*_{F^c}) \tag{4}$$

which, by the Moebius inversion, is equivalent to

$$p(i_E, i^*_{E^c}) = \exp \sum_{F \subseteq E} \theta(i_F) \ . \tag{5}$$

Remark that $\theta_\emptyset(i) = \log p(i^*),\ i \in \mathcal{I}$. We denote $\theta_\emptyset(i^*) = \theta_\emptyset$ and $p(i^*) = p_\emptyset = \exp \theta_\emptyset$. It is easy to see that the following lemma holds.

Lemma 1 *If for $\gamma \in E, E \subseteq V$ we have $i_\gamma = i^*_\gamma = 0$, then $\theta(i_E) = 0$.*

3.2 The Multinomial for Hierarchical Log-linear Models

Consider the hierarchical log-linear model m generated by the class $\mathcal{A} = \{A_1, \ldots, A_k\}$ of subsets of V, which, without loss of generality, can be assumed to be maximal with respect to inclusion. We write $\mathcal{D} = \{E \subseteq_\ominus A_i, i = 1, \ldots, k\}$ for the indexing set of all possible interactions in the model, including the main effects. It follows from the theory of log-linear models (for example, see Darroch and Speed (1983)) and from Lemma 1 that the following constraints hold:

$$\theta(i_E) = 0, \quad E \notin \mathcal{D} \tag{6}$$

Therefore, in this case, for $i_E \in \mathcal{I}^*_E$, (5) becomes

$$\log p(i_E, i^*_{E^c}) = \log p(i(E)) = \theta_\emptyset + \sum_{F \subseteq E, F \in \mathcal{D}, i_F \in \mathcal{I}^*_F} \theta(i_F) \ . \tag{7}$$

and after the change of variable $(n) = (n(i), i \in \mathcal{I}^*) \mapsto (n(i_E), E \in \mathcal{E}_\ominus)$, the multinomial distribution for the hierarchical log-linear model becomes the distribution for the random

variable $Y = (n(i_D), D \in \mathcal{D}, i_D \in \mathcal{I}^*_D)$ with density

$$f_{\mathcal{D}}(y;\theta) \propto \exp\left\{\sum_{D\in\mathcal{D}}\sum_{i_D\in\mathcal{I}^*_D}\theta(i_D)n(i_D) - N\log\left(1+\sum_{E\in\mathcal{E}_\ominus, i_E\in\mathcal{I}^*_E}\exp\sum_{F\subseteq_{\mathcal{D}}E}\theta(i_F)\right)\right\} \tag{8}$$

with respect to a measure $\mu_\infty(y)$. It is important to note here that

$$\theta_{\mathcal{D}} = (\theta(i_D),\ D \in \mathcal{D},\ i_D \in \mathcal{I}^*_D) \tag{9}$$

is the canonical parameter and

$$p_{\mathcal{D}} = (p(i(D)),\ D \in \mathcal{D},\ i_D \in \mathcal{I}^*_D)\ , \tag{10}$$

is the cell probability parameter of the multinomial distribution of m. The other cell probabilities $p(i(E)), E \notin \mathcal{D}$ are not free and are a function of $p_{\mathcal{D}}$.

3.3 The Diaconis-Ylvisaker Conjugate Prior

The natural exponential family form of the distribution of the marginal counts $Y = (n(i_E), E \in \mathcal{D}, i_E \in \mathcal{I}^*_E)$ of a contingency table with cell counts $n(i), i \in \mathcal{I}$ is given in Equation (8). The conjugate prior for θ as introduced by Diaconis and Ylvisaker (1979) is

$$\pi_{\mathcal{D}}(\theta_{\mathcal{D}}|s,\alpha) = I_{\mathcal{D}}(s,\alpha)^{-1}h(\theta_{\mathcal{D}}), \tag{11}$$

where $I_{\mathcal{D}}(s,\alpha) = \int_{\mathbb{R}^{d_{\mathcal{D}}}} h(\theta_{\mathcal{D}})d\theta_{\mathcal{D}}$ is the normalizing constant of $\pi_{\mathcal{D}}(\theta_{\mathcal{D}}|s,\alpha)$, the dimension of the parameter space $d_{\mathcal{D}}$ is $\sum_{D\in\mathcal{D}}\prod_{\gamma\in D}(|I_\gamma|-1)$ and

$$h(\theta_{\mathcal{D}}) = \exp\left\{\sum_{D\in\mathcal{D}}\sum_{i_D\in\mathcal{I}^*_D}\theta(i_D)s(i_D)\right. \tag{12}$$

$$\left. -\alpha\log\left(1+\sum_{E\in\mathcal{E}_\ominus}\sum_{i_E\in\mathcal{I}^*_E}\exp\sum_{F\subseteq_{\mathcal{D}}E}\theta(i_F)\right)\right\}\ . \tag{13}$$

The corresponding hyper-parameters are:

$$(s,\alpha) = (s(i_D), D \in \mathcal{D}, i_D \in \mathcal{I}^*_D, \alpha),\quad s \in \mathbb{R}^{d_{\mathcal{D}}},\quad \alpha \in \mathbb{R}. \tag{14}$$

Massam et al. (2009) give a necessary and sufficient condition for the distribution in Equation (11) to be proper as well as two methods to choose hyper-parameters (s,α) such that $I_{\mathcal{D}}(s,\alpha) < +\infty$.

From the similarity between the form (8) of the distribution of the marginal cell counts Y and the form (11) of the prior on θ, one can think of the hyper-parameters s as the marginal cell entries of a fictive contingency table whose cells contain positive real numbers. Consequently, α can be taken to be the grand total of this fictive table. The lack of prior information can be expressed through a non-informative prior specified by taking all the fictive cell entries to be equal to $\frac{\alpha}{|\mathcal{I}|}$ so that

$$s(i_D) = \sum_{j \in \mathcal{I}, j_D = i_D} \frac{\alpha}{|\mathcal{I}|} \, . \tag{15}$$

In the case of decomposable log-linear models, this approach to constructing a conjugate prior is equivalent to eliciting hyper-Dirichlet priors – see, for example, Dawid and Lauritzen (1993) and Madigan and York (1997). While the hyper-Dirichlet priors are restricted to decomposable log-linear models, the properties of the Diaconis-Ylvisaker conjugate priors extend naturally to graphical and hierarchical log-linear models.

Given the prior $\pi_{\mathcal{D}}(\theta_{\mathcal{D}}|s,\alpha)$, from the form (8) of the distribution of the marginal cell counts Y, the posterior distribution of $\theta_{\mathcal{D}}$ given $Y = y = (n(i_D), D \in \mathcal{D}, i_D \in \mathcal{I}^*_D)$ is

$$\begin{aligned} \pi_{\mathcal{D}}(\theta_{\mathcal{D}}|y,s,\alpha) &= \frac{1}{I_{\mathcal{D}}(s+y,\alpha+N)} \exp\left\{ \sum_{D \in \mathcal{D}} \sum_{i_D \in \mathcal{I}_{D^*}} \theta(i_D)(s(i_D)+n(i_D)) \right. \\ &\left. -(\alpha+N)\log\left(1 + \sum_{E \in \mathcal{E}_\ominus, i_E \in \mathcal{I}^*_E} \exp \sum_{F \subseteq_{\mathcal{D}} E} \theta(i_F)\right)\right\}. \end{aligned} \tag{16}$$

If we look at the posterior $\pi_{\mathcal{D}}(\theta_{\mathcal{D}}|y,s,\alpha)$ as a function of (s,α), $s(i_D)+n(i_D)$ represent the (i_D)-counts in the D-marginal entries of a table having the same structure as (n) and whose entries are obtained by augmenting the actual cell counts in (n) with the fictive cell counts from the prior. The grand total of this augmented table is $\alpha + N$.

3.4 Computing the Marginal Likelihood of a Regression

Let $Y = X_r$, $r \in V$ be a response variable and X_A, $A \subset V \setminus \{r\}$ are a set of explanatory variables. Denote by $(n)_{A\cup\{r\}}$ and $(n)_A$ the marginals of the full contingency table (n). Here (n), $(n)_{A\cup\{r\}}$ and $(n)_A$ are cross-classifications involving X_V, $X_{A\cup\{r\}}$ and X_A, respectively. Dobra and Massam (2010) explored the connection between log-linear models and the regressions derived from them. They expressed the regression parameters of the conditional $[Y|X_A]$ as a function of a log-linear model for Y and X_A. To assure the consistency of the distributions for regression parameters associated with various subsets of explanatory variables, we assume a saturated log-linear model for the full table (n). After collapsing across variables $X_{V\setminus(A\cup\{r\})}$, we obtain a saturated log-linear model for $(n)_{A\cup\{r\}}$. This means that the DY conjugate prior for the saturated log-linear model for (n) reduces to the DY conjugate prior for the saturated log-linear model for $(n)_{A\cup\{r\}}$ by setting some of the θ parameters to zero.

A similar statement can be made about the posterior distribution in Equation (16).

The properties shared by the DY priors and posteriors imply that the marginal likelihood of the regression $[Y|X_A]$ is the ratio between the marginal likelihoods of the saturated models for $(n)_{A\cup\{r\}}$ and $(n)_A$ – see Dawid and Lauritzen (1993); Geiger and Heckerman (2002); Dobra and Massam (2010):

$$\Pr(r|A) = \frac{I_{\mathcal{D}^{A\cup\{r\}}}((s+y)^{A\cup\{r\}}, \alpha+N)}{I_{\mathcal{D}^A}((s+y)^A, \alpha+N)} \cdot \frac{I_{\mathcal{D}^A}(s^A, \alpha)}{I_{\mathcal{D}^{A\cup\{r\}}}(s^{A\cup\{r\}}, \alpha)}. \quad (17)$$

Dobra and Massam (2010) show that, for any set $B \subset V$, we have

$$I_{\mathcal{D}^B}(s^B, \alpha) = \Gamma(\alpha^B_\emptyset) \prod_{D\in\mathcal{D}^B} \prod_{i_D\in\mathcal{I}^*_B} \Gamma(\alpha^B(i_D, i^*_{D^c})),$$

where

$$\alpha^B(i_D, i^*_B) = \sum_{B\supseteq F\supseteq D} \sum_{\substack{j_F\in\mathcal{I}^*_F \\ (j_F)_D=i_D}} (-1)^{|F\setminus D|} s(j_F),$$

$$\alpha^B_\emptyset = \alpha + \sum_{D\subseteq B} (-1)^{|D|} \sum_{i\in\mathcal{I}^*_D} s(i_D).$$

4 Discretizing Ordered Variables

The methods developed in this chapter perform well for datasets involving covariates of almost any type. The only constraints we impose on the nature of the observed covariates is that they have to be either categorical or their values can be ordered. Variables of the latter type are transformed in categorical variables as follows. Let $\{x_1, x_2, \ldots, x_n\}$ be the possible values of x. Denote by $\tau_1 < \tau_2 < \ldots < \tau_{k-1} < \tau_k = \infty$ a set of levels of x called *splits*. Here τ_1 is the minimum value of x and $k \geq 3$. We replace each value x_i, $1 \leq i \leq n$, by $j \in \{1, \ldots, k-1\}$ if $\tau_j \leq x_i < \tau_{j+1}$. This implies that x is transformed in a categorical covariate with $(k-1)$ levels. A sequential discretization of all the observed variables leads to a multi-way contingency table (n) with the same dimension as the original data. The table (n) is invariant to monotonic transformations of covariates – see Hu et al. (2009).

The splits used for discretization can be chosen either based on background information that might be available or can be defined to maximize the individual predictive ability of each covariate with respect to a categorical response variable y. Let x_τ be the categorical version of a predictor x, where τ are a set of splits for x. Following Pittman et al. (2004) and Hu et al. (2009), we use percentiles as a possible choice of splits for continuous covariates such as gene expression levels. A segregating SNP site has three possible genotypes: 0/0, 0/1 and 1/1, where 0 is the wild type and 1 is the mutant allele. Diallelic SNPs can be represented as three-category discrete variables, or can be dichotomized as presence of 0 vs. absence of 0, or as presence of 1 vs. absence of 1.

We choose the splits τ having the largest marginal likelihood in Equation (17) of the

regression of y on x_τ. Ideally we would like to consider combinations of choices of splits for all available predictors and choose those splits that lead to large values of the marginal likelihood of regressions of y on one, two or more categorical predictors. Each predictor will have several categorical versions associated with various splits sets. Unfortunately the datasets resulting from genomewide studies precludes us to perform such large scale computations, and we need to choose splits sequentially for each covariate.

The conjugate priors of Massam et al. (2009) are appropriate for the analysis of general multi-way tables in which each variable is allowed to have any number of categories. In order to increase the mean number of observations per cell we make the constraint that the data needs to be transformed in a dichotomous rather than a polychotomous contingency table. The response variable y is either dichotomous by design or has been dichotomized based on background information. Each predictor is replaced by a dichotomous version constructed with respect to $20, 21, \ldots, 80$-th percentiles. The smallest and largest percentile to be used as a split candidate is chosen such that at least one sample falls in each of the two categories. However, we emphasize that our theoretical framework can be employed to develop multi-class predictive models – see Yeung et al. (2005) for a related approach.

5 Bayesian Model Averaging

Let (n) be the multi-way contingency table obtained by discretizing the observed covariates as described in Section 4. As before, the variables cross-classified in (n) are X_γ, $\gamma \in V$. Denote by $Y = X_r$, $r \in V$, a response variable of interest. We use MOSS to identify regressions $[Y|X_A]$, $A \subset V \backslash \{r\}$ that maximize the marginal likelihood $\Pr(r|A)$ in Equation (17). We assume that all the models are apriori equally likely, so that the posterior probability of each regression is proportional with its marginal likelihood. However, we are only interested in regressions involving a relatively small number of predictors (less than five) because the available number of samples is usually small. One should evaluate the mean number of observations per cell in a marginal $(n)_{A\cup\{r\}}$, where A has the maximum number of predictors allowed. If the resulting mean is too low, the maximum number of predictors should be decreased.

MOSS identifies a set of regressions $\mathcal{S}$ for some $c \in (0, 1)$. The cutoff c is chosen so that $\mathcal{S} = \mathcal{S}(c)$ can be exhaustively enumerated. A regression $m_j \in \mathcal{S}(c)$ is $[Y|X_{A_j}]$ with $A_j \subset V \setminus \{r\}$ and $j \in B$. Here B is a set of indices. The regression of Y on the remaining variables is a weighted average of the regressions in $\mathcal{S}$, where the weights represent the posterior probability of each regression (see, for example, Yeung et al. (2005)):

$$\Pr(Y = y|(n)) = \sum_{j\in B} \Pr(Y = y|(n)_{A_j}) \cdot \Pr(m_j|(n)). \tag{18}$$

Since we assumed that all models are apriori equally likely, the posterior probability of each regression is equal to its marginal likelihood normalized over all the models in $\mathcal{S}$:

$$\Pr(m_j|(n)) \;=\; \frac{\Pr(r|A_j)}{\sum\limits_{l\in B} \Pr(r|A_l)}.$$

Madigan and Raftery (1994) show that the weighted average of regressions in (18) has a better predictive performance than any individual model in $\mathcal{S}$. The revelance of each predictor X_j can be quantified by its posterior inclusion probability defined as the sum of the posterior probabilities of all the models that include X_j.

There is an additional step we perform to assure the most parsimonious versions of regressions $[Y|A_j]$ are employed in Equation (18). Dobra and Massam (2010) developed a version of MOSS for hierarchical log-linear models with the DY conjugate priors. We use their stochastic search procedure to determine the hierarchical log-linear model with the highest posterior probability for each marginal $(n)_{A_j \cup \{r\}}$, where A_j identifies a regression in $\mathcal{S}$. Dobra and Massam (2010) also determine the regression induced by a hierarchical log-linear model for a given response variable Y. It is possible that some variables X_γ, $\gamma \in A_j$ interact with Y indirectly through other variables in A_j and in this case $[Y|X_{A_j}]$ reduces to $[Y|X_{E_j}]$ for $E_j \subset A_j$. Each regression $[Y|X_{A_j}]$ in Equation (18) is the regression induced by the highest posterior probability log-linear model for the marginal $(n)_{A_j \cup \{r\}}$. Since the same choice of conjugate priors is employed in the stochastic search for regressions and in the subsequent stochastic search for log-linear models, the overall consistency of our variable selection method coupled with log-linear model determination is guaranteed.

Dobra and Massam (2010) describe an algorithm called the Bayesian iterate proportional fitting (Bayesian IPF, henceforth) for sampling from the joint posterior in Equation (16) of parameters θ corresponding with a log-linear model. The resulting posterior samples can be used to estimate the coefficients of the regressions in $\mathcal{S}(c)$. The coefficients for the Bayesian model averaging regression $\Pr(Y = y|(n))$ from Equation (18) are estimated by sampling from the joint posterior of the coefficients for each individual regression $m_j = [Y|X_{A_j}]$, $j \in B$, for a number of iterations proportional with $\Pr(m_j|(n))$.

6 Simulated Examples

6.1 First Simulated Example

We apply the MOSS algorithm to perform model selection when the set of potential predictors exhibit strong pairwise correlations. This is a variation of the example suggested by Nott and Green (2004). As in George and McCulloch (1997), we generate $Z_1, \ldots, Z_{15}, Z \sim N_{300}(0, I)$, where $N_{300}(0, I)$ is the 300-dimensional normal distribution with zero mean and identity covariance matrix. Let $X_i = Z_i + 2Z$, $i = 1, 3, 5, 8, 9, 10, 12, 13, 14, 15$ and also $X_2 = X_1 + 0.15Z_2$, $X_4 = X_3 + 0.15Z_4$, $X_6 = X_5 + 0.15Z_6$, $X_7 = X_8 + X_9 - X_{10} + 0.15Z_7$ and $X_{11} = X_{14} + X_{15} - X_{12} - X_{13} + 0.15Z_{11}$. George and McCulloch (1997) point out that this design matrix leads to correlations of about 0.998 between X_i and X_{i+1} for $i = 1, 3, 5$. There are also strong linear associations between (X_7, X_8, X_9, X_{10}) and $(X_{11}, X_{12}, X_{13}, X_{14}, X_{15})$. We let $\tilde{X} = [X^{(1)}X^{(2)}]$ be a 300×30 design matrix obtained by independently simulating two instances $X^{(1)}$ and $X^{(2)}$ of the 300×15 design matrix X.

Consider the 30-dimensional vector of regression coefficients β defined by

$$\beta_j = \begin{cases} 1.5, & \text{if } j = 1, 3, 5, 7, 11, 12, 13, \\ -1.5, & \text{if } j = 8, \\ 0, & \text{otherwise.} \end{cases}$$

We generate $Y = \tilde{X}\beta$. The binary response $\tilde{Y}$ is obtained as $\tilde{Y}_j = 1$ if $Y_j \geq 0$ and $\tilde{Y}_j = 0$ if $Y_j < 0$.

We would like to study whether predictors that belong to $X^{(2)}$ are not selected by MOSS. Note that some predictors in $X^{(1)}$ might still be selected even if their regression coefficients are zero because of the complex correlation structure that exists in this block.

We employed MOSS with $c = 0.333$, $c' = 0.0001$, a pruning probability of 0.25 and five different starting points. We allowed MOSS to explore regressions containing at most five predictors. In order to reduce the sample variability, we report the results we obtained by averaging across 100 replicates of this experiment. The percentage of selected predictors that belong to the block $X^{(1)}$ is 100% with a 95% confidence interval of $[50\%, 100\%]$. This implies that MOSS almost never selects predictors from $X^{(2)}$. The mean number of predictors selected is 2 with a 95% confidence interval of $[1, 9]$. The mean number of regressions evaluated by each instance of MOSS was 1403 with a 95% confidence interval of $[957.2, 2615.2]$. The number of possible regressions with 5 predictors or less is 174437 which is indicative of how fast our proposed stochastic search method can move towards models with high posterior probability.

We employed Bayesian model averaging to construct classifiers for $\tilde{Y}$ based on the regressions in the sets $\mathcal{S}(0.333)$. The mean area below the corresponding ROC curves was 0.996 with a 95% confidence interval of $[0.979, 1]$. Therefore MOSS can construct excellent classifiers by identifying predictors that are actually related to the binary response.

6.2 Second Simulated Example

The second example was suggested by George and McCulloch (1993). We generate $Z_1, \ldots, Z_{60}, Z \sim N_{120}(0, I)$ and construct the 120×60 design matrix $X = (X_1, \ldots, X_{60})$ with $X_i = Z_i + Z$ for $i = 1, \ldots, 60$. The average pairwise correlations in X is about 0.5 which makes variable selection difficult. Consider the 60-dimensional vector of regression coefficients β given by $\beta_j = 0$ if $j = 1, \ldots, 15$, $\beta_j = 1$ if $j = 16, \ldots, 30$, $\beta_j = 2$ if $j = 31, \ldots, 45$ and $\beta_j = 3$ if $j = 46, \ldots, 60$. We take $Y = \tilde{X}\beta$. The binary response $\tilde{Y}$ is obtained as $\tilde{Y}_j = 1$ if $Y_j \geq 0$ and $\tilde{Y}_j = 0$ if $Y_j < 0$. Here the goal is to see if we do not select the first 15 variables.

We run MOSS with the same choice of parameters as in the first simulated example and replicate the experiment 100 times to reduce the sample variability of the results. The mean percentage of predictors that are among the first 15 variables is 20% with a 95% confidence interval of $[0\%, 40\%]$. This is very close to 25% that represents the actual percentage of "unwanted" predictors in X. We can explain these results by the large correlations among the candidate predictors and by the loss of information that occurs during dichotomization of the response variable and of the predictors themselves. The mean number of predictors

selected is 4 with a 95% confidence interval of $[1, 20]$. The mean number of regressions evaluated by each instance of MOSS was 3681 with a 95% confidence interval of $[1841, 7370]$. This is much smaller than 5985198 – the number of all possible regressions with 5 variables or less. Again, this indicates the computational efficiency achieved by MOSS. The Bayesian model averaging classifiers for $\tilde{Y}$ have excellent predictive accuracy. The mean area below the corresponding ROC curves was 1 with a 95% confidence interval of $[0.997, 1]$.

7 Real Examples: Gene Expression

7.1 Breast Cancer Data

We analyze the breast cancer prognosis dataset from van't Veer et al. (2002). Here the goal is to develop a gene expression classifier to predict which patients are likely to develop metastases within 5 years. Yeung et al. (2005) identified 4919 significantly regulated genes in the training set of 76 samples. The test set comprises 19 samples. van't Veer et al. (2002) select 70 genes based on their high correlation with the response and report that only two samples in the test set were incorrectly classified based on the expression levels of these genes. Yeung et al. (2005) used Bayesian model averaging to produce a classifier that involves only 6 genes. Their predictive model gives 3 classification errors on the test set.

We employed MOSS with $c = 0.5$, $c' = 0.0001$, a pruning probability of 0.1 and five different starting points. We searched for regression models containing at most 3 predictors. The number of models evaluated by MOSS in each of the five instances were: 368779, 368781, 486787, 565463, and 206521. These counts are very small compared to the total number of possible regressions 19837151360.

Using the ordering of the genes of Yeung et al. (2005), we identify seven regressions in the set $\mathcal{S}(0.5)$. These regressions involve 11 genes as follows: TSPYL5 (1), NM_021199 (0.88), Contig31010_RC (0.44), Contig16367_RC (0.22), AA555029.RC (0.1), IFIT3 (0.07), PDIA4 (0.07), Contig1829 (0.06), CASC3 (0.06), GDS1048 (0.05), NUP210 (0.05). The numbers of parentheses represent the posterior inclusion probabilities of each gene. Remark that only gene TSPYL5 was also selected by van't Veer et al. (2002) and Yeung et al. (2005) and is the gene with the highest posterior probability in our ranking. The remaining 10 genes do not appear in the list of Yeung et al. (2005). We were not able to determine the actual overlap between our list and the candidate genes of van't Veer et al. (2002). It is likely that this overlap is empty because van't Veer et al. (2002) select genes based on a univariate dependency measure (correlation), while we take into account combinations of at most three genes and the response. Yeung et al. (2005) consider similar combinations in their stochastic search algorithm.

Figure 1 shows the performance of the classifier we produced by employing MOSS for hierarchical log-linear models to determine the most relevant models for the 2^4 contingency tables associated with the binary response and the dichotomized expression levels of the genes in each regression. We chose $c = 0.01$, $c' = 0.0001$, a pruning probability of 0.1 and five starting points. We generated 10000 samples from the posterior distributions of the

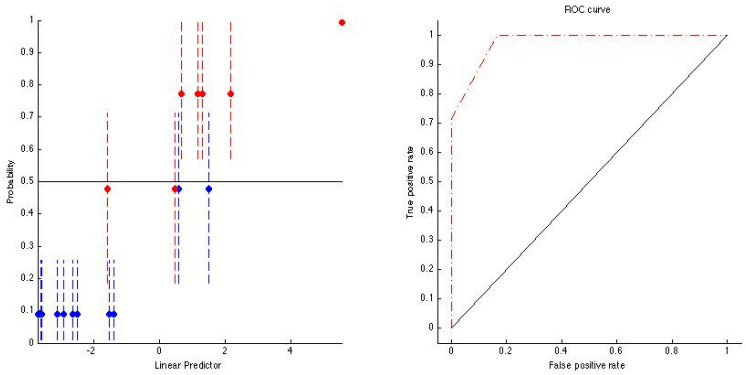

Figure 1: The left panel shows the performance of the 11 gene classifier of disease-free patients after at least 5 years (blue) versus patients with metastases within 5 years (red). The horizontal lines represent 80% confidence intervals. The right panel shows the ROC curve associated with this classifier.

induced logistic regression parameters using Bayesian IPF. These samples are necessary for parameter estimation and to quantify prediction uncertainty. Only two test samples are misclassified in Figure 1. The area below the ROC curve in the right panel of Figure 1 is 0.98 which indicates an excellent predictive performance. The Brier score is 1.73 with a standard deviation of 0.19. For comparison, the Brier Score reported by Yeung et al. (2005) is 2.04. Smaller values of the Brier score indicate a better performance.

7.2 Leukemia Data

The leukemia dataset of Golub et al. (1999) comprise samples from patients with acute lymphoblastic leukemia (ALL) or acute myeloid leukemia (AML). The initial pre-processing leaves the expression levels of 3051 genes on 38 training samples and 34 test samples – see Yeung et al. (2005). We employed MOSS with $c = 0.5$, $c' = 0.0001$, a pruning probability of 0.1 and five different starting points. We searched for regression models containing at most 2 predictors which gives a total number of possible regressions of 4655827. The number of models evaluated by MOSS in each of the five instances were considerably smaller: 115902, 323301, 189102, 176901, 79302. The six regressions in the resulting $\mathcal{S}(0.5)$ involve eight genes: MGST1 (0.67), APLP2 (0.33), CCND3 (0.17), TRAC (0.17), NCAPD2 (0.17), ACADM (0.17), MAX (0.17) and PSME1 (0.16). As before, the posterior inclusion probabilities are indicated in parentheses.

We used MOSS to identify relevant hierarchical log-linear models for the corresponding 2^3 contingency tables. We chose $c = 0.01$, $c' = 0.0001$, a pruning probability of 0.1 and five starting points. We generated 10000 samples from the posterior distributions of the induced logistic regression parameters using Bayesian IPF. Figure 2 shows the prediction results for the test data of the classifier obtained by Bayesian model averaging of the six

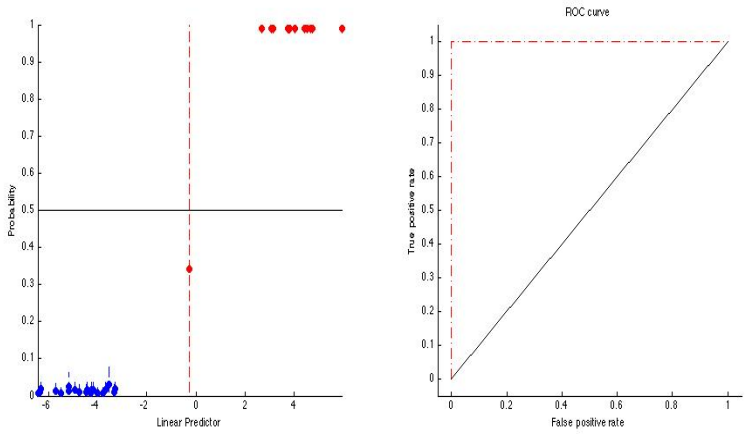

Figure 2: The left panel shows the performance of the 8 gene classifier for the 34 test samples in leukemia data. The ALL samples are coded in blue, while the AML samples are coded in red. The horizontal lines represent 80% confidence intervals. The right panel shows the ROC curve associated with this classifier.

logistic regressions. Only one sample is incorrectly predicted. The Brier score is 0.7 with a standard error of 0.48. The area below the ROC curve is almost 1. Yeung et al. (2005) reports two classification erros based on 20 selected genes and a Brier score of 1.5. Lee et al. (2003) misclassifies one sample based on the expression levels of five genes, while Nguyen and Rocke (2002) reports $1-3$ misclassified samples based on $50-1500$ genes.

7.3 Lymph Node Data

We predict lymph node positivity status in human breast cancer based on the expression levels of 4512 genes – see Hans et al. (2007); Pittman et al. (2004). There are 100 low-risk (node-negative) samples and 48 high-risk (high node-positive). There are two additional predictors: estimated tumor size (in centimeters) and estrogen receptor status (binary variable determined by protein assays).

We employed MOSS with $c = 0.01$, $c' = 0.0001$, a pruning probability of 0.1 and five different starting points. We searched for regression models containing at most 5 predictors due to the larger number of available samples. This choice corresponds with a mean number of samples per cell in the corresponding 2^6 contingency tables of 2.31. The number of models evaluated by MOSS in each of the five instances (2896078, 1109722, 2057022 and 721769) are significantly smaller than the total number of possible regressions $1.56 \cdot 10^{16}$. MOSS identifies 49 regressions in $\mathcal{S}(0.01)$. These regressions involve 69 predictors, but only eight have posterior inclusion probabilities greater than 0.1: ENTPD4 (0.89), tumor size (0.62), ADD3 (0.34), MGLL (0.32), ST6GALNAC2 (0.32), HEXA (0.18), MEGF9 (0.11) and FAM38A (0.1). Remark that tumor size is identified as the second most important predictor. The tree models of Pittman et al. (2004) also found tumor size to be one of the most relevant

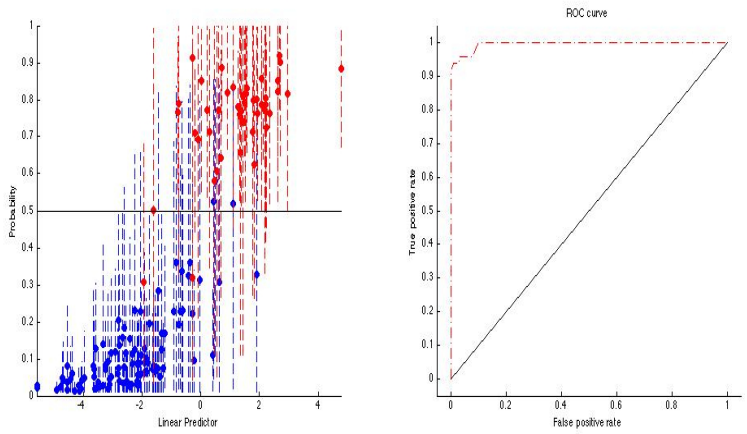

Figure 3: Fitted prediction probabilities for the lymph node data. The node-negative samples are blue, while the node-positive samples are red. The horizontal lines represent 80% confidence intervals. The right panel shows the ROC curve associated with this classifier.

predictors, in contrast with the logistic regressions of Hans et al. (2007) that place a different set of eight genes at the top of their list.

We used MOSS to identify relevant hierarchical log-linear models for the 2^6 contingency tables corresponding with the 49 regressions. We chose $c = 0.01$, $c' = 0.0001$, a pruning probability of 0.1 and five starting points. We generated 10000 samples from the posterior distributions of the induced logistic regression parameters using Bayesian IPF. Figures 3 and 4 show the excellent predictive performance of our classifier. For the fitted values, the area below the ROC curve is almost one, while the Brier score is 11.68 with a standard deviation of 2.27. Five samples were incorrectly predicted: 97% of the positives are above 0.5, and 95.8% of the negatives are below 0.5. For comparison, Hans et al. (2007) evaluate the fit of their model to 96% of the positives above 0.5 and 89% of the negatives below 0.5. We also performed a five-fold cross-validation check of our models. The area below the ROC curve is 0.98, the Brier score increases slightly to 16.59 with a standard deviation of 2.64. In this case six samples were incorrectly predicted which gives a prediction accuracy of almost 96%.

8 Real Examples: Genome-wide Analysis of Estrogen Response with Dense SNP Array Data

A long-term goal of pharmacogenetics research is the accurate prediction of patient response to drugs, as it would facilitate the individualization of patient treatment. Such an approach is particularly needed in cancer therapy, where currently used agents are ineffective in many patients, and side effects are common. The recent development of genome-wide approaches such as high-density SNP arrays enables the simultaneous measurement of thousands of

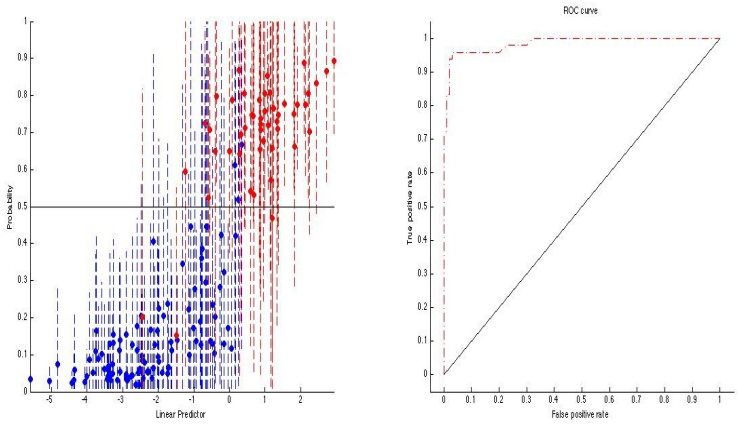

Figure 4: Five-fold cross-validation prediction probabilities for the lymph node data. The node-negative samples are blue, while the node-positive samples are red. The horizontal lines represent 80% confidence intervals. The right panel shows the ROC curve associated with this classifier.

genes in a single experiment and raises the possibility of an unbiased genome-wide approach to the genetic basis of drug response. To facilitate pharmacogenetics research, a panel of 60 cancer cell lines has been used extensively by the Development Therapeutics Program of the National Cancer Institute (http://dtp.nci.nih.gov). These cell lines have been analyzed for their sensitivity and resistance to a broad range of chemical compounds and thus offer an extensive source of information for testing prediction models of chemotherapy response.

Jarjanazi et al. (2008) focused on the chemosensitivity and resistance to estrogen response (one of the several compounds), which is extremely important in the treatment of breast cancer by hormonal therapy. Determining a genetic signature of estrogen response could help to tailor treatment programs based on patient's genetic background and thus reduce considerably the chance of treatment failure. Each cell line was exposed to estrogen for 48 hours, and the concentration of estrogen required for 50% growth inhibition (GI50 henceforth) was scored. The $\log_{10}$ of the GI50 values were retrieved and normalized to obtain a mean of zero and standard deviation of one across the different cell lines. Using a density estimation of the GI50 values (Wand and Jones (1995)), we labeled 25 cell lines as resistant and 17 cell lines as sensitive. Drug sensitivity tests for estrogen were performed. Genotypes of SNPs in these 42 cell lines were obtained from the Affymetrix 125K chip data – see Garraway et al. (2005) for more details. We retained only 25530 SNPs that were genotyped in at least 90% of the cell lines and had a minimum allele frequency (MAF) of at least 10%.

We employed MOSS with $c = 0.5$, $c' = 0.0001$, a pruning probability of 0.1 and a maximum number of predictors equal to three. We run MOSS one hundred times starting from random regressions. The mean number of models evaluated by MOSS was 2407299, while the minimum and maximum number of models evaluated were 255288 and 9853813, respectively. These counts are very small compared to $2.77 \cdot 10^{12}$ – the total number of regressions with at

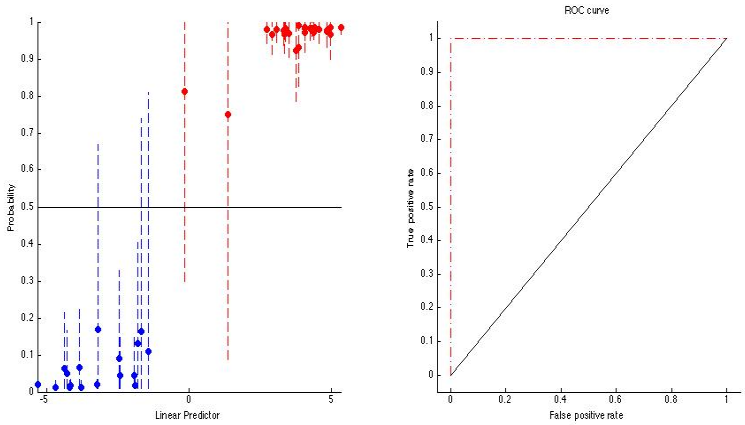

Figure 5: Fitted prediction probabilities for the SNP data when regressions with at most three variables are considered. The control samples are blue, while treatment samples are red. The horizontal lines represent 80% confidence intervals. The right panel shows the ROC curve associated with this classifier.

most three variables. Seven regressions were identified in $\mathcal{S}(0.5)$. These regressions involve 17 SNPs as follows: 1596 (0.39), 8823 (0.33), 5545 (0.33), 7357 (0.33), 14912 (0.25), 14303 (0.25), 7459 (0.18), 11684 (0.13), 23629 (0.12), 4062 (0.12), 9571 (0.12), 394 (0.11), 19083 (0.11), 19132 (0.06), 11103 (0.06), 17403 (0.06), 11775 (0.06). The first SNP is indexed with 1, while the last SNP is indexed with 25530.

Figure 5 shows the fitted values of the model-averaged classifier. The area below the ROC curve is one, the Brier score is 0.993 with a standard error of 0.35. No samples were incorrectly predicted. Figure 6 shows the two-fold cross-validation results of this classifier. The Brier score is 2.41 with a standard error of 1.67 and all samples are correctly predicted.

The description of the interesting SNPs found by MOSS and their physical distance to known genes are summarized in Table 1. Three of these SNPs are located within a known gene, whereas nine were very close to a known gene locus (<150Kb). We also studied the extent of linkage disequilibrium (LD) for the three SNPs that were located within less than 30kb from a known candidate gene region using the software Haploview – see Barrett et al. (2005). In these three cases, we found evidence for LD, suggesting that the associated gene could be implicated in estrogen response. Details about the LD results are given in Tables 2, 3 and 4.

Among the genes identified by MOSS, **Thrombospondin-1 (THBS1)** (also referred to as THBS, TSP, TSP1) has been the most studied in terms of function and association to disease. **THBS1** is a glycoprotein that is involved in various biological processes including the cell motility, cell adhesion, inflammatory response and multicellular organismal development (Go ontology). **THBS1** regulates the activity of the estrogen in the cell – see Sarkar et al. (2007); Sengupta et al. (2004); Slater et al. (1995). **THBS1** is also extensively implicated in cancer and metastasis in several reviews. Although functionally characterized, the asso-

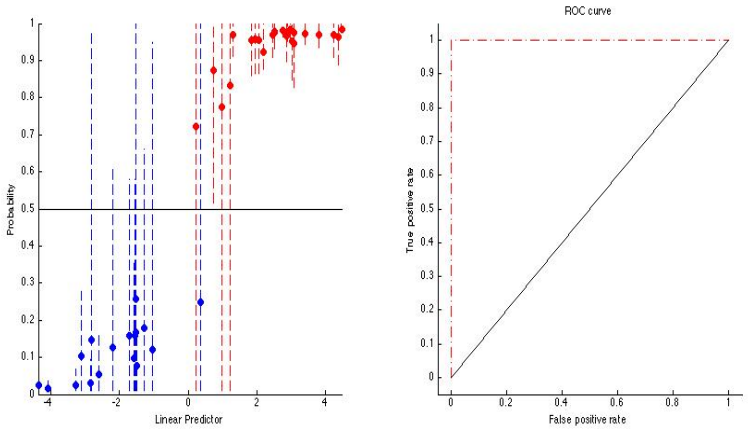

Figure 6: Two-fold cross-validation prediction probabilities for the SNP data when regressions with at most three variables are considered. The control samples are blue, while treatment samples are red. The horizontal lines represent 80% confidence intervals. The right panel shows the ROC curve associated with this classifier.

ciation of the two other candidate genes (**BLNK** and **NSF**) with estrogen response is not well studied. The B cell linker protein **BLNK** at 10q24.1 is a SH2/SH3 containing adaptor protein which serves as a scaffold to assemble downstream targets of antigen activation. It is involved in the regulation of B cell receptor signaling during inflammatory and humoral immune response of the cell. It has been shown to be associated with leukemia, lymphoma and solid tumors. **N-ethylmaleimide-sensitive factor (NSF)** at 17q21 is required for vesicle-mediated transport, from the endoplasmic reticulum to the Golgi stack. It is also involved in proteolysis. The functions of the transmembrane protein 26 (TMEM26) at 10q21.2 and other candidate genes in our study are not well studied.

While the identification of **THBS1** has demonstrated that our method is capable of identifying a gene that is extensively involved in estrogen response and cancer development, the remaining genes with limited knowledge represent novel candidates to be further studied. Several important remarks about the application of MOSS to this study are worth noting. First, **THBS1**, the most functionally relevant SNP identified here, would not have been detected using an exhaustive testing of all single markers. Indeed, the associated p-value after correction for multiple testing is clearly non significant (p=0.35). Second, the proportion of markers that are biological relevant (i.e. within or close to known genes) is relatively large. This is somewhat unusual compared to other GWAS that have found a substantial fraction of associations in regions that do not contain annotated genes or that are outside of transcriptional units of the genome – see Altshuler and Daly (2007). Third, the predictive ability of the selected subset of SNPs on estrogen response is almost perfect, which implies very promising applications of MOSS in predictive genetic testing. Finally, MOSS provides a unique and comprehensive statistical framework for analyzing GWAS, filling an important gap in the search for combinations of SNPs that have high predictive values in

Table 1: Important SNPs associated with estrogen response from resistant and sensitive NIC60 cell lines

Variable #	**SNP ID**	**Chr #**	**Location**	**Allele1**	**Allele2**	**MAF**[1]	**p-value**[2]	**FDR_BH**[3]	**Rank**[4] **(p-value)**	**Rank**[5] **(MOSS)**	**Related gene**[6]	**Distance**[7] **from closest gene**
8823	rs199449	17	45283725	A	G	0.23	0.0238	0.5058	1201	2	NSF	0 kb
11684	rs1915437	10	62530796	G	A	0.24	0.0005	0.3299	34	8	TMEM26	0 kb
17403	rs7077062	10	97646812	T	C	0.65	0.0002	0.3020	17	16	BLNK	0 kb
19083	rs1037182	9	110087746	C	T	0.47	0.0015	0.3668	101	13	UGCG	<20 kb
19132	rs7861999	9	110502843	A	G	0.30	0.0047	0.4037	289	14	HSDL2	<20 kb
7459	rs4923835	15	37542125	G	A	0.37	0.0009	0.3448	66	7	THSB1	<30 kb
14912	rs4703882	5	81799438	T	A	0.13	0.6648	0.9308	18211	5	FLJ41309	35 kb
11103	rs1370230	5	58973526	A	G	0.52	0.0576	0.5924	2478	15	MIRN582	40 kb
23629	rs1985065	4	165780046	G	A	0.18	0.9513	0.9919	24471	9	ANP32C	70 kb
11775	rs953654	1	62919199	A	G	0.42	0.2111	0.7609	7073	17	loc199897	95 kb
394	rs2279768	10	2933651	G	C	0.00	0.0004	0.3299	28	12	PFKP	120 kb
9571	rs1830876	6	51112593	C	T	0.21	0.0064	0.4210	385	11	FTHP1	150 kb
1596	rs200032	5	8744094	C	T	0.29	0.0001	0.2686	8	1	SEMA5A	400 kb
14303	–	16	78592091	G	C	0.17	0.0064	0.4210	383	6	WWOX	400 kb
4062	rs1577053	13	21136254	T	C	0.07	0.0000	0.2673	2	10	FTHL7	500 kb
5545	rs1474519	21	28715801	G	A	0.26	0.0071	0.4210	422	3	none	–
7357	rs1401154	8	37083828	T	C	0.13	0.0020	0.3716	139	4	none	–

[1] Minor allele frequency
[2] P-value from single marker test using Cochran-Armitage trend test
[3] False discovery rate from Benjamini and Hochberg (1995)
[4] Rank based on single-marker p-value
[5] Rank based on MOSS posterior probability for each single marker
[6] Closest gene based on physical distance
[7] Distance in kilo bases (Kb)

Table 2: Linkage Analysis of rs4923835 (marker 932558) with THBS1 and FSIP1. LD was mapped using HapMap Caucasian Data (HapMap Data Release 22, NCBI B36 assembly, dbSNP build 126). The LD of rs4923835 was constructed with 7 markers within the THBS1 and FSIP1. All the markers have shown strong LD ($r^2 \geq 0.9$) indicating the linkage of the significant marker to both THBs1 and FSIP1. The distance between the rs4923835 and the other SNPs ranged between 59-85Kb.

Significant Marker	**Markers within Genes**	**Marker Location**	**D′***	**LOD***	**r^2***	**CI†**	**Distance (Kb)**
rs4923835	rs2228263	exon 18 - THBS1	1	15.31	0.92	(0.84, 1)	59232
rs4923835	rs1051442	3′-UTR - THBS1	1	16.27	0.925	(0.85, 1)	61580
rs4923835	rs17633107	3′-UTR - THBS1	1	16.27	0.925	(0.85, 1)	62212
rs4923835	rs6492905	3′-UTR - FSIP1	1	16.06	0.924	(0.85, 1)	66333
rs4923835	rs17633210	intron 1 - FSIP1	1	15.61	0.92	(0.84, 1)	74497
rs4923835	rs17705806	intron 1 - FSIP1	1	16.27	0.925	(0.85, 1)	74402
rs4923835	rs17706083	intron 1 - FSIP1	1	16.27	0.925	(0.85, 1)	85005

* Measures of linkage disequilibrium (Devlin and Risch, 1995)

† Confidence interval of r^2

Table 3: Linkage Analysis of rs7861199 with SNPs in HSDL2, EPF5 and ROD1.

Significant Marker	**Markers within Genes**	**Marker Location**	**D′***	**LOD***	**r^2***	**CI†**	**Distance (Kb)**
rs7861999	rs3813855	5′ UTR - HSDL2	0.918	13.05	0.844	(0.75, 0.98)	18973
rs7861999	rs7852741	intron 1 - HSDL2	1	12.72	0.79	(0.81, 1)	33384
rs7861999	rs10759544	intron 6 - ROD1/EPF5	0.771	5.9	0.534	(0.51, 0.91)	100518

* Measures of linkage disequilibrium (Devlin and Risch, 1995)

† Confidence interval of r^2

Table 4: Linkage Analysis of rs1037182 with UGCG (linked to 5′ sequences).

Significant Marker	**Markers within Genes**	**Marker Location**	**D′***	**LOD***	**r^2***	**CI†**	**Distance (Kb)**
rs1037182	rs10817244	∼10Kb 5′-end of UGCG	1	18.75	0.718	(0.88, 1)	1022
rs1037182	rs7028129	∼20Kb 5′-end of UGCG	1	19.1	0.799	(0.88, 1)	286

* Measures of linkage disequilibrium (Devlin and Risch, 1995)

† Confidence interval of r^2

high-dimensional data problems.

9 Discussion

We summarize the key steps of our proposed methodology. The use of the conjugate priors for log-linear parameters of Massam et al. (2009) make the variable selection step and the log-linear model selection step coherent. The variable selection step is needed to allow us to focus on the most important covariates, while the model selection step is crucial to determine the most parsimonious representation of the underlying interactions. The efficiency of MOSS in identifying regions of high-posterior probability in the models space allow our method to scale to genomewide studies with data on $\geq$ 500K assayed SNPs as well as gene expression and clinical information. Bayesian model averaging is crucial to produce classifiers based on the most relevant small subsets of regressors. These subsets embody complex interactions among covariates of any type.

We allowed any combination of covariates as a possible candidate in our variable selection step. The accuracy of our method can be improved by incorporating prior knowledge related to spatial dependencies among markers or genes. Our methodology does not allow for the presence of missing data, hence the individuals for whom complete information is not available would have to be discarded. This could potentially lead to a significant decrease in the total sample size, hence our ability to detect higher-order associations or more complex combinations of covariates might be severely diminished. One alternative would be to use one time imputation methods such as the PHASE software of Stephens et al. (2001); Stephens and Donnelly (2003) for missing genotype/haplotype data. Another alternative that also takes into account the uncertainty related to missing genotype data is the Bayesian model selection and model averaging method proposed by Lunn et al. (2006). Unfortunately their MCMC approach does not seem to scale to whole genome scans.

C++ and Matlab code implementing the methods described in this chapter can be downloaded from

`http://www.stat.washington.edu/adobra/software/largetables/`

Acknowledgments

The authors would like to thank Ka Yee Yeung who provided us with the data for the breast cancer and leukemia examples and Chris Hans who provided the data for the lymph node status example.

References

Altshuler, D., Daly, M., 2007. Guilt beyond reasonable doubt. Nature Genetics 39 (7), 813–4.

Barrett, J. C., Fry, B., Maller, J., Daly, M. J., 2005. Haploview: analysis and visualization of ld and haplotype maps. Bioinformatics 21 (2), 263–265.

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: Series B 57, 289–300.

Carlin, B. P., Chib, S., 1995. Bayesian model choice via Markov chain Monte Carlo. Journal of the Royal Statistical Society, Series B 57, 473–484.

Chipman, H., George, E. I., McCullogh, R. E., 2001. The practical implementation of Bayesian model selection (with discussion). In: Lahiri, P. (Ed.), Model Selection. IMS: Beachwood, OH, pp. 66–134.

Christensen, K., Murray, J. C., 2007. What genome-wide association studies can do for medicine. New England Journal of Medicine 356, 1094–1097.

Clark, T. G., De Iorio, M., Griffiths, R. C., 2007. Bayesian logistic regression using a perfect phylogeny. Biostatistics 8, 32–52.

Clyde, M., George, E. I., 2004. Model uncertainty. Statistical Science 19, 81–94.

Cordell, H. J., Clayton, D. G., 2002. A unified stepwise regression procedure for evaluating the relative effects of polymorphisms within a gene using case/control or family data: application to HLA in type 1 diabetes. American Journal of Human Genetics 70, 124–141.

Darroch, J. N., Speed, T. P., 1983. Additive and multiplicative models and interaction. The Annals of Statistics 11, 724–738.

Dawid, A. P., Lauritzen, S. L., 1993. Hyper Markov laws in the statistical analysis of decomposable graphical models. The Annals of Statistics 21, 1272–1317.

Dellaportas, P., Forster, J. J., 1999. Markov chain Monte Carlo model determination for hierarchical and graphical log-linear models. Biometrika 86, 615–633.

Devlin, B., Risch, N., 1995. A comparison of linkage disequilibrium measures for fine-scale mapping. Genomics 29 (2), 311–22.

Diaconis, P., Ylvisaker, D., 1979. Conjugate priors for exponential families. The Annals of Statistics 7, 269–281.

Dobra, A., Hans, C., Jones, B., Nevins, J. R., Yao, G., West, M., 2004. Sparse graphical models for exploring gene expression data. Journal of Multivariate Analysis 90, 196–212.

Dobra, A., Massam, H., 2010. The mode oriented stochastic search (MOSS) for log-linear models with conjugate priors. Statistical MethodologyTo appear DOI:10.1016/j.stamet.2009.04.002.

Dudoit, S., Fridlyand, J., Speed, T. P., 2002. Comparison of discrimination methods for the classification of tumors using gene expression data. Journal of the American Statistical Association 97, 77–87.

Edwards, D. E., Havranek, T., 1985. A fast procedure for model search in multidimensional contingency tables. Biometrika 72, 339–351.

Efron, B., Tibshirani, R., 2002. Empirical Bayes methods and false discovery rates for microarrays. Genetics Epidemiology 23, 70–86.

Fernández, C., Ley, E., Steel, M. F., 2003. Benchmark priors for Bayesian model averaging. Journal of Econometrics 75, 317–343.

Furnival, G. M., Wilson, R. W., 1974. Regression by leaps and bounds. Technometrics 16, 499–511.

Garraway, L. A., Widlund, H. R., Rubin, M. A., Getz, G., Berger, A. J., Ramaswamy, S., Beroukhim, R., Milner, D. A., Granter, S. R., Du, J., Lee, C., Wagner, S. N., Li, C., Golub, T. R., Rimm, D. L., Meyerson, M. L., Fisher, D. E., Sellers, W. R., 2005. Integrative genomic analyses identify MTF as a lineage survival oncogene amplified in malignant melanoma. Nature 436, 117–122.

Geiger, D., Heckerman, D., 2002. Parameter priors for directed acyclic graphical models and the characterization of several probability distributions. The Annals of Statistics 30, 1412–1440.

George, E. I., McCulloch, R. E., 1993. Variable selection via Gibbs sampling. Journal of the American Statistical Association 88, 881–889.

George, E. I., McCulloch, R. E., 1997. Approaches for Bayesian variable selection. Statistica Sinica 7, 339–373.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., Lander, E. S., 1999. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. Science 286, 531–537.

Green, P. J., 1995. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika 82, 711–732.

Hans, C., Dobra, A., West, M., 2007. Shotgun stochastic search for "large p" regression. Journal of the American Statistical Association 102, 507–516.

Hoggart, C. J., Clark, T. G., De Iorio, M. D., Whittaker, J. C., Balding, D. J., 2008. Genome-wide significance for dense SNP and resequencing data. Genetic Epidemiology 32, 179–185.

Hu, J., Joshi, A., Johnson, V. E., 2009. Log-linear models for gene association. Journal of the American Statistical Association 104, 597–607.

Jarjanazi, H., Kiefer, J., Savas, S., Briollais, L., Tuzmen, S., Pabalan, N., Ibrahim-Zada, I., Mousses, S., Ozcelik, H., 2008. Discovery of genetic profiles impacting response to chemotherapy: Application to gemcitabine. Human Mutations 29, 461–467.

Jones, B., Carvalho, C., Dobra, A., Hans, C., Carter, C., West, M., 2005. Experiments in stochastic computation for high-dimensional graphical models. Statistical Science 20, 388–400.

Kass, R., Raftery, A. E., 1995. Bayes factors. Journal of the American Statistical Association 90, 773–795.

Lauritzen, S. L., 1996. Graphical Models. Clarendon Press, Oxford.

Lee, K. E., Sha, N., Dougherty, E. R., Vanucci, M., Mallick, B. K., 2003. Gene selection: a Bayesian variable selection approach. Bioinformatics 19, 90–97.

Liang, F., Paulo, R., Molina, G., Clyde, M., Berger, J. O., 2008. Mixtures of g-priors for Bayesian variable selection. Journal of the American Statistical Association 103, 410–423.

Lunn, D. J., Whittaker, J. C., Best, N., 2006. A Bayesian toolkit for genetic association studies. Genetic Epidemiology 30, 231–247.

Madigan, D., Raftery, A. E., 1994. Model selection and accounting for model uncertainty in graphical models using Occam's window. Journal of the American Statistical Association 89, 1535–1546.

Madigan, D., York, J., 1995. Bayesian graphical models for discrete data. International Statistical Review 63, 215–232.

Madigan, D., York, J., 1997. Bayesian methods for estimation of the size of a closed population. Biometrika 84, 19–31.

Marchini, J., Donnelly, P., Cardon, L. R., 2005. Genome-wide strategies for detecting multiple loci that influence complex diseases. Nature Genetics 37, 413–417.

Massam, H., Liu, J., Dobra, A., 2009. A conjugate prior for discrete hierarchical log-linear models. The Annals of Statistics 37, 3431–3467.

Nguyen, D. V., Rocke, D. M., 2002. Tumor classification by partial least squares using microarray gene expression data. Bioinformatics 18, 39–50.

Nott, D. J., Green, P. J., 2004. Bayesian variable selection and the Swendsen-Wang algorithm. Journal of Computational and Graphical Statistics 13, 1–17.

Pittman, J., Huang, E., Dressman, H., Horng, C. F., Cheng, S. H., Tsou, M. H., Chen, C. M., Bild, A., Iversen, E. S., Huang, A. T., Nevins, J. R., West, M., 2004. Integrated modeling of clinical and gene expression information for personalized prediction of disease outcomes. Proceedings of the National Academy of Sciences 101, 8431–8436.

Raftery, A. E., Madigan, D., Hoeting, J., 1997. Bayesian model averaging for linear regression models. Journal of the American Statistical Association 92, 1197–1208.

Ruczinski, I., Kooperberg, C., LeBlanc, M., 2003. Logic regression. Journal of Computational and Graphical Statistics 12, 475–511.

Sarkar, A. J., Chaturvedi, K., Chen, C. P., Sarkar, D. K., 2007. Changes in thrombospondin-1 levels in the endothelial cells of the anterior pituitary during estrogen-induced prolactin-secreting pituitary tumors. Journal of Endocrinology 192, 395–403.

Schaid, D., 2004. Evaluating associations of haplotypes with traits. Genetic Epidemiology 27, 348–364.

Sengupta, K., Banerjee, S., Saxena, N. K., Banerjee, S. K., 2004. Thombospondin-1 disrupts estrogen-induced endothelial cell proliferation and migration and its expression is suppressed by estradiol. Molecular Cancer Research 2, 150–8.

Slater, M., Patava, J., Mason, R. S., 1995. Thrombospondin co-localises with TGF beta and IGF-I in the extracellular matrix of human osteoblast-like cells and is modulated by 17 beta estradiol. Experientia 51 (3), 235–44.

Stephens, M., Donnelly, P., 2003. A comparison of Bayesian methods for haplotype reconstruction from population genotype data. American Journal of Human Genetics 73, 1162:1169.

Stephens, M., Smith, N. J., Donnelly, P., 2001. A new statistical method for haplotype reconstruction from population data. American Journal of Human Genetics 68, 978:989.

Storey, J. D., Tibshirani, R., 2003. Statistical significance for genomewide studies. Proceedings of the National Academy of Sciences 100, 9440–9445.

Swartz, M. D., Duncan, D. C., Daw, E. W., 2007. Model selection and Bayesian methods in statistical genetics: Summary of Group 11 contributions to Genetic Analysis Workshop 15. Genetics Epidemiology 31 (Supplement 1), S96–S102, on behalf of Group 11.

Thomas, D. C., Haile, R. W., Duggan, D., 2005. Recent developments in genomewide association scans: a workshop summary and review. American Journal of Human Genetics 77, 337–345.

Tusher, V. G., Tibshirani, R., Chu, G., 2001. Significance analysis of microarrays applied to the ionizing radiation response. Proceedings of the National Academy of Sciences 98, 5116–5121.

van't Veer, L. J., Hongyue, D., van de Vijver, M. J., He, Y. D., Hart, A. A. M., Mao, M., Peterse, H. L., van der Kooy, K., Marton, M. J., Witteveen, A. T., Schreiber, G. J., Kerkhoven, R. M., Roberts, C., Linsley, P. S., Bernards, R., Friend, S. H., 2002. Gene expression profiling predicts clinical outcome of breast cancer. Nature 415, 530–536.

Verzilli, C. J., Stallard, N., Whittaker, J. C., 2006. Bayesian graphical models for genomewide association studies. American Journal of Human Genetics 79, 100–112.

Wand, M. P., Jones, M. C., 1995. Kernel Smoothing. Chapman and Hall.

Wang, K., Abbott, D., 2008. A principal components regression approach to multilocus genetic association studies. Genetic Epidemiology 32, 108–118.

Yeung, K., Bumgarner, R. E., Raftery, A. E., 2005. Bayesian model averaging: Development of an improved multi-class, gene selection and classification tool for microarray data. Bioinformatics 21, 2394–2402.

Zhang, Y., Liu, J. S., 2007. Bayesian inference of epistatic interactions in case-control studies. Nature Genetics 39, 1167–73.