BAYESIAN STATISTICS 9
J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid,
D. Heckerman, A. F. M. Smith and M. West (Eds.)

Bayesian Models for Variable Selection that Incorporate Biological Information

Marina Vannucci and Francesco C. Stingo
Rice University, USA
marina@rice.edu fcs1@rice.edu

Summary

Variable selection has been the focus of much research in recent years. Bayesian methods have found many successful applications, particularly in situations where the amount of measured variables can be much greater than the number of observations. One such example is the analysis of genomics data. In this paper we first review Bayesian variable selection methods for linear settings, including regression and classification models. We focus in particular on recent prior constructions that have been used for the analysis of genomic data and briefly describe two novel applications that integrate different sources of biological information into the analysis of experimental data. Next, we address variable selection for a different modeling context, i.e. mixture models. We address both clustering and discriminant analysis settings and conclude with an application to gene expression data for patients affected by leukemia.

Keywords and Phrases: Classification and Clustering; Discriminant Analysis; Gene Networks; Markov Random Field Priors; Pathways; Regression Models; Variable Selection.

1. INTRODUCTION

The practical utility of variable selection is well recognized and this topic has been the focus of much research. Variable selection can help assessing the importance of explanatory variables, improving prediction accuracy, providing a better understanding of the underlying mechanisms generating data and reducing the cost of measurement and storage for future data. Bayesian methods for variable selection have several appealing features. They address the selection and prediction problems in a unified manner, they allow rich modeling via the implementation of MCMC stochastic search strategies and incorporate optimal model averaging prediction strategies; they extend quite naturally to multivariate responses and many linear and nonlinear settings; they can handle the "small n - large p" setting, *i.e.*,

M. Vannucci is Professor of Statistics and F.C. Stingo is Postdoctoral Fellow at Rice University, Houston, TX, USA. Work partially supported by NIH and NSF.

situations where the number of measured covariates is much larger than the sample size; they allow past and collateral information to be easily accommodated into the model through the priors.

In this paper we first consider modeling frameworks that express a response variable as a linear combination of predictors and offer a review of Bayesian methods for variable selection that use mixture priors with a spike at zero. The key idea of the approach is to introduce latent binary vectors, representing the possible subsets of predictors, that induce mixture priors on the regression coefficients of the model. The approach was first developed for the commonly used regression setting and it extends quite easily to other linear settings via data augmentation strategies.

The flexibility of the approach and the fact that it can handle the"large p - small n" paradigm have made the Bayesian methods particularly relevant for the analysis of genomic studies, where high-throughput technologies allow thousands of variables to be measured on individual samples. We briefly discuss recent contributions that focus on developing prior constructions that incorporate biological information into the models. We present in some details two novel applications: One considers a linear model that predicts a phenotype based on predictors synthesizing the activity of genes belonging to same pathways. The prior model encodes information on gene-gene networks, as retrieved from available databases. The other application concerns a statistical procedure that aims at inferring a biological network of very high dimensionality, where microRNAs, small RNAs, are supposed to down-regulate mRNAs, also called targets, and where sequence and structure information is integrated into the model via the prior formulation.

In the second part of the paper we briefly describe how some of the key ideas of the variable selection methods for linear settings can be used in a different modeling context, i.e. mixture models. We treat both unsupervised, i.e. clustering, and supervised settings for pattern recognition. Latent binary vectors are introduced again to achieve the selection. However, the inclusion of the latent indicators into the model is done via the likelihood rather than a prior model on regression coefficients.

The rest of the paper is organized as follows. In Section 2 we briefly review Bayesian methods for variable selection in linear modeling settings and briefly describe extensions and applications that take into account specific characteristics of genomics data. In Section 3 we discuss variable selection in the context of mixture models, for both unsupervised and supervised pattern recognition, and present an application to DNA microarray data.

2. MIXTURE PRIORS FOR VARIABLE SELECTION

2.1. *Review of the Approach for Linear Regression Models*

Let us start with the classical linear regression model

$$\mathbf{Y} = \mathbf{1}\alpha + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{1}$$

with $\boldsymbol{\epsilon} \sim N(0, \sigma^2\mathbf{I}_n)$ and where $\mathbf{Y}$ is the $n \times 1$ response vector, $\mathbf{X}$ the $n \times p$ matrix of predictors and $\boldsymbol{\beta}$ the $p \times 1$ vector of regression coefficients. Often, in applications, not all p covariates play an important role in explaining changes of the response and one goal of the analysis is to identify the important variables. This is a problem of variable selection.

In the Bayesian paradigm variable selection can be achieved by imposing mixture priors on the regression coefficients of model (1) via a latent binary vector,

$\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)$, as

$$\beta_j|\sigma^2 \sim (1-\gamma_j)\delta_0(\beta_j) + \gamma_j N(0, h_j\sigma^2), \tag{2}$$

where $\delta_0(\cdot)$ is the Dirac function at zero and the h_j's hyperparameters to be chosen. With this prior, if $\gamma_j = 0$ then β_j is set to 0, whereas if $\gamma_j = 1$ a nonzero estimate of β_j corresponds to an important predictor. In addition, conjugate priors can be imposed on α and σ^2, *i.e.*,

$$\alpha|\sigma^2 \sim N(\alpha_0, h_0\sigma^2) \tag{3}$$

$$\sigma^2 \sim IG(\nu/2, \lambda/2) \tag{4}$$

with α_0, h_0, ν and λ to be chosen. Mixture priors of type (2) for univariate linear regression models were originally proposed by Leamer (1978) and Mitchell and Beauchamp (1988) and made popular by George and McCulloch (1993,1997), Geweke (1996), Clyde *et al.* (1996), Smith and Kohn (1996), Carlin and Chib (1995) and Raftery *et al.* (1997). Brown *et al.* 1998a,2002) extended the construction to multivariate linear regression models with q response variables. Reviews of special features of the selection priors and on computational aspects can be found in Chipman *et al.* (2001) and Clyde and George (2004).

Common choices of the hyperparameters h_j's in the prior model (2) assume that the β_j's are a priori independent given $\boldsymbol{\gamma}$, for example, by choosing $h_j = c$ for every j. Brown *et al.* (1998a) investigate the case of h_j chosen to be proportional to the j-th diagonal element of $(\mathbf{X}'\mathbf{X})^{-1}$, while Smith and Kohn (1996) propose the use of a Zellner's g-prior, see Zellner (1986), of the type

$$\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\sigma^2 \sim N(0, c(\mathbf{X}'_{\boldsymbol{\gamma}}\mathbf{X}_{\boldsymbol{\gamma}})^{-1}\sigma^2). \tag{5}$$

Priors of type (5) have an intuitive interpretation as they use the design matrix of the current experiment. Recently, Liang *et al.* (2008) and Cui and George (2008) have investigated formulations that use a fully Bayesian approach by imposing mixtures of g-priors on c. They also propose hyper-g priors for c which lead to closed form marginal likelihoods and nonlinear shrinkage via Empirical Bayes procedures.

Prior construction (2) also requires the choice of a prior distribution for $\boldsymbol{\gamma}$. The simplest and most common choice adopted in the literature is a product of independent Bernoulli's of the type

$$p(\boldsymbol{\gamma}) = \prod_{j=1}^{p} w_j^{\gamma_j}(1-w_j)^{1-\gamma_j}, \tag{6}$$

with $w_j = p(\gamma_j = 1)$ the prior probability of inclusion of the j-th variable in the model. A suitable choice is $w_j = w$ which implies that $p \times w$ is the number of variables expected *a priori* to be included in the model. Uncertainty on w can be modeled by imposing a Beta hyperprior, $w \sim Beta(a, b)$, with a, b to be chosen, see for example Brown *et al.* (1998b). An attractive feature of these priors is that appropriate choices of w that depend on p impose an a priori multiplicity penalty, as argued in Scott and Berger (2010). Recent contributions to the application of Bayesian variable selection models in the analysis of genomic data have featured priors on γ that exploit the complex dependence structure between genes (variables)

linked via underlying biological processes and/or networks. Some of these contributions are described below.

Efficient schemes for posterior inference can be obtained by integrating out the model parameters to obtain the posterior distribution of $\boldsymbol{\gamma}$,

$$p(\boldsymbol{\gamma}|\mathbf{Y},\mathbf{X}) \propto p(\mathbf{Y}|\boldsymbol{\gamma},\mathbf{X})p(\boldsymbol{\gamma}). \tag{7}$$

When a large number of predictors makes the full exploration of the model space unfeasible, Monte Carlo Markov chain methods can be used as stochastic searches to quickly and efficiently explore the posterior distribution looking for "good" models, *i.e.*, models with high posterior probability, see George and McCulloch (1997). The most popular is the Metropolis scheme (MC^3), proposed by Madigan and York (1995) in the context of model selection for discrete graphical models and subsequently adapted to variable selection, see Raftery *et al.* (1997) and Brown *et al.* (1998b,2002), among others. Improved MCMC schemes have been proposed to achieve an even faster exploration of the posterior space, see for example the *shotgun* algorithm of Hans *et al.* (2007) and the evolutionary Monte Carlo schemes combined with parallel tempering proposed by Bottolo and Richardson (2010).

The MCMC procedure results in a list of visited models, $\boldsymbol{\gamma}^{(0)},\cdots,\boldsymbol{\gamma}^{(T)}$ and their corresponding posterior probabilities. Variable selection can then be achieved either by looking at the $\boldsymbol{\gamma}$ vectors with largest joint posterior probabilities among the visited models or, marginally, by calculating frequencies of inclusion for each γ_j and then choosing those γ_j's with frequencies exceeding a given cut-off value. Finally, prediction of future observations Y^f can be done based on the selected models, either via least squares on single models or by using the *model averaging* idea of Madigan and York (1995). This procedure is based on the predictive distribution $p(Y^f|Y,X^f)$ and exploits the conjugacy of the model. After integrating α, β and σ out it is possible to calculate Y^f as weighted mean of the expected values of $p(Y^f|Y,X^f)$ given different configurations of γ, with the weights being the posterior probabilities of these configurations. Only the best k configurations, according to the posterior probabilities, are typically used for prediction.

2.2. *Extensions to Other Linear Settings*

The prior models for variable selection described above can be easily applied to other modeling settings, where a response variable is expressed as a linear combinations of the predictors. For example, probit models were considered by Sha *et al.* (2003,2004) and Kwon *et al.* (2007). In this setting data augmentation approaches allow to express the model in the linear framework (1), with latent responses, and conjugate priors allow to integrate the model parameters out, therefore facilitating the implementation of very efficient MCMC schemes. Holmes and Held (2006) considered logistic models and a data augmentation approach that uses latent variables to write the model in linear form. Gustafson and Lefebvre (2008) extended methodologies to settings where the subset of predictors associated with the propensity to belong to a class varies with the class. Sha *et al.* (2006) considered accelerated failure time models for survival data.

Probit and logit models, in particular, belong to the more general class of generalized linear models (GLMs) of McCullagh and Nelder (1989), that assume the distribution of the response variable as coming from the exponential family. Conditional densities in the general GLM framework cannot be obtained directly and the

resulting mixture posterior may be difficult to sample using standard MCMC methods due to multimodality. Some attempts to Bayesian variable selection methods for GLMs were done by Raftery (1996), who proposed approximate Bayes factors, and by Ntzoufras *et al.* (2003), who developed a method to jointly select variables and the link function. See also Ibrahim *et al.* (2000) and Chen *et al.* (2003).

Among possible extensions of linear models, we also mention the class of mixed models, that include random effects capturing heterogeneity among subjects, Laird and Ware (1982). One challenge in developing SSVS approaches for random effects models is the constraint that the random effects covariance matrix needs to be semi-definite positive. Chen and Dunson (2003) imposed mixture priors on the regression coefficients of the fixed effects and achieve simultaneous selection of the random effects by imposing variable selection priors on the components in a special LDU decomposition of the random effects covariance. Cai and Dunson (2006) extended the approach to generalized linear mixed models (GLMM).

2.3. *Priors that Incorporate Biological Information*

The flexibility of the prior models for variable selection and the fact that the inferential methods can handle the"large p - small n" paradigm have made these techniques particularly relevant for the analysis of genomic studies, where high-throughput technologies allow thousands of variables to be measured on individual samples. Recent contributions in particular have focused on developing prior constructions that incorporate biological information, typically available via online databases, into the models.

Chen *et al.* (2010) consider the problem of finding genes that relate to a response variable. In their approach the authors take into account that recent interest in biology has moved from the analysis of single genes to the analysis of known groups of genes, called pathways. Many databases exist now where information on pathways, including gene-pathway memberships, and on gene-gene networks can be retrieved. In the proposed model formulation pathway "scores" that synthesize the activity of each pathway are defined via partial least square techniques and used as predictors in a model of type (1). Gene network information is then encoded through the prior distribution on γ. In particular, gene-gene relations are modeled using a Markov random field (MRF) model, where genes are represented by nodes and relations between them by edges. One possible parametrization of the MRF, used in Chen *et al.* (2010), is represented by the following probabilities:

$$p(\gamma_j|\mu,\eta,\gamma_k,k\in N_j)=\frac{\exp(\gamma_j F(\gamma_j))}{1+\exp(F(\gamma_j))}\ , \tag{8}$$

where $F(\gamma_j)=\mu+\eta\sum_{k\in N_j}(2\gamma_k-1)$ and N_j is the set of direct neighbors of variable j in the MRF. The global distribution on the MRF is given by

$$p(\gamma|\mu,\eta)\propto\exp(\mu\ n_1-\eta\ n_{01}), \tag{9}$$

where n_1 is the number of selected variables and n_{01} is the number of edges linking nodes with different values of γ_j (*i.e.*, edges linking included and non-included nodes),

$$n_1=\sum_{j=1}^{q}\gamma_j,\qquad n_{01}=\frac{1}{2}\sum_{k=1}^{q}\left[\sum_{j=1}^{q}r_{kj}-\left|\sum_{j=1}^{q}r_{kj}(1-\gamma_k)-\sum_{j=1}^{q}r_{kj}\gamma_j\right|\right].$$

The parameter μ controls the sparsity of the model, while higher values of η result in neighboring variables taking on the same γ_j value. If a variable does not have any neighbor, its prior distribution reduces to an independent Bernoulli with parameter $p = \exp(\mu)/[1 + \exp(\mu)]$, which is a logistic transformation of μ.

Other contributions to the use of MRF priors for genomic data include Telesca *et al.* (2008), who have proposed a model for the identification of differentially expressed genes that takes into account the dependence structure among genes from available pathways while allowing for correction in the gene network topology. Also, Li and Zhang (2010) incorporate the dependence structure of transcription factors in a regression model with gene expression outcomes; in their approach a network is defined based on the Hamming distance between candidate motifs and used to specify a Markov random field prior for the motif selection indicator. A different parametrization of the MRF is used, corresponding to the following distribution for γ:

$$p(\gamma|D, G) \propto \exp(D'\gamma + \gamma' G\gamma) \tag{10}$$

with $D = d1_p$, 1_p the unit vector of dimension p and G a matrix with elements $\{g_{ij}\}$ usually set to some constants. While d plays the same role as μ in (9), G and η affect the probability of selection of a variable in different ways. This is evident from the conditional probability

$$P(\gamma_j|d, g, \gamma_k, k \in N_j) = \frac{\exp(\gamma_j(d + g\sum_{k\in N_j}\gamma_k))}{1 + \exp(d + g\sum_{k\in N_j}\gamma_k)}\ , \tag{11}$$

which can only increase as a function of the number of selected neighbor genes. In contrast, with the parametrization in (8), the prior probability of selection for a variable does not decrease if none of the neighbors are selected. Although the parametrization is somewhat arbitrary, some care is needed in deciding whether to put a prior distribution on G. Allowing G to vary can lead to a phase transition problem, that is, the expected number of variables equal to 1 can increase massively for small increments of G. This problem can happen because equation (11) can only increase as a function of the number of the x_j's equal to 1.

2.4. *A Graphical Model Formulation for Regulatory Network Inference*

Variable selection methods have also been extended to graphical models. These focus on identifying latent graphical structure that encodes conditional independencies, see Whittaker (1990) and Cowell *et al.* (1999) among others. A graph is formed by nodes and arcs; nodes represent random variables and the lack of arcs represents conditional independence. Hence graphical models provide a compact representation of joint probability distributions. Arcs can be undirected or directed. Undirected graphical models are also called Markov Random Field (MRF) models. Directed graphical models are also called Bayesian Network (BN). Directed acyclic graph (DAG), in particular, do not allow for the presence of cycles. Conditional independencies in a DAG depend on the ordering of the variables. When the joint distribution is a multivariate normal the model is called Graphical Gaussian model (GGM). Nodes that are directly connected to node j and precede j in the ordering are called parents of j. In a Bayesian Network, X_j is independent, given its parents, of the set of all the other variables in the graph, except its parents.

Bayesian treatments of model selection for discrete graphical models, such as DAG, were first considered by Madigan and Raftery (1994) and Madigan and

York (1995). With multivariate Gaussian data the selection of an edge is equivalent to setting equal to zero the corresponding element of the concentration matrix, see also Giudici and Green (1999). Efficient stochastic search procedures can be implemented when the graph is decomposable using an hyper Inverse Wishart prior for the covariance matrix that allows to explicitly obtain the marginal likelihoods, as first noted by Clyde and George (2004). Jones *et al.* (2005) describe how to perform Bayesian variable selection for both decomposable and non decomposable undirected Gaussian graphical models in a high dimensional setting, underlining the computational difficulties for the latter case, see also Roverato (2002) and Dobra *et al.* (2004).

When the goal of the analysis is to recover the structure of a directed graphical model, with the ordering of the variables known a priori, it is possible to write the model in terms of a system of linear equations and therefore employ the spike and slab prior formulation (2) for the regression coefficients to achieve variable selection. Exploiting this idea, Stingo *et al.* (2010) put forward a graphical model formulation of a multivariate regression model which is used to infer a biological network of very high dimensionality, where microRNAs, small RNAs, are supposed to down-regulate mRNAs, also called target genes. The main goal of the model is to understand which elements of the network are connected and which ones are not. In addition, specific biological characteristics/constraints need to be considered. Their model formulation includes constraints on the regression coefficients and selection priors that incorporate biological knowledge. The variable selection formulation they adopt overcomes the somehow rigid structure of the model in Brown *et al.* (1998a), which does not allow to select different predictors for different responses. See also Monni and Tadesse (2009) for an approach based on partition models.

Briefly, Stingo *et al.* (2010) define a DAG and impose an ordering of the variables such that each target gene can be affected only by the miRNAs and that the miRNAs can affect only the targets. Let $\mathbf{Z} = (\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_G, \mathbf{X}_1, \ldots, \mathbf{X}_M)$ with $\mathbf{Y} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_G)$ the matrix representing the targets and $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_M)$ the miRNAs. In their application the data consist of $G = 1,297$ targets and $M = 23$ miRNAs observed on $N = 11$ units. Matrix $\mathbf{Z}$ is assumed to be a matrix-variate normal variable with zero mean and a variance matrix Ω for its generic row, that is, following the notation of Dawid (1981), $\mathbf{Z} - \mathbf{0} \sim \mathcal{N}(I_N, \Omega)$. In addition, the assumption that the target genes are independent conditionally upon the miRNAs, that is, $\mathbf{Y}_i \perp\!\!\!\perp \mathbf{Y}_j | \mathbf{X}_1, \ldots, \mathbf{X}_M$ is made. Note that assumptions on the marginal distribution of $(\mathbf{X}_1, \ldots, \mathbf{X}_M)$ do not affect the regulatory network. In a Bayesian Network framework these assumptions imply an ordering of the nodes and, consequently, a likelihood factorization of the type:

$$p(\mathbf{Z}) = \prod_{g=1}^{G} p(\mathbf{Y}_g | \mathbf{X}) \prod_{m=1}^{M} p(\mathbf{X}_m),$$

where $p(\mathbf{Y}_g|\mathbf{X}) \sim N(\mathbf{X}\beta_g, \sigma_g I_N)$ and $p(\mathbf{X}_m) \sim N(0, \sigma_m I_N)$, with $\beta_g = \Omega_{\mathbf{XX}}^{-1}\Omega_{\mathbf{XY}_g}$ and $\sigma_g = \omega_{gg} - \Omega_{\mathbf{XY}_g}^T \Omega_{\mathbf{XX}}^{-1} \Omega_{\mathbf{XY}_g}$. Here ω_{gg} indicates the g-th diagonal element of Ω and $\Omega_{\mathbf{XX}}$, $\Omega_{\mathbf{XY}}$ are the appropriate blocks of the covariance matrix. For $m = 1, \ldots, M$ we have $\sigma_m = \omega_{mm}$. This graphical model formulation is equivalent to a system of G linear regression models.

Knowledge about the fact that miRNAs down-regulate gene expression can be incorporated into the model by specifying negative regression coefficients via the

prior choice, i.e. $(\tilde{\beta}_{gm}|\sigma_g) \sim Ga(1, c\,\sigma_g)$ and $\sigma_g^{-1} \sim Ga((\delta + M)/2, d/2)$, with $\beta_{gm} = -\tilde{\beta}_{gm}$. Furthermore, the underlying regulatory network can be completely encoded introducing a $(G \times M)$ association matrix $\boldsymbol{R}$ with elements $r_{gm} = 1$ if the mth miRNA is included in the regression of the gth target and $r_{gm} = 0$ otherwise. The regression coefficient parameters are then stochastically independent, given the regulatory network $\boldsymbol{R}$, and have the following mixture prior distribution:

$$\pi(\tilde{\beta}_{gm}|\sigma_g, r_{gm}) = r_{gm} Ga(1, c\,\sigma_g) + (1 - r_{gm})\delta_0(\tilde{\beta}_{gm}). \tag{12}$$

Prior distributions for $\boldsymbol{R}$ can be specified by taking into account biological information encoded by sequence/structure databases available on the internet. Scores of possible gene-miRNA pair associations that come from these sources can be integrated into the model by defining the prior probability of selecting the edge between a gene g and a miRNA m as:

$$P(r_{gm} = 1|\tau) = \frac{\exp[\eta + \tau_1 s^1_{gm} + \tau_2 s^2_{gm} + \ldots + \tau_J s^J_{gm}]}{1 + \exp[\eta + \tau_1 s^1_{gm} + \tau_2 s^2_{gm} + \ldots + \tau_J s^J_{gm}]} ,$$

with $\tau = (\tau_1, \ldots, \tau_J)$ and where the s^j_{gm}'s, with $j = 1, \ldots, J$, denote the J available scores.

For posterior inference, the regression coefficients can be integrated out, reducing the computational complexity of the MCMC algorithm to the sampling of the models space, $\boldsymbol{R}$, the data integration parameters, τ_j, and the variances, σ_g. See Stingo *et al.* (2010) for details.

3. MIXTURE MODELS

In this second part of the paper we address variable selection in a different modeling context, i.e. mixture models for pattern recognition. We treat in particular the unsupervised framework, known in the statistical literature as clustering, and then describe an adaptation to the simpler supervised framework, known as discriminant analysis. For both model formulations we borrow ideas from the linear settings treated in Section 2.1. For example, a latent binary vector $\boldsymbol{\gamma}$ is introduced for variable selection, and stochastic search MCMC techniques are used to explore the space of variable subsets. However, building a variable selection mechanism into mixture models is more challenging than the linear settings. In clustering, for example, there is no observed response to guide the selection and the elements of the matrix $\mathbf{X}$ are viewed as random variables. The inclusion of the latent indicators into the models, therefore, cannot be done like in the linear modeling context, where $\boldsymbol{\gamma}$ is used to induce mixture priors on regression coefficients.

3.1. *Model-based Clustering*

A first attempt to cluster high-dimensional data was done by Liu *et al.* (2003) who addressed the problem by first reducing the dimension of the data using principal component analysis and then fitting a mixture model on the factors, with a fixed number of clusters. They used Markov chain Monte Carlo sampling techniques to update the sample allocations and the number of factors deemed relevant for the clustering. An approach to variable selection for model-based clustering was put forward by Tadesse *et al.* (2005), who formulated the clustering in terms of a finite mixture of Gaussian distributions with an unknown number of components and then

introduced latent variables to identify discriminating variables. The authors used a reversible jump Markov chain Monte Carlo technique to allow for the creation and deletion of clusters. A similar model was considered by Raftery and Dean (2006). Kim *et al.* (2006) proposed an alternative modeling approach that uses infinite mixture models via Dirichlet process priors. Hoff (2006) adopted a mixture of Gaussian distributions where different clusters are identified by mean shifts and Bayes factors are computed to identify discriminating variables. This method allows separate subsets of variables to discriminate different groups of observations.

In the finite mixture model formulation of Tadesse *et al.* (2005) the data are viewed as coming from a mixture of distributions:

$$p(\boldsymbol{x}_i|\boldsymbol{w},\boldsymbol{\phi}) = \sum_{k=1}^{K} w_k\, p(\boldsymbol{x}_i|\boldsymbol{\phi}_k),$$

where $p(\boldsymbol{x}_i|\boldsymbol{\phi}_k)$ is the density of sample $\boldsymbol{x}_i$ from group k and $\boldsymbol{w} = (w_1, \cdots, w_K)^T$ are the cluster weights ($\sum_k w_k = 1, w_k \geq 0$), see McLachlan and Basford (1988). Here K is assumed finite but unknown. Latent variables $\boldsymbol{c} = (c_1, \cdots, c_n)^T$, with $c_i = k$ if the i-th sample comes from group k, are introduced to identify the cluster from which each observation is drawn.

The sample allocations, c_i, are assumed to be independently and identically distributed with probability mass function $p(c_i = k) = w_k$. We assume that the mixture distributions are multivariate normal with component parameters $\boldsymbol{\phi}_k = (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$. Thus, for sample i, we have

$$\boldsymbol{x}_i|c_i = k, \boldsymbol{w}, \boldsymbol{\phi} \;\sim\; \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k). \tag{13}$$

For variable selection, a latent binary vector $\boldsymbol{\gamma}$ is used to identify the discriminating variables. More specifically, variables indexed by a $\gamma_j = 1$, denoted $\mathbf{X}_{(\gamma)}$, define the mixture distribution, while variables indexed by $\gamma_j = 0$, $\mathbf{X}_{(\gamma^c)}$, favor one multivariate normal distribution across all samples. The distribution of sample i is then given by

$$\begin{aligned} \boldsymbol{x}_{i(\gamma)}|c_i = k, \boldsymbol{w}, \boldsymbol{\phi}, \boldsymbol{\gamma} \;&\sim\; \mathcal{N}(\boldsymbol{\mu}_{k(\gamma)}, \boldsymbol{\Sigma}_{k(\gamma)}) \\ \boldsymbol{x}_{i(\gamma^c)}|\boldsymbol{\psi}, \boldsymbol{\gamma} \;&\sim\; \mathcal{N}(\boldsymbol{\eta}_{(\gamma^c)}, \boldsymbol{\Omega}_{(\gamma^c)}), \end{aligned} \tag{14}$$

where $\boldsymbol{\psi} = (\boldsymbol{\eta}, \boldsymbol{\Omega})$.

Priors on $\boldsymbol{\gamma}$ can be specified similarly to what discussed for the linear settings of Section 2. For the vector of component weights, a symmetric Dirichlet prior can be specified. For the unknown number of components, K, a truncated Poisson or a discrete Uniform prior on $[1, \ldots, K_{\max}]$, where $K_{\max}$ is chosen arbitrarily large, are suitable choices. An efficient sampler can be implemented by working with a marginalized likelihood where the model parameters are integrated out. The integration is facilitated by taking conjugate Normal-Wishart priors on both $\boldsymbol{\phi}$ and $\boldsymbol{\psi}$. Some care is needed in the choice of the hyperparameters. In particular, the variance parameters need to be specified within the range of variability of the data. The MCMC procedure is described in Tadesse *et al.* (2005) and requires a sampler that jumps between different dimensional spaces, generalizing the reversible jump approach of Richardson and Green (1997).

3.2. *Discriminant Analysis*

We now show an adaptation of the method to the simpler supervised setting, where, in addition to the observed vectors $\boldsymbol{x}_i$'s, the number of groups K and the classification labels c_i's are also available and where the aim is to derive a classification rule that will assign further cases to their correct groups. When the distribution of $\mathbf{X}$ conditional on the group membership is assumed normal then this statistical methodology is known as discriminant analysis. Fearn *et al.* (2002) proposed a Bayesian decision theory approach to variable selection for this modeling setting.

In discriminant analysis, given the selected variables, the predictive distribution of a new observation $\boldsymbol{x}^f$ is used to classify every new sample into one of the possible K groups. This distribution is a multivariate T-student, see Brown (1993) among others. The probability that a future observation, given the observed data, belongs to the group k is then given by:

$$\pi_k(c^f|\boldsymbol{X}) = p(c^f = k|\boldsymbol{x}^f, \mathbf{X}) \tag{15}$$

where c^f is the group indicator of $\boldsymbol{x}^f$. By estimating the prior probability that one observation comes from group k as $\hat{\pi}_k = n_k/n$, the previous distribution can be written in closed form as:

$$\pi_k(c^f|\boldsymbol{X}) = \frac{p_k(\boldsymbol{x}^f)\hat{\pi}_k}{\sum_{i=1}^K p_i(\boldsymbol{x}^f)\hat{\pi}_i} \ ,$$

where $p_k(\boldsymbol{x}^f)$ indicates the predictive T-student distribution. A new observations is then assigned to the group with the highest posterior probability.

As in the clustering setting, we introduce a latent binary vector γ to perform the selection. As done by Raftery and Dean (2006), extending the approach of Tadesse *et al.* (2005) to avoid any independence assumptions, the following likelihood can be used to separate the discriminant variables from the noisy ones as:

$$L(X, c; \cdot) = \prod_{i=1}^{n} p(\boldsymbol{x}_{i(\gamma^c)}|\boldsymbol{x}_{i(\gamma)}) \prod_{k=1}^{K} w_k^{n_k} \prod_{j=1}^{n_k} p_k(\boldsymbol{x}_{j(\gamma)}). \tag{16}$$

The first factor of the likelihood refers to the non important variables, while the second is formed by variables able to classify observations into the correct groups. Under the normality assumption the likelihood becomes:

$$\prod_{i=1}^{n} N_{|\gamma^c|}(\boldsymbol{x}_{i(\gamma^c)} - \beta \boldsymbol{x}_{i(\gamma)}; \eta_{(\gamma^c)}, \Sigma_{(\gamma^c)}) \prod_{k=1}^{K} w_k^{n_k} \prod_{j=1}^{n_k} N_{|\gamma|}(\boldsymbol{x}_{j(\gamma)}; \mu_{k(\gamma)}, \Sigma_{k(\gamma)}) \tag{17}$$

where β is a matrix of regression coefficients resulting from the linearity assumption on the expected value of the conditional distribution $p(\boldsymbol{x}_{i(\gamma^c)}|\boldsymbol{x}_{i(\gamma)})$, and where $\eta_{(\gamma^c)}$ and $\Sigma_{(\gamma^c)}$ are the mean and covariance matrix, respectively, of $\boldsymbol{x}_{i(\gamma^c)} - \beta \boldsymbol{x}_{i(\gamma)}$.

Murphy *et al.* (2010) use a similar likelihood formulation in a frequentist approach to variable selection in discriminant analysis.

For the parameters corresponding to the non selected variables it is computationally convenient to use the following conjugate priors:

$$\begin{array}{rcl} \eta_{(\gamma^c)}|\Sigma_{(\gamma^c)} & \sim & N(\mu_{0(\gamma^c)}, h_0\Sigma_{(\gamma^c)}) \\ \beta - \beta_0|\Sigma_{(\gamma^c)} & \sim & \mathcal{N}(H_\gamma, \Sigma_{(\gamma^c)}) \\ \Sigma_{(\gamma^c)} & \sim & IW(\delta, \Omega_{0(\gamma^c)}). \end{array} \tag{18}$$

The corresponding MCMC algorithm benefits of this parametrization, since it is possible to integrate out means, variances and regression coefficients and design Metropolis steps that depend only on the selected and proposed variables. Below we show an application of the model to the analysis of microarray data where, as in Chen *et al.* (2010), we use a MRF prior on γ to capture knowledge on the gene network structure.

3.3. *An Application to Microarray Data*

We analyze the widely used leukemia data of Golub *et al.* (1999) that comprise a training set of 38 patients and a validation set of 34 patients. The training set consists of bone marrow samples obtained from acute leukemia patients while the validation set consists of 24 bone marrow samples and 10 peripheral blood samples. The aim of the analysis is to identify genes whose expression discriminate acute lymphoblastic leukaemia (ALL) patients form acute myeloid leukaemia (AML) patients. Following Dudoit *et al.* (2002) we truncate expression measures beyond the threshold of reliable detection at 100 and 16,000, and remove probe sets with intensities such that $max/min \leq 5$ and $max - min \leq 500$. This leaves us with 3,571 genes for the analysis. The expression readings are log-transformed and each variable is rescaled by its range. The results we report here were obtained by specifying a MRF prior on γ that uses the gene network structure downloaded from the public available data base KEGG. Note that some of the genes do not have neighbors.

This dataset was also analyzed by Kim *et al.* (2006) using a mixture model for cluster analysis. As in Kim *et al.* (2006) we assume that the non significant variables are marginally independent of the significant ones. The hyperparameters are taken to be $\delta = 3$, $h_1 = 10$, $h_0 = 100$, $\Omega_1 = 0.6^{-1} \cdot I_{|\gamma|}$ and $k_0 = 10^{-1}$. We set the hyperparameters of the MRF prior, parameterized according to equation (11), as $d = -2.5$ and $g = 0.5$. Two samplers were started with randomly selected starting models that had 10 and 2 included variables, respectively. We ran 150,000 iterations with the first 50,000 used as burn-in. Final inference was performed by pooling the two chains together.

With a threshold of 0.85 on the marginal probability of inclusion we selected 29 genes that were able to correctly classify 33 of the 34 samples. Lowering the threshold to 0.5 selected 72 significant variables that were able to correctly classify 30 of the 34 patients of the validation set. As described in Golub *et al.* (1999), the validation set includes a much broader range of samples, including samples from peripheral blood rather than bone marrow, from childhood AML patients, and from different reference laboratories that used different sample preparation protocols. Their method made a correct prediction for 29 of the 34 samples and the authors considered this result a "notable success" also because some observations came from one laboratory that used a very different protocol for sample preparation. This suggests that including standardization of sample preparation can lead to even better classification results. In addition, our results indicate that the selection of the top genes is not affected by the different protocol used in one laboratory or by other confounding effects. More insights can be found in Stingo and Vannucci (2010).

Some of the selected genes are already known to be implicated with the differentiation or progression of leukemia cells. For example, Secchiero *et al.* (2005) have already found that cyclooxygenase-2 (COX-2), selected with posterior probability of 0.93, increases tumorigenic potential by promoting resistance to apoptosis and Chien *et al.* (2009) have highlighted the pathogenic role of the Vascular endothelial

growth factor (VEGF)-C, a recognized tumor lymphangiogenic factor, in leukemia via regulation of angiogenesis through upregulation of COX-2. Peterson *et al.* (2007) have found that CD44 gene, selected with posterior probability of 0.98, is involved in the growth and maintenance of the AML blast/stem cells. Jin *et al.* (2006), studying the mechanisms underlying the elimination of leukemic stem cells (LSCs), also identified CD44 as a key regulator of AML LSCs.

4. CONCLUSIONS

We have reviewed Bayesian approaches for variable selection for linear settings and for mixture models and have described novel extensions that aim at addressing important problems in the analysis of genomic data. The Bayesian approaches we have presented offer a coherent framework in which variable selection and prediction, classification or clustering of the samples are performed simultaneously. Bayesian variable selection techniques can cope with a large number of regressors and can handle a number of covariates larger than the sample size. These methods allow the evaluation of the joint effect of sets of variables and the use of stochastic search techniques to explore the high-dimensional variable space. In addition, the flexible prior model allows to incorporate additional information in quite a natural way.

REFERENCES

Bottolo, L. and Richardson, S. (2010). Evolutionary stochastic search for Bayesian model exploration. http://ArXiv.org/abs/1002.2706 .

Brown, P. J. (1993). *Measurement, Regression and Calibration.* Oxford: University Press.

Brown, P. J., Vannucci, M. and Fearn, T. (1998a). Multivariate Bayesian variable selection and prediction. *J. Roy. Statist. Soc. B* **60**, 627–641.

Brown, P. J., Vannucci, M. and Fearn, T. (1998b). Bayesian wavelength selection in multicomponent analysis, *J. Chem.* **12**, 173–182.

Brown, P. J., Vannucci, M. and Fearn, T. (2002). Bayes model averaging with selection of regressors. *J. Roy. Statist. Soc. B* **64**, 519–536.

Cai, B. and Dunson, D. (2006). Bayesian covariance selection in generalized linear mixed models. *Biometrics* **62**, 446–457.

Carlin, B. P. and Chib, S. (1995). Bayesian model choice via Markov chain Monte Carlo methods. *J. Roy. Statist. Soc. B* **57**, 473–484.

Chen, M.-H., Ibrahim, J., Shao, Q.-M. and Weiss, R. (2003). Prior elicitation for model selection and estimation in generalized linear mixed models. *J. Statist. Planning and Inference* **111**, 57–76.

Chen, Y., Stingo, F., Tadesse, M. and Vannucci, M. (2010). Incorporating biological information in Bayesian models for the selection of pathways and genes. *Ann. Appl. Statist.*, (invited revision).

Chen, Z. and Dunson, D. (2003). Random effects selection in linear mixed models. *Biometrics* **59**, 762–769.

Chien, M., Ku, C., Johansson, G., Chen, M., Hsiao, M., Su, J., Inoue, H., Hua, K., Wei, L. and Kuo, M. (2009). Vascular endothelial growth factor-c (vegf-c) promotes angiogenesis by induction of cox-2 in leukemic cells via the vegf-r3/jnk/ap-1 pathway. *Carcinogenesis* **30**, 2005–2013.

Chipman, H., George, E. I. and McCulloch, R. (2001). The practical implementation of Bayesian model selection. *Model Selection*, IMS, 67–116.

Clyde, M., DeSimone, H. and Parmigiani, G. (1996). Predition via orthogonalized model mixing. *J. Amer. Statist. Assoc.* **91**, 1197–1208.

Clyde, M. and George, E. I. (2004). Model uncertainty. *Statist. Science* **19**, 81–94.

Cowell, R., Dawid, A., Lauritzen, S. and Spiegelhalter, D. (1999) *Probabilistic Networks and Expert Systems*. Berlin: Springer.

Cui, W. and George, E. I. (2008). Empirical Bayes vs. fully Bayes variable selection. *J. Statist. Planning and Inference* **138**, 888–900.

Dawid, A. (1981). Some matrix-variate distribution theory: Notational considerations and a Bayesian application. *Biometrika* **68**, 265-274.

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

Dudoit, S., Fridlyand, J. and Speed, T. (2002). Comparison of Discrimination Methods for the Classification of Tumors using Gene Expression Data. *J. Amer. Statist. Assoc.* **97**, 77–87.

Fearn, T., Brown, P. and Besbeas, P. (2002). A Bayesian decision theory approach to variable selection for discrimination. *Statist. Computing* **12**, 253–260.

George, E. I. and McCulloch, R. (1993). Variable selection via Gibbs sampling. *J. Amer. Statist. Assoc.* **85**, 398–409.

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

Geweke, J. (1996). Variable selection and model comparison in regression. *Bayesian Statistics 5* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) Oxford: University Press, 609–620.

Giudici, P. and Green, P. (1999). Decomposable graphical Gaussian model determination. *Biometrika* **86**, 785–801.

Golub, T., Slonim, D., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J., Coller, H., Loh, M., Downing, J., Caligiuri, M., Bloomfield, C. and Lander, E. (1999). Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. *Science* **286**, 531–537.

Gustafson, P. and Lefebvre, G. (2008). Bayesian multinomial regression with class specific predictor selection. *Ann. Appl. Statist.* **2**, 1478–1502.

Hans, C., Dobra, A. and West, M. (2007). Shotgun stochastic search for "large p" regression. *J. Amer. Statist. Assoc.* **102**, 507–516.

Hoff, P. (2006). Model-based subspace clustering. *Bayesian Analysis* **1**, 321–344.

Holmes, C. and Held, L. (2006). Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Analysis* **1**, 145–168.

Ibrahim, J., Chen, M.-H. and Ryan, L. (2000). Bayesian variable selection for time series count data. *Statistica Sinica* **10**, 971–987.

Jin, L., Hope, K., Zhai, Q., Smadja-Joffe, F. and Dick, J. (2006). Targeting of cd44 eradicates human acute myeloid leukemic stem cells. *Nature Medicine* **12**, 1167–1164.

Jones, B., Carvalho, C., Dobra, A., Carter, C. H. and West, M. (2005). Experiments in stochastic computation for high-dimensional graphical models. *Statist. Science* **20**, 388–400.

Kim, S., Tadesse, M. G. and Vannucci, M. (2006). Variable selection in clustering via Dirichlet process mixture models. *Biometrika* **93**, 877–893.

Kwon, D., Tadesse, M., Sha, N., Pfeiffer, R. and Vannucci, M. (2007). Identifying biomarkers from mass spectrometry data with ordinal outcomes. *Cancer Informatics* **3**, 19–28.

Laird, N. and Ware, J. (1982). Random effects models for longitudinal data. *Biometrics* **38**, 963–974.

Leamer, E. (1978). Regression selection strategies and revealed priors. *J. Amer. Statist. Assoc.* **73**, 580–587.

Li, F. and Zhang, N. (2010). Bayesian variable selection in structured high-dimensional covariate space with application in genomics. *J. Amer. Statist. Assoc.* , (to appear).

Liang, F., Paulo, R., Molina, G., Clyde, M. and Berger, J. (2008). Mixture of g priors for Bayes variable section. *J. Amer. Statist. Assoc.* **103**, 410–423.

Liu, J., Zhang, J., Palumbo, M. and Lawrence, C. (2003). Bayesian clustering with variable and transformation selections. *Bayesian Statistics 7* (J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith and M. West, eds.) Oxford: University Press , 249–275.

Madigan, D. and Raftery, A. (1994). Model selection and accounting for model uncertainty in graphical models using Occamís window. *J. Amer. Statist. Assoc.* **89**, 1535–1546.

Madigan, D. and York, J. (1995). Bayesian graphical models for discrete data. *Internat. Statist. Rev.* **63**, 215–232.

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models* (2nd ed.) London: Chapman and Hall.

McLachlan, G. and Basford, K. (1988). *Mixture Models: Inference and Applications to Clustering* New York: Marcel Dekker.

Mitchell, T. J. and Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *J. Amer. Statist. Assoc.* **83**, 1023–1036.

Monni, S. and Tadesse, M. G. (2009). A stochastic partitioning method to associate high-dimensional datasets. *Bayesian Analysis* **4**, 413–436.

Murphy, T., Dean, N. and Raftery, A. (2010). Variable selection and updating in model-based discriminant analysis for high dimensional data with food authenticity applications. *Ann. Appl. Statist.* **4**, 396–421.

Ntzoufras, I., Dellaportas, P. and Forster, J. (2003). Bayesian variable and link determination for generalised linear models. *J. Statist. Planning and Inference* **111**, 165–180.

Peterson, L., Wang, Y., Lo, M., Yan, M., Kanbe, E. and Zhang, D. (2007). The multifunctional cellular adhesion molecule cd44 is regulated by the 8;21 chromosomal translocation. *Leukemia* **21**, 2010–2019.

Raftery, A. (1996). Approximate Bayes factors and accounting for model uncertainty in generalized linear models. *Biometrika* **83**, 251–266.

Raftery, A. and Dean, N. (2006). Variable selection for model-based clustering. *J. Amer. Statist. Assoc.* **101**, 168-178.

Raftery, A., Madigan, D. and Hoeting, J. (1997). Bayesian model averaging for linear regression models. *J. Amer. Statist. Assoc.* **92**, 179–191.

Richardson, S. and Green, P. (1997). On Bayesian analysis of mixtures with an unknown number of components. *J. Roy. Statist. Soc. B* **59**, 731–792 (with discussion).

Roverato, A. (2002), Hyper inverse Wishart distribution for nondecomposable graphs and its application to Bayesian inference for Gaussian graphical models. *Scandinavian J. Statist.* **29**, 391–411.

Scott, J. and Berger, J. (2010). Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *Ann. Statist.* **38**, 2587–2619.

Secchiero, P., Barbarotto, E., Gonelli, A., Tiribelli, M., Zerbinati, C., Celeghini,C., Agostinelli, C., Pileri, S. and Zauli, G. (2005). Potential pathogenetic implications of cyclooxygenase-2 overexpression in b chronic lymphoid leukemia cells. *The American Journal of Pathology* **167**, 1559–607.

Sha, N., Tadesse, M. G. and Vannucci, M. (2006). Bayesian variable selection for the analysis of microarray data with censored outcome. *Bioinformatics* **22**, 2262–2268.

Sha, N., Vannucci, M., Brown, P., Trower, M., Amphlett, G. and Falciani, F. (2003). Gene selection in arthritis classification with large-scale microarray expression profiles. *Comp. and Funct. Gen.* **4**, 171–181.

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

Smith, M. and Kohn, R. (1996). Nonparametric regression using Bayesian variable selection. *J. Econometrics* **75**, 317–343.

Stingo, F., Chen, Y., Vannucci, M., Barrier, M. and Mirkes, P. (2010). A Bayesian graphical modeling approach to microRNA regulatory network inference. *Ann. Appl. Statist.*, (to appear).

Stingo, F. and Vannucci, M. (2010). Variable selection for discriminant analysis with Markov Random Field priors for the analysis of microarray data. (submitted).

Tadesse, M. G., Sha, N. and Vannucci, M. (2005). Bayesian variable selection in clustering high dimensional data. *J. Amer. Statist. Assoc.* **100**, 602–617.

Telesca, D., Muller, P., Parmigiani, G. and Freedman, R. (2008). Modeling dependent gene expression. *Tech. Rep.*, University of Texas M.D. Anderson Cancer Center.

Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistic*. New York: Wiley.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti* (P. K. Goel and A. Zellner, eds.) Amsterdam: North-Holland, 233–243.

DISCUSSION

CARLO BERZUINI (*University of Cambridge, UK*)

The paper by Vannucci and Stingo (hereafter VS) is an excellent review of Bayesian variable selection methods in linear regression and graphical model selection, with significant elements of innovation. VS show the relevance of the methods within the current research effort to elucidate molecular mechanisms at the basis of pathogenesis. The following discussion should not be interpreted as a criticism of the approach, as much as an attempt to highlight possible difficulties encountered in its application, in the hope that the points raised may inspire possible enhancements of an already extremely useful method. I shall organize the discussion around the following themes: reverse causation, tissue-specificity and variability of gene expression, MRF prior, MCMC *vs* particles, epistatic interaction and systematic *vs* focused.

Reverse causation

VS analyze Golub's data, where gene expression levels may reflect the general signalling pattern that has *caused* the observed onset of leukaemia, but also *have been influenced by* the specific type of leukaemia developed. In such a "reverse causation" situation, any detected expression difference may represent a *reaction*, as well as a *cause*, of the disease, or both. The MRF prior proposed by the authors will not help here. The impact of a possible reverse causation effect on the conclusions of the study will have to be evaluated in the light of the intended scientific target. Is this to predict disease before it occurs? Is it to get some clues about the pathways that are *causally* involved in pathogenesis? Or is it to predict the future evolution of patients observed when they have reached a given disease stage?

Suppose we wish to bias the model search process towards the inclusion of variables that are supported as causal with respect to the studied disease. A number of strategies may then be adopted. One is to incorporate in the analysis information from large scale experimental studies which interrogate genes individually with respect to their impact on disease-related phenotypes. One example being *in vitro* experiments where *si*RNAs are designed to target a collection of candidate genes and to test the effects of knocking down these genes, one by one, on some disease-related trait. Such (high-throughput) tests should highlight those genes whose expression

level is causally related to disease. Inclusion of these genes into the model should then, in some way, be favored.

Another possibility, which does not require experiments, is to incorporate in the analysis extra information which, under suitable "natural randomization" assumptions, helps discriminate "reactive" from "causal" hits. One way of doing this is to exploit the fact that the expression levels of genes are regulated by specific loci located in gene-specific *regulatory regions* of DNA, typically located in proximity of the gene they regulate. These regions are often well known. One idea is then to incorporate in the analysis *genotypic* information at the *regulating loci* of the studied expression levels. Ideally, each measured expression level, E, would then be accompanied in the model by the genotype G of the corresponding (experimentally verified) regulating locus. The estimated structure of the dependencies between G, E and Y might then point at the underlying causal structure. For example, if both E and G happen to be marginally associated with Y we may safely exclude that the (E, Y) association is reactive. There is, of course, also the possible confounding of the relationship between E and Y. Principles of mendelian randomization and principal stratification methods may be relevant here. I would be strongly tempted to think of an MCMC (or particle) algorithm which incorporates the above ideas to bias the search for an optimal set of regressors towards models where predictors are supported as causal with respect to the predictand.

Tissue-specificity and variability of gene expression

There is a further reason why I am concerned about using exclusively gene expression levels as regressors. The activity displayed by a gene may often strongly depend on the cell type, the so called *tissue specificity* of gene expression. By contrast, genotype-disease relationships do not vary across tissues. In addition gene expression levels, unlike genotypes, are subject to considerable "measurement error" due to their dependence on such factors as time of the day, laboratory and patient conditions. This is why, in certain circumstances, I would expect a better predictive accuracy to be obtained when gene expression information is accompanied in the model by genotypic information at the corresponding regulating loci, as stated previously. This point is illustrated by the final example in this discussion. As a final, and slightly different type of, consideration we would suggest that the methods discussed by VS could be used to systematically address dependence of disease risk on *both* gene expression and cell type.

MRF prior

The proposed MRF prior lets the variable selection process be guided by prior knowledge about the way genes are functionally organized into groups, or pathways. The underlying assumption here being that the selected variables are more likely true positives if they cluster together within the same pathway. I think this is a wonderful idea, not least because one would expect genes in the same pathway to be co-regulated, and therefore the device of biasing search towards pathway-homogeneous clusters of predictors increases our chances to find the disease-relevant co-regulation patterns. Concerning in particular the use of KEGG, where a pathway is represented by a collection of nodes and (directed or undirected) edges connecting them, I have perhaps some concern about the extent to which KEGG topology is captured by the proposed MRF prior. Finally, how important is, in practice, the MRF prior suggested by VS Does it lead to models with greater predictive accuracy? Or to models that are biologically more meaningful? Has this been assessed empirically?

MCMC vs *particles*

VS use the above discussed MRF prior to incorporate aspects of the structure of KEGG (or of any other relevant biological net). I have two questions. Would it be sensible to

- use the same information also to implement smart MCMC moves?
- 'pretend" that sample individuals arrive sequentially, one after another, and deal with this via sequential Monte Carlo updating of the posterior, using *particles*?

Also, I wonder whether these two ideas may be combined in some way. Each 'particle" would, at any stage of the sequential process, represent a particular selection of covariates and a particular realization of the regression coefficients associated with these covariates. Changes in the posterior, induced by incoming individuals, would be reflected by corresponding changes in the distribution of the particles, possibly involving Metropolis-Hastings particle moves within the parameters' space, and jumps across the model space (each particular selection of covariates being called a "model"). Sequential updating opens the door to prequential model assessment, in the sense of Philip Dawid, and may be used to prevent the waste of samples for pure purpose of testing. Moreover each particle, corresponding to a particular selection of covariates, could be "expanded" to contain the *biological pathway locations* of those covariates. At any stage of the sequential updating process, the current *ensemble* of particles would thus resemble a "map" of interesting areas within the studied pathways, and this map could be used to inform clever cross-model move proposals, designed to allow each individual particle to quickly reach biologically interesting areas of the model space.

Epistatic interaction

VS restrict attention to regression models based on main effects. I confess I am afraid that the exclusion of interaction terms might prevent the model from capturing an important (if not the most important) part of the biological logic we are studying. And for a number of reasons. Biologists have recently become aware of the central role of *epistatic* interactions between genes in the complex architecture of cellular systems. The term "epistasis" has come to describe various types of gene×gene (or gene×expression) interactions that have a biological explanation. The explanation may be *functional*, for example, when the interaction reflects the interplay and the mutual compensation relationships that occur between proteins that bind together into a complex. Or *compositional*, for example when the effect of one gene, A say, is *blocked* by a mutation in another gene, B. This may occur if A operates downstream of B in a common pathway, and if the B-mutation causes the downstream part of the pathway to collapse, thereby causing genetic/expression variation at A to be no longer longer relevant to biological function. Hence we should not lightheartedly ignore the potential importance of expression×expression, expression×genotype, and genotype×genotype interaction terms in the model. In the presence of readily available database information (KEGG, GO, etc.), we could consider for inclusion in the model also any interaction term (for example, gene×gene) which could be supported by evidence of epistasis. For example, the fact that two proteins are known from KEGG to interact with each other suggests that there may be epistasis between the corresponding genetic markers.

Of course, not any statistically significant interaction term will be interpretable as evidence of some form of epistasis. However Philip Dawid and I have developed a formal theory of epistatic interaction.In the general case of two causal factors, A and B, defined over quite general spaces, we make the "deep determinism" assumption that the binary disease phenotype Y can be expressed as $Y = f(A, B, U)$, where U is an unobserved "context" variable, and f is a *deterministic* function taking values 0 and 1 only, independent of how the causal factors of interest A and B are generated, be it interventionally or observationally. In this setting, we say that "A and B interact biologically" if there is some u^* and values (a^*, b_0, b_1) such that changing the value of B from b_0 to b_1, when A remains fixed to a^*, changes the value of Y, and that the same is true when we interchange A and B, a and b, and allow a different context, u^{**}. Let α [resp., β] be a given set of possible values for A [resp., B]. We also re-use α [resp., β], to denote the truth-value (0 or 1) of the event $A \in \alpha$ [resp., $B \in \beta$]. Define, for $i, j = 0, 1$, the *observational* risk

$$R_{ij} := \Pr(Y = 1 \mid \alpha = i, \beta = j), \tag{19}$$

directly estimable (under prospective or cross-sectional sampling) from observed proportions in data. Then, if $\Pr(A \in \alpha) < 0.5$ and $\Pr(B \in \beta) < 0.5$, the condition

$$R_{11} - R_{10} - R_{01} > 0 \tag{20}$$

implies that A and B interact biologically in the previously state sense.

Systematic vs *focused*

It is perhaps appropriate to describe VS's approach as *systematic*, in that is is based on the application of comprehensive, overarching, models to (often observational) data generated by high-throughput platforms. This should be contrasted to the *focused* approach of so many research groups in biology, who concentrate their efforts on a specific disease mechanism involving a tiny portion of a relevant pathway, and proceed by a self-adapting sequence of small observational studies, experiments and bioinformatic investigations. The two approaches would have to be integrated in some way.

In the remaining part of this discussion, I would like to provide an example of the "focused" approach, by drawing on a study by N. Marziliano, M.F. Notarangelo, P.A. Merlini, S. Veronese, F. Orsini, D. Lina, D. Ardissino and myself, on human atherosclerosis. This example will also illustrate some of the above points, notably the importance of incorporating genotypic information and information at multiple phenotypic levels.

Recent association studies have highlighted an association between genetic variants tagged by SNP rs1333040 in chromosomal region 9p21.3 and ischemic heart disease. Subsequently, knock out mice experiments have provided evidence that 9p21.3 is involved in normal cardiac expression of cell cycle inhibitor gene CDK2NB, thus suggesting that genetic variation closely tagged by rs1333040 has a regulatory effect on that gene, and supporting the hypothesis that CDK2NB disregulation is part of the mechanism through which genetic variation in 9p21.3 affects risk of coronary artery disease. In the light of this, we have chosen a new disease-related subphenotype, indicating whether the patient suffer from angina only during physical exercise (Y="S") or also at rest (Y="A"), these two categories being associated with different risk of infarction. Define the pair (G, E) where G represents the

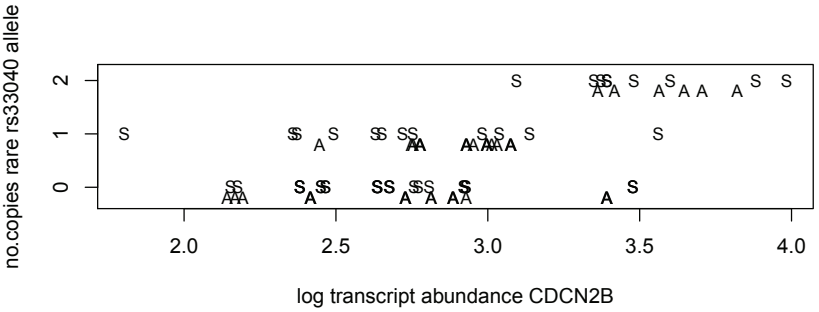

Figure 1: *Data for the illustrative example*

rs1333040 genotype and E represents the expression of gene CDK2NB. (Y, G, E) data have been collected on a sample of patients, and summarized in Figure 1. Regression of E on G confirms that rs1333040 is a strong regulator of the expression of CDK2NB also in humans. Logistic regression of Y upon E provides no evidence of an association between Y and the level of expression, whereas a logistic regression of Y on G provides statistically significant evidence of a higher risk of $Y =$ "A'' in patients who are heterozygous at rs1333040.

REPLY TO THE DISCUSSION

We thank Carlo for the very thoughtful and stimulating discussion. Many of the points he raises have indeed provided us with the opportunity to think more broadly about possible applications and enhancements of our methodologies. We have organized our rejoinder following the list of themes of Carlo's discussion.

Reverse causation. "Causality" is certainly a question of great interest in genomic studies. In our work, however, we have been more interested in the development of models that can deal with the high dimensionality of the data, therefore functioning more as exploratory models rather than models able to assess causality. In the case studies we describe in this paper the inferential goal is to find sets of genes that are differentially expressed at a given time point, or genes that are able to correctly classify subjects. Our aim is not to establish whether it is the expression of these genes that causes the disease or viceversa. Indeed, our overall objective is the selection of important biomarkers and the identification, for each subject, of the best medical treatment.

Of course, information about important, (or *causally* important) genes, when available, can be incorporated into our prior distibution on the selection indicators, perhaps in a similar manner to what we do in Stingo *et al.* (2010). Indeed, the idea of incorporating in the analysis *genotypic* information at the *regulating loci* of the expression levels under study is very interesting and could lead also to the definition of a model for causal analysis. More and more research is now done on the development of models able to integrate several types of data, which is also a feature of our methodologies. In our modeling strategies, however, we have taken the approach of biasing the search towards promising models by modifying the prior probability model accordingly, rather than modifying the MCMC algorithm as suggested in the discussion.

Tissue-specificity and variability of gene expression. Some of our proposed methodologies can be adapted to the case of genotypic covariates, and we are indeed currently working in this direction. The integration of the cell type information into the model is also an interesting suggestion.

MRF prior. Incorporating biological knowledge into our prior models is one of the most innovative features of our work. First attempts at incorporating biological network information into a probabilistic model, such as in Wei and Li (2007), have adopted the strategy of translating the "functional" network of KEGG into an undirected graphical model, see also Telesca *et al.* (2008) for a more sophisticated approach. However, different approaches are possible relatively to the way that a network is translated into a graphical model (directed, undirected, chain graph, etc...) and also to the way the marginalization with respect to the non observed genes included in the network is performed. We expect future investigations to focus more on these aspects.

In Chen *et al.* (2010) we show how our model, including both pathway scores as regressors and the MRF prior based on the KEGG network, improves on the predictive performances with respect to a model that does not incorporate any biological information. More important, our modeling strategy leads to a better understanding of the biological process, because it allows to find pathways and genes that are related to a particular phenotype, together with an indication of whether these pathways share some of the selected genes and of whether the selected genes are connected in the KEGG network.

MCMC vs *particles.* In Chen *et al.* (2010) we actually use the information about genes grouping into pathways to construct MCMC moves that take into account constraints specified both for the interpretability and the identifiability of the model. Additional details can be found in the paper. The use of particles, as suggested by Carlo, is certainly an interesting idea that would however imply a substantial modification of the inferential algorithm we are using. We have no insights on how such change could effect our posterior inference.

Epistatic interaction. Our models allow to add both interactions and others terms to the list of predictors. In practice, this is possible if the number of additional terms is limited, as the algorithm becomes computationally prohibitive when the number of regressors explodes. In addition, when interactions are added into the model it is necessary to add constraints that exclude the selection of an interaction without selecting the two main effects, see also Chipman (1996).

Systematic vs *focused.* The integration of *systematic* approaches, like ours, with more *focused* ones represents an interesting idea that would enhance the data integration aspect of the proposed methodologies.

ADDITIONAL REFERENCES IN THE DISCUSSION

Chipman, H. (1996). Bayesian variable selection with related predictors. *The Canadian Journal of Statistics*, **24**, 17–36.