

Bayesian Variable Selection in Multinomial Probit Models to Identify Molecular Signatures of Disease Stage

Naijun Sha,¹ Marina Vannucci,^{2,*} Mahlet G. Tadesse,² Philip J. Brown,³ Ilaria Dragoni,⁴
Nick Davies,⁵ Tracy C. Roberts,⁶ Andrea Contestabile,⁷ Mike Salmon,⁸ Chris Buckley,⁸
and Francesco Falciani⁵

¹Department of Mathematical Sciences, University of Texas at El Paso, El Paso, Texas 79968-0514, U.S.A.

²Department of Statistics, Texas A&M University, College Station, Texas 77843-3143, U.S.A.

³Institute of Mathematics and Statistics, University of Kent at Canterbury, Canterbury, Kent CT2 7NF, U.K.

⁴Novartis Institute for Molecular Sciences, London W1B 1BN, U.K.

⁵School of Biosciences, University of Birmingham, Birmingham B15 2TT, U.K.

⁶GlaxoSmithKline, Genomics and Proteomics Science, Stevenage SG1 2NY, U.K.

⁷Department of Human and General Physiology, University of Bologna, 40126 Bologna, Italy

⁸MRC Centre for Immune Regulation, Birmingham University, Birmingham B15 2TT, U.K.

* *email*: mvannucci@stat.tamu.edu

SUMMARY. Here we focus on discrimination problems where the number of predictors substantially exceeds the sample size and we propose a Bayesian variable selection approach to multinomial probit models. Our method makes use of mixture priors and Markov chain Monte Carlo techniques to select sets of variables that differ among the classes. We apply our methodology to a problem in functional genomics using gene expression profiling data. The aim of the analysis is to identify molecular signatures that characterize two different stages of rheumatoid arthritis.

KEY WORDS: Bayesian variable selection; Discrimination; DNA microarrays; Latent variables; MCMC; Multinomial probit model; Truncated sampling.

1. Introduction

In this article, we describe a methodology for the general problem of variable selection in classification, where the response variable is categorical, with two or more categories, and where a large number of predictors, typically much larger than the number of observations, is available. We focus on nominal (i.e., unordered qualitative) responses. Our aim is the classification of samples as well as the identification of important variables characterizing the different classes. We achieve both goals simultaneously by combining multinomial probit models for classification with Bayesian variable selection methods for the identification of important predictors. Following Albert and Chib (1993), we adopt a data augmentation approach to inference, Tanner and Wong (1987), transforming the probit model into a normal linear regression. We build into the model a variable selection mechanism by using mixture priors for the regression coefficients (see George and McCulloch, 1997). Conditional on the latent responses, our model is equivalent to that of Brown, Vannucci, and Fearn (1998a,b) for regression models with multivariate responses. Inference in our model is complicated by the presence of the unknown latent responses. We combine truncated sampling techniques with a Metropolis algorithm to sample from the marginal distribu-

tion of single models. We also explore ways to predict class allocations based on single models as well as model averaging.

We apply our methodology to a problem from functional genomics using microarrays. The recent development of genome-wide technologies has created an unprecedented situation in biology. A significant proportion of an organism's genome can be monitored in single experiments, leading to data characterized by a large number of variables (gene expressions) measured in a relatively small number of experimental conditions. Classification problems, in particular, have received considerable attention, starting from the work of Golub et al. (1999). The goal of the analysis is to find sets of genes that are, for example, related to different kinds of diseases, so that future tissue samples can be correctly classified. Among Bayesian contributions, Ibrahim, Chen, and Gray (2002) proposed a Bayesian univariate selection method, for binary responses only, that primarily models gene expression of individual genes given disease status. A Bayesian approach to dimension reduction in discrimination with probit models for disease status given expression was proposed by West et al. (2000). There, rather than selecting actual genes, as does the method we propose here, a singular-value decomposition is applied to the design matrix to reduce the dimension.

In this article, we study the classification of two groups of rheumatoid arthritis patients at different stages of the disease using gene expression profiling data.

1.1 Case Study: Rheumatoid Arthritis

Rheumatoid arthritis (RA) is an autoimmune disease characterized by chronic synovial inflammation and destruction of cartilage and bone in the joints. At the cellular level the characteristic features of the disease are synovial tissue hyperplasia, and chronic infiltration of immune cells (T and B lymphocytes, monocyte/macrophages, and granulocytes). The hyperplastic layer is mainly composed of fibroblast-like synoviocytes (FLS) and macrophage cells whereas the infiltrating lymphocytes are found in the sublining tissue around the blood vessels. FLS and T lymphocytes (primarily CD4+ memory cells) share the same microenvironment but seem to be spatially separated, suggesting that their interactions must be mediated by diffusible factors. The local production of pro-inflammatory cytokines supports the formation of a clearly defined microenvironment with distinct microarchitectural features, which support the ongoing persistent inflammatory response. Inflammation is also systemic in RA. This means that cells in distant compartments of the body (peripheral blood) will display characteristic features linked to inflammation.

Combining functional genomics technologies with statistical techniques allows us to perform genome-wide searches for multigene predictors of physiological readouts. In Sha et al. (2003), we identified markers for the classification of two forms of arthritis, rheumatoid arthritis and osteoarthritis, which have a similar clinical endpoint but different underlying molecular mechanisms. Here we identify molecular signatures that are predictive of the stage of rheumatoid arthritis, with the goal of understanding how the immune cells in the peripheral blood modify their molecular profiles with the progression of the disease. Our theory allows multiple stages although our illustration is for two stages.

The outline of the article is as follows. Section 2 describes the multinomial probit model with latent variables. Section 3 presents the Bayesian variable selection approach, the posterior analysis, and the prediction strategies. Guidelines for choice of hyperparameters in the prior distributions are also given. Section 4 describes the case study analyses and the biological findings. Section 5 concludes the article.

2. Multinomial Probit with Latent Variables

Multinomial models are commonly used in the social and biological sciences for the analysis of categorical response variables; see Agresti (1990), Chapter 9, among others. Here we introduce the model from the point of view of data augmentation, i.e., using latent variables, as in Albert and Chib (1993). Let (\mathbf{Z}, \mathbf{X}) indicate the observed data, with $\mathbf{X}_{n \times p}$ the predictor matrix and $\mathbf{Z}_{n \times 1}$ a (categorical) response vector coded as $0, \dots, J - 1$, for J classes. Each outcome z_i is associated with a vector $(p_{i,0}, \dots, p_{i,J-1})$, with $p_{i,j}$ the probability that the i th respondent falls into the j th category.

Let us consider the first category as a “baseline” category. A data augmentation approach to inference introduces latent

data \mathbf{Y} into the problem. For the simple binary probit case, we have

$$y_i = \alpha + x_i' \beta + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2), \quad i = 1, \dots, n. \quad (1)$$

The correspondence between y_i and the binary outcome z_i is

$$z_i = \begin{cases} 0 & \text{if } y_i \leq 0 \\ 1 & \text{otherwise.} \end{cases} \quad (2)$$

It is evident that multiplying α and β by a constant c and σ by the same constant leaves the model unchanged. Thus the constraint $\sigma^2 = 1$ is often used to identify the model.

The latent variable approach can be extended to cope with a multinomial response as follows. Let $\mathbf{Y}_{n \times q}$ with $q = J - 1$, be a latent matrix for the $\mathbf{Z}_{n \times 1}$ observed categorical vector. The element $y_{i,j}$ is the unobserved propensity of the i th subject to belong to the j th class. Let us assume a multivariate normal distribution for \mathbf{Y} with common covariance across the different groups

$$\mathbf{Y}_i = \alpha' + \mathbf{X}_i' \mathbf{B} + \epsilon_i, \quad \epsilon_i \sim N(0, \Sigma), \quad i = 1, \dots, n \quad (3)$$

with $\mathbf{Y}_i = (y_{i,1}, \dots, y_{i,J-1})$ the row vector of \mathbf{Y} corresponding to the i th subject. The relationship between z_i and the unobserved \mathbf{Y}_i becomes

$$z_i = \begin{cases} 0 & \text{if } \max_{1 \leq k \leq J-1} \{y_{i,k}\} \leq 0 \\ j & \text{if } \max_{1 \leq k \leq J-1} \{y_{i,k}\} > 0 \text{ and } y_{i,j} = \max_{1 \leq k \leq J-1} \{y_{i,k}\}. \end{cases} \quad (4)$$

Obviously, in the special case $J = 2$ we obtain the univariate latent model (1) and (2) for binary responses. Also by the same argument as in that case, there is a scalar indeterminacy in the model for each response. Possible ways to address the problem are to prespecify a single parameter in the regression for each response or to set the diagonal of Σ as the identity matrix, but at the expense of losing the general structure.

3. Bayesian Variable Selection

We are interested in situations with a large number of predictors, typically much greater than the sample size, $p \gg n$, and where it is desirable to reduce dimensionality. The use of latent variables has allowed us to write a multinomial model in the form of a linear regression. In this framework we can develop a Bayesian variable selection approach that uses mixture priors for the regression coefficients, similar to the variable selection methods in multivariate regression of Brown et al. (1998a,b).

Without loss of generality, we assume in the sequel that \mathbf{X} has been centered, so that the columns of \mathbf{X} have entries that sum to zero. Thus $\text{rank}(\mathbf{X}) \leq \min\{n - 1, p\}$.

3.1 Prior Distributions

Using the notation of Dawid (1981), conditionally on α , \mathbf{B} , and Σ , the standard multivariate normal regression model (3) can be written as

$$\mathbf{Y} - \mathbf{1}\alpha' - \mathbf{X}\mathbf{B} \sim \mathcal{N}(\mathbf{I}_n, \Sigma) \quad (5)$$

with \mathbf{Y} an $n \times q$ random matrix, $\mathbf{1}$ an $n \times 1$ vector of 1's, \mathbf{X} the $n \times p$ design matrix, regarded as fixed, and \mathbf{B} the $p \times q$ matrix of regression coefficients. In the notation $\mathcal{N}(\cdot, \cdot)$ both arguments are proportional to covariance matrices, by rows

and by columns, respectively. This avoids the need to string out matrices as vectors and the use of Kronecker products for covariance.

Conjugate priors for the parameters α , \mathbf{B} , and Σ are discussed in Brown et al. (1998a). Variable selection is achieved through the introduction of a binary p -vector γ with j th element γ_j either 1 or 0 according to whether the j th variable is included or not in the model. For this selection prior each column of \mathbf{B} has a singular p -variate distribution and, given γ , selecting the variables with $\gamma_j = 1$ gives

$$\mathbf{B}_\gamma - \mathbf{B}_{0\gamma} \sim \mathcal{N}(\mathbf{H}_\gamma, \Sigma), \tag{6}$$

where \mathbf{B}_γ and \mathbf{H}_γ are just \mathbf{B} and \mathbf{H} with the rows and, in the case of \mathbf{H} , columns for which $\gamma_j = 0$ deleted. We also have $\alpha' - \alpha'_0 \sim \mathcal{N}(h, \Sigma)$ and $\Sigma \sim \mathcal{IW}(\delta; \mathbf{Q})$, with α and \mathbf{B} independent of each other. The notation $\mathcal{IW}(\delta; \mathbf{Q})$, with shape parameter $\delta = n - q + 1$, indicates the inverse Wishart distribution with n degrees of freedom and q dimensions. The simplest form of the prior distribution for γ is $\pi(\gamma | w) = w^{p_\gamma} (1 - w)^{p - p_\gamma}$ where p_γ indicates the number of chosen variables, i.e., the number of ones in γ . A further Beta prior distribution can be imposed on w .

3.2 Posterior Inference

In our model the observed data are (\mathbf{Z}, \mathbf{X}) , the parameters of interest are $\theta = (\alpha, \mathbf{B}, \Sigma, \gamma)$, and \mathbf{Y} is a matrix of unobserved latent variables. Here, with variable selection being our main focus, we implement a fast inference scheme by deriving the posterior distribution of γ given (\mathbf{Y}, \mathbf{X}) , essentially integrating out α , \mathbf{B} , and Σ from the joint posterior. Conditional on the latent responses, our model is equivalent to that of Brown et al. (1998a) for regression models with multivariate responses. Inference in our model, however, is complicated by the presence of the unknown latent variables. We use the following MCMC procedure:

- The latent matrix $\mathbf{Y}(n \times q)$ is treated as missing and imputed from its marginal distribution. First, conditionally on Σ and γ , we have $\mathbf{Y} - \mathbf{1}\alpha'_0 - \mathbf{X}_\gamma \mathbf{B}_{0\gamma} \sim \mathcal{N}(\mathbf{P}_\gamma, \Sigma)$ with $\mathbf{P}_\gamma = \mathbf{I}_n + h\mathbf{1}\mathbf{1}' + \mathbf{X}_\gamma \mathbf{H}_\gamma \mathbf{X}'_\gamma$. Averaging over Σ , setting $\alpha_0 = 0$ and $\mathbf{B}_0 = 0$, and letting h large, under constraint (4) we have the matrix Student distribution, in the notation of Brown (1993),

$$\mathbf{Y} | (\gamma, \mathbf{X}, \mathbf{Z}) \sim \mathcal{T}(\delta; \mathbf{P}_\gamma, \mathbf{Q}), \tag{7}$$

or a truncated normal distribution under the constraint (2) for the binary case. Samples $\mathbf{Y}(n \times q)$ from these truncated distributions can be drawn using a sub-chain Gibbs sampler on univariate full conditionals. See Geweke (1991) for the optimized exponential rejection sampling method.

- The vector γ can be drawn from the posterior distribution of γ given $(\mathbf{Y}, \mathbf{X}, \mathbf{Z})$, as in equation (20) of Brown et al. (1998a)

$$\begin{aligned} \pi(\gamma | \mathbf{Y}, \mathbf{X}, \mathbf{Z}) &\propto g(\gamma) \\ &= \pi(\gamma) | \mathbf{I}_n + \mathbf{X}_\gamma \mathbf{H}_\gamma \mathbf{X}'_\gamma |^{-q/2} | \mathbf{Q}_\gamma |^{-(\delta+n+q-1)/2} \end{aligned} \tag{8}$$

with $\mathbf{Q}_\gamma = \mathbf{Q} + \mathbf{Y}'(\mathbf{I} - \mathbf{X}_\gamma \mathbf{K}_\gamma^{-1} \mathbf{X}'_\gamma) \mathbf{Y}$ and $\mathbf{K}_\gamma = \mathbf{X}'_\gamma \mathbf{X}_\gamma + \mathbf{H}_\gamma^{-1}$ and where \mathbf{X} and \mathbf{Y} are both centered. The distribution (8) is not of a known form and γ can be sampled using a Metropolis algorithm as in Brown et al.

(1998b). The method visits a sequence of models that differ successively in one or two variables. At a generic iteration, given the previous visited vector, the algorithm randomly chooses among a set of transition moves, by adding or deleting a variable or swapping two variables. A fast updating scheme that suitably augments the data and uses QR decompositions, with QR-deletion or addition algorithms to remove or add single columns, is developed in Brown, Vannucci, and Fearn (2002).

3.3 Classification of Future Cases

The missing value \mathbf{Y} can be imputed using the mean of all \mathbf{Y} 's sampled during the MCMC. Let us indicate the estimate of \mathbf{Y} as $\hat{\mathbf{Y}}$. The normalized conditional posterior probabilities $\pi(\gamma | \hat{\mathbf{Y}}, \mathbf{X}, \mathbf{Z})$ for all distinct γ 's visited by the MCMC can be easily computed. Marginal probabilities of inclusion of single variables $\pi(\gamma_j = 1 | \hat{\mathbf{Y}}, \mathbf{X}, \mathbf{Z}), j = 1, \dots, p$, can be derived from these posterior probabilities. Various prediction methods are then possible.

Let us assume we have available n_f further measurements $\mathbf{X}_f(n_f \times p)$ for which we want to predict the corresponding \mathbf{Y} -vectors. We center \mathbf{X}_f at the training means. For a given γ , we can evaluate the joint distribution of Y and Y_f and then use properties of the matrix- T distribution to derive the conditional distribution of Y_f given Y . A single model prediction is

$$\hat{\mathbf{Y}}_f = \mathbf{1}\hat{\alpha}' + \mathbf{X}_{f(\gamma)} \tilde{\mathbf{B}}_\gamma, \tag{9}$$

where γ can be chosen as the model with the highest posterior probability among those visited, and where $\hat{\alpha} = \hat{\mathbf{Y}}$ and $\tilde{\mathbf{B}}_\gamma = \mathbf{K}_\gamma^{-1} \mathbf{X}'_\gamma \hat{\mathbf{Y}}$. Alternatively, the Bayesian approach to variable selection allows prediction via model averaging, by using a set of a posteriori likely models, Madigan and Raftery (1994) and Brown et al. (1998a), as

$$\hat{\mathbf{Y}}_f = \sum_\gamma (\mathbf{1}\hat{\alpha}' + \mathbf{X}_{f(\gamma)} \tilde{\mathbf{B}}_\gamma) \pi(\gamma | \hat{\mathbf{Y}}, \mathbf{X}, \mathbf{Z}). \tag{10}$$

Recent developments in Bayesian model averaging include methods that incorporate variable selection for prediction, Brown et al. (2002), and single model approximations, Barbieri and Berger (2004).

Having obtained an estimate of \mathbf{Y}_f , the corresponding predicted categorical value for the i th future observation can be computed via (4).

In practice, the predictive performance of the selected models is typically assessed by splitting the available data into a training and a validation set, fitting the model on the training data and using the validation set to compute mean-squared errors of the prediction estimates obtained as described above. In the case study presented here, however, we have a limited number of samples available, which is typical of experiments involving expression level profiling. We therefore resort to sampling-based methods for crossvalidation prediction (Gelfand, 1996). In the binary case, let $\mathbf{Z}_{(i)}$ be the vector \mathbf{Z} without the i th element. A crossvalidation predictive probability can be calculated as

$$\begin{aligned} P(Z_i = 1 | \mathbf{X}, \mathbf{Z}_{(i)}) &= \int_\gamma \int_{\mathbf{Y}} P(Z_i = 1 | \mathbf{X}, \mathbf{Z}_{(i)}, \gamma, \mathbf{Y}) p(\gamma, \mathbf{Y} | \mathbf{X}, \mathbf{Z}_{(i)}) d\gamma d\mathbf{Y}. \end{aligned}$$

By using $p(\gamma, \mathbf{Y} | \mathbf{X}, \mathbf{Z})$ as importance sampling density for $p(\gamma, \mathbf{Y} | \mathbf{X}, \mathbf{Z}_{(i)})$ we have

$$\begin{aligned} \hat{P}(Z_i = 1 | \mathbf{X}, \mathbf{Z}_{(i)}) &= \frac{1}{M} \sum_{t=1}^M P(Y_i > 0 | \mathbf{X}, \mathbf{Z}_{(i)}, \gamma^{(t)}, \mathbf{Y}^{(t)}) \\ &= \frac{1}{M} \sum_{t=1}^M \Phi(\tilde{\alpha}^{(t)'} + \mathbf{X}'_{i(\gamma^{(t)})} \tilde{\mathbf{B}}_{\gamma}^{(t)}) \end{aligned} \quad (11)$$

with $\tilde{\alpha}^{(t)} = \bar{\mathbf{Y}}^{(t)}$, $\tilde{\mathbf{B}}_{\gamma}^{(t)} = \mathbf{K}_{\gamma^{(t)}}^{-1} \mathbf{X}'_{\gamma^{(t)}} \mathbf{Y}^{(t)}$, $\mathbf{X}_{i(\gamma^{(t)})}$ being the measurements for the i th individual at the variables selected by $\gamma^{(t)}$. $\mathbf{Y}^{(t)}$ and $\gamma^{(t)}$ are the MCMC samples at the t^{th} iteration and $\Phi(\cdot)$ is the normal CDF.

3.4 Hyperparameter Settings

Generally, we set proper prior distributions to cope with the indeterminacy in the model and high correlations in the columns of \mathbf{X} . Our aim is to provide priors for a variety of similar settings, and so we choose suitably vague parameter values whenever possible with impunity. We leave harder choices of informative priors to a form of bracketing suggested by the likely prior range.

A vague prior can be assigned to the intercept parameter vector α by specifying the hyperparameter h as a large value tending to ∞ , so that the value ascribed to the prior mean α_0 becomes irrelevant. We choose $\alpha_0 = 0$. We set $\mathbf{B}_0 = 0$. The prior distribution for \mathbf{B} given γ depends on the matrix \mathbf{H}_{γ} . Brown et al. (2002) discuss relative merits and drawbacks of different specifications, such as $\mathbf{H} = c \text{Diag}((\mathbf{X}'\mathbf{X})^{-})$, a diagonalized version of a full g-prior $\mathbf{H} = c(\mathbf{X}'\mathbf{X})^{-}$, and a simpler $\mathbf{H} = cI$. The latter we would recommend in general by the arguments of that paper and because it is easier to calibrate, as seen below.

Some care in the choice of the parameter c is needed and we offer some guidelines for the $H = cI$ case. We can imagine orthogonally transforming the prior and model so that there are just $r = \text{rank}(\mathbf{X})$ nonzero parameters. The $r \times r$ covariance matrix for the data is proportional to the diagonal matrix $\text{Diag}[1/\lambda_1, \dots, 1/\lambda_r]$. The prior precision (inverse of variance) for the i th parameter is $1/c$; the posterior precision is $1/c + \lambda_i$. Thus the *total* relative precision of prior to posterior is, using *trace* of the precision matrix as the total information, $(r/c)/(r/c + \sum_{i=1}^r \lambda_i)$ or $(1/c)/(1/c + \bar{\lambda})$. The range of c is implied by the ratio of prior to posterior precision being between say 0.1 and 0.005, that is

$$c^*(\bar{\lambda}, 0.1) < c < c^*(\bar{\lambda}, 0.005), \quad (12)$$

where $c^*(\lambda, p) = (1 - p)/(p\lambda)$. The parameter c in fact regulates the amount of shrinkage in the model (as does the prior on the number of nonzero regression coefficients). Indeed $1/c$ is like the ridge parameter in ridge regression (see Hoerl and Kennard, 1970). This guideline range we contend avoids too much regularization, as well as large values, that could induce nonlinear shrinkage as a result of Lindley's paradox (Lindley, 1957). At least this will be the case provided the nonzero eigenvalues are not very different. In cases where the condition number (largest to smallest nonzero eigenvalue) is say more than 100 we would recommend focusing more on ill-estimated parameters and increasing the upper bound on c with $1/c^*$ set to say the lower decile of the $n - 1$ nonzero

eigenvalues (the point such that 10% of eigenvalues are less than it). Thus (12) becomes more generally

$$c^*(\bar{\lambda}, 0.1) < c < \max\{c^*(\bar{\lambda}, 0.005), c^*(\lambda_{0.1}, 0.5)\}. \quad (13)$$

4. Case Study

4.1 Experimental Protocols

We have data available from 20 rheumatoid arthritis patients, recruited from the rheumatology clinic at the Addenbrooks Hospital in Cambridge, U.K. The patients were representative of an early (11 patients with disease duration less than 2 years) and late stage of the disease (9 patients with over 15 years of disease duration). The erythrocytes sedimentation rate (ESR) of each patient was measured as a general indicator of infection.

The experimental study was performed using custom-made nylon high-density arrays with 999 cDNA clones representative of the major functional categories (i.e., cell cycle, adhesion molecules, apoptosis, cytoskeleton, extracellular matrix, homeostasis, cytokines, growth factors, homeobox, inflammation, lipids, shear stress, signal transduction, and transcription factors). Each clone was replicated four times on the array. The mRNA was extracted from peripheral blood using the mRNA Isolation Kit for Blood/Bone Marrow (Boehringer) and subsequently amplified using the SMART cDNA amplification system (Clontech), following the manufacturers' instructions. The resulting cDNA was labeled using RediprimeII random prime labeling system (Amersham) in the presence of radioactive ^{32}P dCTP. Hybridization on high-density arrays was performed in Techne bottles with 10 ml DIG Easy Hybsolution (Boehringer) for 3 days at 45°C . The membranes were then washed 3×15 minutes in $0.1\times$ SSC, 0.1% SDS at 65°C , mounted in a cassette (sandwiched between two layers of thin plastic/clingfilm), and exposed for 2 weeks at the photoscreen. Image data were captured using a Storm Scanner (Molecular Dynamics). For each clone the expression estimates were obtained by averaging over the intensities of the replicates. The data were then log-transformed and normalized using the quantile-normalization procedure to make the distribution of probe intensities across arrays similar (Bolstad et al., 2003).

4.2 Identification of Predictive Biological Markers

For our Bayesian analysis, we set $h = 10^7$ and $H = cI$ with $c = 5$ and chose the binomial prior to have an expectation of 5, since we expected models with very few genes to perform well. The choice of c did not appear to be so critical as long as the parameter is chosen in a guideline range calculated from (12), the condition number for this data being around 10^4 . Six MCMC chains with 200,000 iterations each were run. We allowed ample burn-in time by discarding the first 100,000 iterations. Starting γ vectors were taken with (i) 1, (ii) 10, (iii) 50, (iv) 100, (v) 500, and (vi) 999 randomly selected genes included.

For each chain, we ordered the list of the distinct visited subsets of genes according to their normalized conditional posterior probabilities $\pi(\gamma | \hat{\mathbf{Y}}, \mathbf{X}, \mathbf{Z})$ with $\hat{\mathbf{Y}}$ computed as the mean of the sampled \mathbf{Y} 's. We also looked at the marginal probabilities of inclusion of single variables. These allow us to locate sets of genes that can be of interest for further

Table 1

Genes included in the 10 best models of each chain and in the 10 best models of the union of the six chains. The last column reports the crossvalidated misclassification error rates and the labels of the misclassified samples.

Chain	Genes included in the 10 best models	Error rate (obs-label)
1	Filamin, Calcineurin A1, MT3-MMP, Adenosine A2B, receptor, Connexin 40, Jun-B (clone 1), Jun-B (clone 2), Paxillin, Notch 4	0.05 (20-late)
2	Profilin, Calcineurin A1, Complement Component 7, MT3-MMP, Adenosine A2B receptor, Jun-B (clone 1), Cbl-b, Jun-B (clone 2), Lymphocyte interferon alpha, Calmodulin-I	0.05 (20-late)
3	Profilin, Calcineurin A1, MT3-MMP, Adenosine A2B, receptor, Connexin 40, Rasf-A Pla2, Jun-B (clone 1), Jun-B (clone 2), Paxillin	0.05 (20-late)
4	Filamin, Calcineurin A1, MT3-MMP, Adenosine A2B receptor, Connexin 40, Jun-B (clone 1), Jun-B (clone 2), Paxillin, Notch 4	0.05 (20-late)
5	Filamin, MT3-MMP, Adenosine A2B receptor, Connexin 40, Rasf-A Pla2, Jun-B (clone 1), Jun-B (clone 2), Paxillin, Notch 4	0.1 (2-early, 20-late)
6	Filamin, Profilin, Calcineurin A1, MT3-MMP, Rasf-A Pla2, Jun-B, Jun-B, Paxillin, Notch 4	0.1 (2-early, 20-late)
Overall	Filamin, Profilin, Calcineurin A1, MT3-MMP, Adenosine A2B, receptor, Connexin 40, Jun-B (clone 1), Jun-B (clone 2), Paxillin	0.05 (20-late)

investigation, simply by considering the genes with marginal posterior probability greater than a certain value. Interesting sets can also be found by exploring the visited models that have the highest values of the posterior probability. In general, we found the selection based on marginal probabilities to give very similar sets to those obtained by inspecting the “best”

visited models. Table 1 reports the genes that appeared in the 10 best models of each chain. We also considered the pooled set of visited models obtained by taking the union of the sets visited by the six chains. Genes that appeared in the 10 best models of this pooled set are also reported in Table 1. Heat maps of all selected sets in Table 1 are shown in Figure 1.

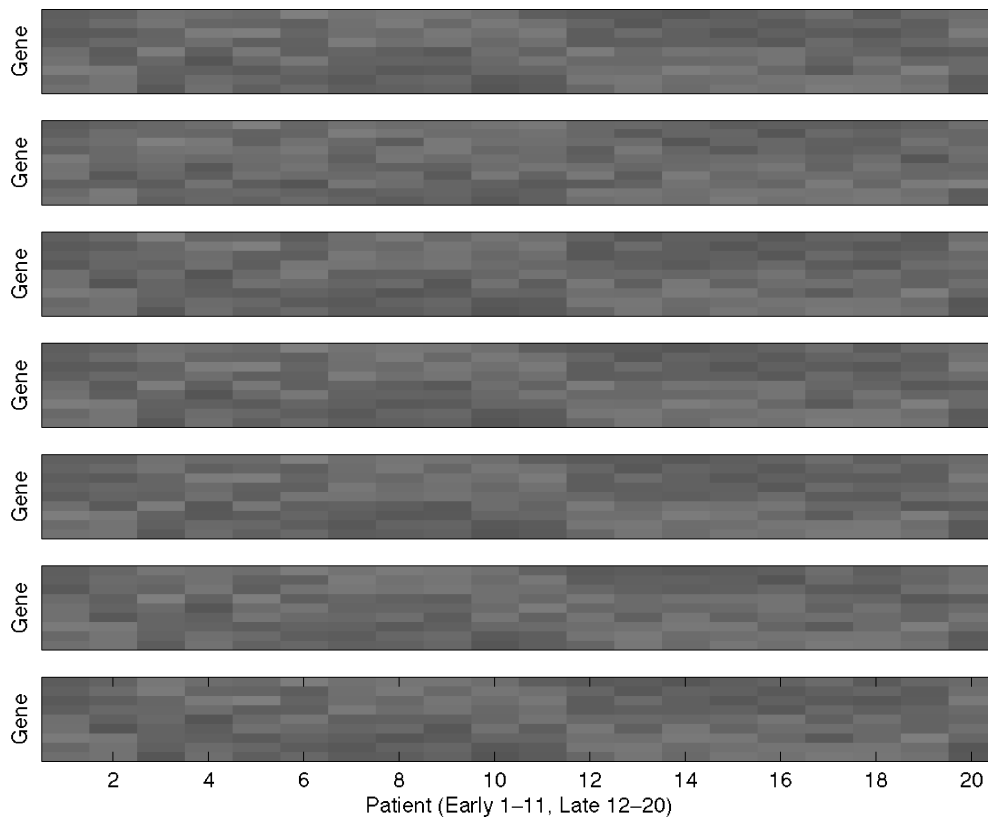


Figure 1. Heat maps of selected genes for early–late disease stage discrimination. From top to bottom: genes included in the 10 best models of single chains and in the 10 best of the union of all chains.

There is sufficient partial overlap among gene sets to allow an unambiguous interpretation (see biological findings). As for prediction, our analysis produced very low crossvalidated errors (5% for chains 1, 2, 3, and 4 and 10% for chains 5 and 6). Misclassification errors and labels of the misclassified samples are reported in Table 1.

4.3 Biological Findings

From a purely biological perspective, two main and related cellular functions are represented in our models. Figure 2 provides a graphical representation of the biological system under study and related functions of the genes selected by our method.

The first aspect of cell physiology that clearly emerges from our models is represented by a number of actin-associated proteins that are involved in cell motility and cytoskeleton re-arrangement. Four genes involved in cytoskeleton remodeling and motility are included in our models. These are: profilin, paxillin, filamin, and calmodulin. Profilin can promote polymerization of actin filaments by transporting monomers to the fast-growing barbed ends of filaments (Dos Remedios et al., 2003). Paxillin is a focal adhesion protein that serves as an adapter molecule, providing docking sites for both structural and regulatory proteins (Dos Remedios et al., 2003). In lymphocytes paxillin has been proven to be directly associated with the cytoplasmic tail of $\alpha 4$ integrin (Herreros et al., 2003), establishing a direct link between cytoskeleton remodeling and migration of leukocytes in the site of inflammation. Filamin is also regulating the cytoskeleton rearrangement by acting on integrins, transmembrane receptor complexes, and second messengers (Stossel et al., 2001). Calmodulin is only present in one of the sets but it is interesting to notice its link with motility and in particular its essen-

tial role in regulating human T-cell aggregation (Fagerholma et al., 2001).

The second aspect of cell physiology that emerges from our findings is represented by a set of genes known to influence the ability of human lymphocytes to respond to activation and in particular to express the pro-inflammatory cytokine interleukin 2 (IL-2). These are notch 4 receptor, the gap junction protein connexin 40, and the adenosine receptor. Notch is a developmental gene that has been recently implicated in the development of human lymphocytes (McKenzie et al., 2003). Interestingly, there is a demonstrated link between the notch receptor and at least two actin-binding proteins (Zhang et al., 1998). It has also been shown that immunoglobulin and cytokine expression in mixed lymphocyte cultures is reduced by disruption of gap junctions of which connexin 40 is a structural component (Oviedo-Orta, Gasque, and Evans, 2001). Similarly, the expression of adenosine A2B receptor is regulated in T-cell activation suggesting that the role of adenosine in lymphocyte deactivation is mediated by A2BRs (Mirabet et al., 1999). It is of interest to notice that the expression of notch and adenosine receptor is higher in late RA than in early RA whereas the expression of connexin is lower in early RA. This clearly shows that the overall effect of gene regulation in late-stage RA blood cells may result in T-cell unresponsiveness to stimulation, a feature of RA lymphocytes that has been previously described.

Collectively, the components of our models are tightly associated, structurally (profilin, filamin, paxillin, and notch 4) and/or via a functional association. Our results clearly define two main characteristics of late RA blood cells. The first seems to be associated to a potential increase in the ability to polymerize actin filaments and the second is a reduced ability to respond to activation by expressing the

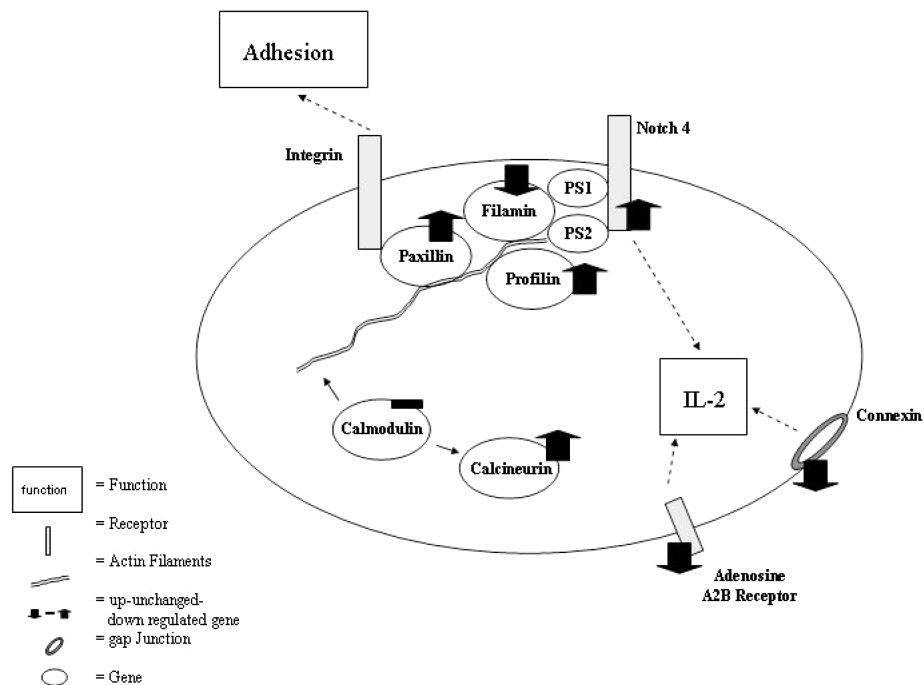


Figure 2. Functional pathways related to early-late disease progression.

pro-inflammatory cytokine IL-2. The ability to migrate is crucial to many cells in the immune system. The hypothesis we generate from this model is that blood cells in late-stage RA may have a higher capacity to rearrange their cytoplasm (for example in response to activation) and to migrate. These have profound repercussions on the disease progression. Increased mobility may result in a larger amount of cells concentrated in the primary focus of inflammation. On the other hand notch 4, connexin 40, and the adenosine receptor point in the direction of an “anergic” phenotype, a well-known feature of RA lymphocytes.

A further aspect of our findings is that the transcription factor Jun B plays a central role in all selected sets. Transgenic mice ubiquitously expressing Jun B develop a disease that resembles a natural form of the human chronic myeloid leukemia (CML; Weitzman, 2001). The in vitro analysis of transgenic cells revealed that the constitutive expression of Jun B is associated with increased proliferation upon treatment with GM-CSF and with a reduced number of apoptotic granulocytes. Although these effects may be indirect, these data suggest that the higher level of Jun B and its relevance in discriminating between early- and late-stage arthritis could be related to an increased ability of immune cells to proliferate and escape apoptosis.

5. Discussion

We have developed an approach to discrimination with multinomial probit models that uses latent variables and Bayesian mixture priors for variable selection. Information on the size of models for prediction can easily be included in our Bayesian search. Our method, at the same time, has the flexibility of allowing the identification of larger sets of genes, via the inspection of the best visited models or the marginal probabilities of single genes, as we have shown. We have applied our methodology to the problem of identifying molecular signatures of disease stage for the classification of rheumatoid arthritis patients. While it is certainly true that some gene substitution may be possible with our MCMC search, we have identified a subset of very predictive models which make sense biologically. It should be stressed that our methodology aims to get models that predict well, not models that happen by chance to discriminate perfectly internally to the actual data. Our crossvalidatory procedure demonstrates that this is so. The predictive sets we have identified are based on molecular signatures of blood cells and suggest hypotheses on the physiology of these cells. Further experimental work is needed to validate our selections. Moreover, our models are based on blood expression profiling. The contribution of individual cell types remains to be determined. This is currently the focus of our laboratory work. A possible extension of our methodology may be in trying to tease out interaction effects among the genes by elaborating the prior on γ . This will require historical data or expert opinion to specify the hyperparameters.

ACKNOWLEDGEMENTS

N. Sha is partially supported by BBRC/RCMI grant 2G12RR08124 from the NIH. M. Vannucci is supported by the National Science Foundation, CAREER award DMS-0093208. M. G. Tadesse is supported by NCI grant CA90301. We thank

Prof Hill Gaston for useful discussions and for providing the clinical samples.

RÉSUMÉ

Nous nous intéressons ici à des problèmes de discrimination lorsque le nombre de prédicteurs dépasse largement la taille de l'échantillon, et nous proposons une approche bayésienne de sélection de variables à des modèles probit multinomiaux. Notre méthode utilise des mélanges d'a-priori et des méthodes MCMC pour sélectionner les ensembles de variables qui diffèrent selon les classes. Nous appliquons notre méthodologie à un problème de génomique fonctionnelle utilisant des données profilant l'expression des gènes. Le but de l'analyse est d'identifier les signatures moléculaires qui caractérisent deux états différents de la polyarthrite rhumatoïde.

REFERENCES

- Agresti, A. (1990). *Categorical Data Analysis*. New York: John Wiley & Sons.
- Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association* **88**, 669–679.
- Barbieri, M. M. and Berger, J. O. (2004). Optimal predictive model selection. *Annals of Statistics* **32**, 870–897.
- Bolstad, B. M., Irizarry, R. A., Astrand, M., and Speed, T. P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on bias and variance. *Bioinformatics* **19**, 185–193.
- Brown, P. J. (1993). *Measurement, Regression, and Calibration*. Oxford: Clarendon Press.
- Brown, P. J., Vannucci, M., and Fearn, T. (1998a). Multivariate Bayesian variable selection and prediction. *Journal of the Royal Statistical Society, Series B* **60**, 627–641.
- Brown, P. J., Vannucci, M., and Fearn, T. (1998b). Bayesian wavelength selection in multicomponent analysis. *Journal of Chemometrics* **12**, 173–182.
- Brown, P. J., Vannucci, M., and Fearn, T. (2002). Bayes model averaging with selection of regressors. *Journal of the Royal Statistical Society, Series B* **64**, 519–536.
- Dawid, A. P. (1981). Some matrix-variate distribution theory: Notational considerations and a Bayesian application. *Biometrika* **68**, 265–274.
- Dos Remedios, C. G., Chabra, D., Kekic, M., Dedova, I. V., Tsubakihara, M., Berry, D. A., and Nosworthy, N. J. (2003). Actin binding proteins: Regulation of cytoskeletal microfilaments. *Physiological Reviews* **83**, 433–473.
- Fagerholma, S. C., Prescott, A., Cohena, P., and Gahmberg, C. G. (2001). An essential role for calmodulin in regulating human T cell aggregation. *FEBS Letters* **491**, 131.
- Gelfand, A. E. (1996). Model determination using sampling-based methods. In *Markov Chain Monte Carlo in Practice*, W. R. Gilks, S. Richardson, and D. J. Spiegelhalter (eds). London: Chapman & Hall.
- George, E. I. and McCulloch, R. E. (1997). Approaches for Bayesian variable selection. *Statistica Sinica* **7**, 339–373.
- Geweke, J. (1991). Efficient simulation from the multivariate normal and student-t distributions subject to linear constraints. In *Computing Science and Statistics: Proceedings of the Twenty-Third Symposium on the Interface*, 571–578. Alexandria, Virginia: American Statistical Association.

- 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., and Lander, E. S. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science* **286**, 531–537.
- Herreros, L., Rodriguez-Fernandez, J. L., Brown, M. C., Alonso-Lebrero, J. L., Cabanas, J. C., Sanchez-Madradi, F., Longo, N., Turner, C. E., and Sanchez-Mateos, P. (2000). Paxillin localizes to the lymphocyte microtubule organizing center and associates with the microtubule cytoskeleton. *The Journal of Biological Chemistry* **34**, 26436–26440.
- Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12**, 55–67.
- Ibrahim, J. G., Chen, M. H., and Gray, R. (2002). Bayesian models for gene expression with DNA microarray data. *Journal of the American Statistical Association* **97**, 88–99.
- Lindley, D. V. (1957). A statistical paradox. *Biometrika* **44**, 187–192.
- Madigan, D. and 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.
- McKenzie, G. J., Young, L. L., Briend, E., Lamb, J. R., Dallman, M. J., and Champion, B. R. (2003). Notch signalling in the regulation of peripheral T-cell function. *Seminars in Cell Development Biology* **14**, 127–134.
- Mirabet, M., Herrera, C., Cordero, O. J., Mallol, J., Lluís, C., and Franco, R. (1999). Expression of A2B adenosine receptors in human lymphocytes: Their role in T cell activation. *Journal of Cell Science* **112**, 491–502.
- Oviedo-Orta, E., Gasque, P., and Evans, W. H. (2001). Immunoglobulin and cytokine expression in mixed lymphocyte cultures is reduced by disruption of gap junction intercellular communication. *FASEB Journal* **51**, 768–774.
- Sha, N., Vannucci, M., Brown, P. J., Trower, M. K., Amphlett, G., and Falciani, F. (2003). Gene selection in arthritis classification with large-scale microarray expression profiles. *Comparative and Functional Genomics* **4**, 171–181.
- Stossel, T. P., Condeelis, J., Cooley, L., Hartwig, J. H., Noegel, A., Schleicher, M., and Shapiro, S. S. (2001). Filamins as integrators of cell mechanics and signalling. *Nature Reviews on Molecular and Cell Biology* **2**, 138–145.
- Tanner, T. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association* **82**, 528–549.
- Weitzman, J. B. (2001). Life and death in the JUNgle. *Trends in Molecular Medicine* **7**, 141–142.
- West, M., Blanchette, C., Dressman, H., Huang, E., Ishida, S., Spang, R., Zuzan, H., Marks, J. R., and Nevins, J. R. (2001). Predicting the clinical status of human breast cancer using gene expression profiles. *Proceedings of the National Academy of Sciences* **98**, 11462–11467.
- Zhang, W., Woo Han, S., McKeel, D. W., Goate, A., and Wu, J. Y. (1998). Interaction of presenilins with the filamin family of actin-binding proteins. *The Journal of Neuroscience* **18**, 914–922.

Received May 2002. Revised October 2003.

Accepted January 2004.