Outcome-Driven Clustering of Microarray Data

Dissertation Advisor: Professor Dianne Finkelstein Jessie Jann Hsu

Outcome-Driven Clustering of Microarray Data

Abstract

The rapid technological development of high-throughput genomics has given rise to complex high-dimensional microarray datasets. One strategy for reducing the dimensionality of microarray experiments is to carry out a cluster analysis to find groups of genes with similar expression patterns. Though cluster analysis has been studied extensively, the clinical context in which the analysis is performed is usually considered separately if at all. However, allowing clinical outcomes to inform the clustering of microarray data has the potential to identify gene clusters that are more useful for describing the clinical course of disease.

The aim of this dissertation is to utilize outcome information to drive the clustering of gene expression data. In Chapter 1, we propose a joint clustering model that assumes a relationship between gene clusters and a continuous patient outcome. Gene expression is modeled using cluster specific random effects such that genes in the same cluster are correlated. A linear combination of these random effects is then used to describe the continuous clinical outcome. We implement a Markov chain Monte Carlo algorithm to iteratively sample the unknown parameters and determine the cluster pattern. Chapter 2 extends this model to binary and failure time outcomes. Our strategy is to augment the data with a latent continuous representation of the outcome and specify that the risk of the event depends on the latent variable. Once the latent variable is sampled, we relate it to gene expression via cluster specific random effects and apply the methods developed in Chapter 1. The setting of clustering longitudinal microarrays using binary and survival outcomes is considered in Chapter 3. We propose a model that incorporates a random intercept and slope to describe the gene expression time trajectory. As before, a continuous latent variable that is linearly related to the random effects is introduced into the model and a Markov chain Monte Carlo algorithm is used for sampling. These methods are applied to microarray data from trauma patients in the Inflammation and Host Response to Injury research project. The resulting partitions are visualized using heat maps that depict the frequency with which genes cluster together.

Contents

List of Figures

List of Tables

Acknowledgments

I owe a great debt of gratitude to my advisors, Dianne Finkelstein and David Schoenfeld, for their guidance through the research process. I am immensely thankful for their invaluable expertise and unwavering faith in me. Working with them has truly been a wonderful learning experience. I would also like to thank Rebecca Betensky for providing me with many insightful suggestions and ideas.

I will forever be grateful to my friends who have supported me every step of the way. I especially want to thank Leesa Lin for showing me the meaning of true friendship and Shannon Stock for bringing so much sunshine into my life. I will also always cherish the joyful memories and friendships that HSPH Dance Club has given me over the years.

I would like to thank all my friends in the department, including Kate Jie Hu, Sabrina Khan, Rui Wang, Alisa Stephens, Anna Snavely, Lei Quanhong, Sharon Lutz, Miguel Marino, Mariel Finucane, Betsy Ogburn, Bonnie Zhang, Jeanne Jiang, Huan Huang, Zeynep Coban, Rui Zhao, Keith Betts, Roland Matsouaka, Alfa Yansane, and Alane Izu.

I would also like to thank my friends from home, including Arpi Shaverdian, Shirley Chou, Leigh Momii, Cheryl Hou, and Michelle Chan.

Most of all, I am grateful to my mom, dad, and sister for their unconditional love and support. Thank you Sherrie for being my best friend and always making me laugh. No matter where I am or what I am doing, my family has always given me a place to call home.

A Bayesian Approach to Informatively Clustering Microarray Data

Jessie J. Hsu, Dianne M. Finkelstein, and David A. Schoenfeld

Department of Biostatistics, Harvard School of Public Health
Biostatistics Center, Massachusetts General Hospital

1.1 Introduction

In the past decade, new technologies for high-throughput genomics and proteomics have developed with the potential of revolutionizing medicine. Gene expression microarrays are one such technology that measure the levels of RNA expression in a cell. These expression levels are constantly changing, producing a rich influx of information. Due to the wealth of potential knowledge encoded in the human genome that is captured in microarray experiments, there is substantial interest in identifying differential gene expression patterns and relating gene activity to phenotypic information.

Our goal is to reduce a microarray dataset into clusters of genes that are biologically meaningful and to use those clusters to predict patient outcome. We would like to find clusters of genes that are both correlated with each other as well as associated with patient outcome, and we hypothesize that using outcome information to drive the pattern discovery can potentially result in gene clusters that are more coherent and biologically meaningful. We are motivated by the Inflammation and Host Response to Injury research program, also known as the Glue Grant (http://www.gluegrant.org). The Glue Grant is a large-scale interdisciplinary study of inflammation following severe trauma or burn injury. The immune system reacts to injury by activating the inflammation response in an attempt to prevent further damage to the body, and presumably the chain of events that takes place as the body tries to stabilize and recover is reflected in differential gene expression. The general aims of the Glue Grant are to uncover the biological reasons why patients have such varying responses following their injury, to understand the genomic and proteomic markers that predict clinical outcomes, and to determine the relationship between changes in gene expression and clinical features. For this paper, we focus on the association between patterns in differential gene expression and metabolic recovery in patients with severe trauma.

Many methods have been developed for relating gene expression to clinical outcomes, most of which involve reducing the dimensionality of the gene expression data. One way to go about this is to identify a subset of genes that are predictive markers of outcome. The simplest method for subset selection is univariate variable selection, where each gene is individually tested for significance and the top ranked ones are included in a multivariate model. Stepwise selection procedures achieve the same end but can be unstable for high-dimensional datasets. Increased stability and reduced prediction error can be obtained by penalized regression methods which operate by imposing a constraint on the parameters, leading to coefficient shrinkage (Tibshirani, 1996). In particular, lasso simultaneously obtains parameter estimates and achieves variable selection because the absolute value constraint causes some coefficients to be estimated at exactly zero. Dimension reduction can also be accomplished by principal components regression (Hastie *et al.*, 2001a). This is an unsupervised procedure that reduces the gene expression values down to their principal components and incorporates the first few components that explain the majority of the predictor variation into a regression model. A supervised version of this approach is partial least squares regression (PLSR) (Park *et al.*, 2002). Here, both the predictors and outcome are decomposed such that the latent vectors used in the decomposition maximize their covariance. Given that our goal of using outcome information to drive the data reduction is partially addressed by PLSR, we will use it for comparison to our method.

Clustering is another widely used form of microarray dimension reduction that is based on the assumption that groups of genes are more similar to each other than others for reasons such as related functionality, shared biological pathways, or a similar effect on outcome. One approach, though computationally burdensome, is to perform a stochastic search across the entire space of possible partitions and select the true clustering to be the one with the highest likelihood. Another approach is to cluster the genes across patient samples via a technique such as K-means and then to use the cluster expression averages in a regression model (Eisen *et al.*, 1998). K-means clustering is a classic clustering algorithm that finds the partition of K sets that minimizes the distance of each observation to its center, where each cluster center is the mean of the observations in that cluster (Hartigan, 1975). Achieving the optimal clustering using K-means with a Euclidean distance metric is equivalent to maximizing the likelihood that corresponds to modeling gene expression as a normally distributed cluster specific fixed effect. The maximum likelihood occurs when each gene is assigned to its nearest cluster center such that the within cluster sum of squares is minimized. This approach operates under the assumption that all the genes to be clustered are independent. This is appropriate for clustering independent individuals but is flawed for clustering features that have a correlation structure. Rather, it is more reasonable to state that genes in the same cluster are correlated while genes across different clusters are independent. Furthermore, K-means assumes that there is only one correct clustering pattern and does not provide a measure of uncertainty associated with the cluster assignments.

A related formulation of the clustering problem is the normal mixture model, where each observation is viewed as arising from a mixture of distributions. Fraley and Raftery (1998) and Ghosh and Chinnaiyan (2002) discussed model-based clustering where the gene expression data is modeled as a normal mixture and clusters are determined by the Expectation-Maximization (EM) algorithm. A Bayesian approach can also be used to fit the mixture model (Vogl *et al.*, 2005). In these approaches, the probability distribution of each gene is modeled as the sum of K weighted underlying distributions, each representing the distribution of a gene conditional on membership in each cluster. The entire data likelihood is then a product across all the genes. Once again, this approach fails to specify any sort of correlation between genes in the same cluster. These types of mixture models are valid for clustering patients, but do not reasonably extend to the setting of clustering features measured on each patient.

The statistically sound approach for model-based clustering is to include a random effect such that highly correlated genes fall in the same cluster. Ng *et al.* (2006) implemented an EM algorithm to fit the random effects model for clustering. Alternatively, the Bayesian paradigm provides a unified framework for fitting complex hierarchical models. For example, Booth *et al.* (2008) proposed a random effects clustering model and performed a stochastic search for clusters using the posterior distribution of the unknown partition as the objective function. Tadesse *et al.* (2005) presented a Markov chain Monte Carlo (MCMC) sampling scheme for simultaneously selecting discriminating genes and clustering patients. The advantage of the Bayesian approach is that it accounts for uncertainty in all of the parameters, including variation about cluster membership. It can incorporate prior information and naturally allows outcome to drive the clustering of the genes when fitting the joint model.

The notion of outcome-informed clustering has been studied less extensively. Hastie *et al.* (2001b) touched upon the idea of informative clustering in his proposal of a supervised approach called 'tree-harvesting' where clusters of genes are explored in a stepwise fashion and related to outcome using the intermediate results of hierarchical clustering. Dettling and Bühlmann (2002) discussed a strategy that directly incorporates the response variable into the clustering process by using a rank-based test statistic for finding groups of genes that discriminate a categorical response. Ideally, one would like to simultaneously find clusters and model the outcome such that each part is influenced by the other.

In this chapter, we propose a joint model for simultaneously clustering correlated gene expression data and predicting a continuous patient outcome. We use a random effects model for describing gene expression cluster membership and relate the latent cluster effects to a continuous patient outcome via a linear model. We develop a MCMC clustering algorithm for model fitting and parameter inference based on a marginalized likelihood. By simultaneously modeling patient outcome with gene expression and developing a clustering algorithm that makes use of clinical data, we will generate clusters that are more useful for describing the clinical course of injury.

Our methodology is described in Section 1.2. The results of simulation studies are presented in Section 1.3, and an analysis of the Glue Grant data is presented in Section 1.4.

We conclude with a discussion in Section 1.5.

1.2 Methods

1.2.1 Model Specification

We propose a joint model for clustering correlated gene expression data that is driven by a continuous patient outcome. Consider representing the microarray dataset as a $N \times J$ matrix consisting of gene expression values for J genes measured on N patients. Let Y_{ij} be the gene expression value for patient i and gene j belonging in cluster k. Conditional on membership of gene j in the k^{th} cluster, the random effects model for describing gene expression is

$$Y_{ij} = c_{ik(j)} + \epsilon_{ij} \tag{1.1}$$

where $i = 1, ..., N, j = 1, \ldots, J$, and $k = 1, \ldots, K$. Here, $c_{ik(j)}$ are patient-cluster specific random effects that represent the cluster centers and induce correlation between genes in the same cluster. We assume $c_{ik(j)} \sim N(0, \tau^2)$ after the data have been log-transformed and centered to have mean zero. We also assume that c_{ik} and $c_{ik'}$ are independent for $k \neq k'$. Thus, for a given patient, the covariance between genes in the same cluster is τ^2 while genes in different clusters and across different patients remain independent. The ϵ_{ij} are measurement errors, assumed to be distributed $N(0, \sigma^2)$. To link the gene clusters to patient outcome, we specify a linear relationship between the clusters and Z_i, a continuous outcome for patient i,

$$Z_i = \sum_{k=1}^{K} \beta_k c_{ik(j)} + \xi_i. \tag{1.2}$$

The cluster effects relate gene expression and patient outcome to each other by acting as covariates in the regression model. The β_k are the respective regression coefficients for each cluster, and the error terms are assumed to be $\xi_i \sim N(0, \gamma^2)$.

Latent variables $\phi = (\phi_{11}, ..., \phi_{1K}, \phi_{21}, ..., \phi_{JK})$ are introduced into the model, where

ϕ_{jk} is an indicator denoting membership of gene j in cluster k. Additionally, let $\omega = (\omega_1, \ldots, \omega_K)$ be the cluster weights with $\omega_k > 0$ for all k and $\sum\limits_k \omega_k = 1$. These weights represent the probability of belonging in each cluster.

1.2.2 Joint Likelihood

We will work with the marginal likelihood where the random effects c are integrated out:

$$f(Y, Z|\sigma, \tau, \beta, \gamma, \phi, \omega) = \int f(Y, Z|c, \sigma, \tau, \beta, \gamma, \phi, \omega) f(c|\tau) dc. \tag{1.3}$$

This is for ease of computation, since the random effects are nuisance parameters and our model fitting procedure is facilitated by not having to estimate all of them. A closed form expression for (1.3) is readily achieved, as described next, because the random effects are normally distributed.

Let $X_i = (Y_i, Z_i)$, the vector of observations associated with patient i, where $Y_i = (Y_{i1}, ..., Y_{iJ})$. Let Θ denote the set of parameters $\{\sigma, \tau, \beta, \gamma, \phi, \omega\}$. The resulting complete data likelihood for (Y, Z) is given by a multivariate normal distribution,

$$f(Y, Z|\Theta) = \prod_{i=1}^{N} \frac{exp\{-\frac{1}{2} X_i' \Sigma^{-1} X_i\}}{(2\pi)^{(J+1)/2} |\Sigma|^{1/2}}.$$

The covariance matrix Σ is a symmetric $(J+1) \times (J+1)$ matrix that is block diagonal in all but the last row and column. It is represented by

$$\begin{aligned}
\Sigma_{u,v} &= \sigma^2 I(u = v) + \tau^2 \sum_{k=1}^{K} I(u, v \in S_k) \\
\Sigma_{u,J+1} &= \tau^2 \sum_{k=1}^{K} I(u \in S_k) \beta_k \\
\Sigma_{J+1,J+1} &= \tau^2 \sum_{k=1}^{K} \beta_k^2 + \gamma^2
\end{aligned} \tag{1.4}$$

where the subscripts index the matrix elements. Here, $u = (1, \ldots, J), v = (1, \ldots, J)$, and S_k denotes the k^{th} cluster set.

A closed form expression exists for both the inverse and determinant of Σ. Therefore, the expression for the multivariate normal distribution simplifies substantially, speeding up computation of the Metropolis-Hastings algorithm.

1.2.3 Prior Distributions for Model Parameters

We specify a non-informative prior distribution for every parameter. The prior for σ is set to be uniform on a wide range. We also specify a uniform prior on a wide range for the hierarchical parameter τ, as recommended by Gelman (2006). The standard non-informative prior is used for the regression parameters $(\beta, \gamma^2) \propto 1/\gamma^2$.

Non-informative conjugate priors are specified for ω and ϕ. A symmetric Dirichlet prior is set for the weights, $P(\omega_1, \ldots, \omega_K) \propto$ Dirichlet$(\alpha, \ldots, \alpha)$. Larger values of α reflect the presence of more clusters, while smaller values of α reflect fewer clusters. Lastly, the cluster membership variable ϕ has a multinomial prior that depends on the weights, $P(\phi_{jk} = 1) = \omega_k$.

1.2.4 MCMC Clustering Algorithm

We fit the model by implementing a MCMC algorithm that consecutively samples every parameter until a sufficient representation of the posterior distribution is achieved. The MCMC sampling procedure consists of repeating the following six steps until convergence:

1. Sample σ^2.
2. Sample τ^2.
3. Sample ϕ.
4. Sample ω.

5. Sample β.
6. Sample γ^2.

Gibbs sampling is used for sampling the parameters that have an available full conditional posterior distribution. The set of samples obtained through multiple iterations estimates the posterior distribution of that parameter. When the full conditional distribution cannot be directly sampled from, we use the Metropolis-Hastings algorithm. Candidate values are drawn from a proposal distribution and accepted with probability proportional to the ratio of the posterior density evaluated at the current value to the posterior density evaluated at the new value. That is, samples are accepted with probability

$$min(1, \frac{P(\theta^*|Y,Z)/Q(\theta^*|\theta')}{P(\theta'|Y,Z)/Q(\theta'|\theta^*)})$$

where Q is the proposal density, P is the posterior likelihood, θ' is the current parameter value, and θ^* is the candidate parameter value.

Update of variance parameters

The Metropolis-Hastings algorithm is used to sample σ^2, τ^2, and γ^2. We use an inverse gamma proposal distribution with shape parameter s and scale parameter s/θ. These tuning parameters are determined experimentally during initial runs to accept proposed samples at the recommended rate of $40\% - 45\%$ (Gelman *et al.*, 2004).

Update of cluster membership and weights

Cluster membership ϕ is sampled from a multinomial distribution with probabilities proportional to the likelihood given the current parameter values. For every gene, we calculate the likelihood of belonging in each of the K clusters. The value of the likelihood weighted by the current value of ω then becomes the updated multinomial sampling probabilities. We sample directly from the full conditional distribution, given by

$$f(\phi_j|Y, Z, \sigma, \tau, \beta, \gamma, \omega) \propto \prod_{k=1}^{K}(f(Y, Z|\sigma, \tau, \beta, \gamma, \omega) * \omega_k)^{\phi_{jk}}.$$

After sampling the cluster memberships of all the genes, ω is sampled via a Gibbs step. The full conditional distribution of ω is Dirichlet$(\alpha + n_1, ..., \alpha + n_K)$, where n_k is the current number of genes in the k^{th} cluster.

Update of regression coefficients

The regression coefficients β are sampled as a block by a Metropolis-Hastings algorithm. We specify a multivariate normal proposal distribution with an independent mean that is continuously updated whenever ϕ_j is updated in step 3. When a gene is assigned to a different cluster, the average of the coefficients of the original cluster and the new cluster become the new coefficients for the respective clusters. The resulting value after updating all the ϕ_j is then used as the proposal mean. As for the proposal covariance, we found through initial experimental runs that a covariance of 0.2*I, where I is the identity matrix, produces an acceptance rate near the recommended rate of $20\% - 25\%$.

1.2.5 Model Without Outcome

The random effects clustering model in (1.1) can stand alone as a special case of the full joint model. In this simple case, $f(Y|\sigma, \tau, \phi, \omega)$ is a multivariate normal density with covariance matrix equal to (1.4) with the last row and column removed. Simplifying the likelihood expression leads to

$$f(Y|\sigma, \tau, \phi, \omega) = \frac{exp\{-\frac{1}{2}\sum_{i=1}^{N}[\sigma^{-2}\sum_{k=1}^{K}(\sum_{j\in S_k}(Y_{ij}^2) - \frac{\tau^2}{\sigma^2+n_k\tau^2}(\sum_{j\in S_k}Y_{ij})^2)]\}}{[(2\pi)^J(\sigma^2)^{J-K}\prod_{k=1}^{K}(\sigma^2+n_k\tau^2)]^{N/2}}.$$

Estimates of the unknown parameters are obtained by Metropolis-Hastings sampling in the same manner as previously described. We found that the random effects model does in fact cluster genes differently from the fixed effects model. To illustrate analytically, we make the simplifying assumption that there is an equal number of genes in every cluster and derive an expression for the log-likelihood that has the variance components profiled out. As expected, the profile log-likelihood for the fixed effects model is maximized when each gene is as close as possible to its cluster center,

$$L \propto \sum_{i} \sum_{k} \sum_{j \in S_k} (Y_{ij} - \overline{Y}_{ik})^2.$$

Interestingly, for the random effects model,

$$L \propto \sum_{i} \sum_{k} \sum_{j \in S_k} (Y_{ij} - \overline{Y}_{ik})^2 * n \sum_{i} \sum_{k} (\overline{Y}_{ik})^2.$$

This result implies that genes are clustered in a way that not only minimizes the distance to the cluster centers, but also shrinks the cluster means towards zero.

1.2.6 Determining the Number of Clusters

Though the true number of non-empty clusters is unknown, it does not need to be sampled separately in our algorithm because an estimate of K is obtained at every iteration as an immediate result from the samples of cluster membership. Recall that $P(\phi_{jk} = 1)$ is proportional to the weighted likelihood of belonging in cluster k. These multinomial probabilities are always non-zero because ω_k is positive for all k regardless of cluster size. Due to the probabilistic nature of the allocation, there is always a chance that a cluster will end up with no genes, or that an empty cluster will become filled at any given iteration. Therefore, the only value that needs to be specified in advance is K_{max}, the maximum number of clusters. K_{max} can also be thought of as the total number of both empty and non-empty clusters, where $0 \leq K \leq K_{max}$.

1.2.7 Posterior Inference

Posterior Distributions

The MCMC algorithm outputs samples from the posterior distribution of each of the parameters and can be characterized by its posterior mean and posterior credible interval. We make an exception for ϕ and instead summarize its posterior distribution by the concordance between every pair of genes, where concordance is measured as the percentage of iterations that a pair of genes falls into the same cluster. Displaying cluster membership as a heat map allows the relationship between genes to be captured and circumvents the issue of label switching which can cause the appearance of non-identifiability. The other cluster-dependent parameters, β and ω, can also appear non-identifiable as a consequence of label switching. A remedy is to index these parameters by genes rather than by clusters.

Prediction

Given microarray data for a new patient, the predictive density of that patient's outcome can be obtained. Since Z_i and Y_i are both normally distributed, $f(Z_i|Y_i)$ is also normally distributed. Its expected value and variance are given by

$$E(Z_i|Y_i) = \tau^2 \sum_{k=1}^{K} \frac{\beta_k}{\sigma^2 + n_k\tau^2} (\sum_{j \in S_k} Y_{ij})$$

$$Var(Z_i|Y_i) = \tau^2\sigma^2 \sum_{k=1}^{K} \frac{\beta_k^2}{\sigma^2 + n_k\tau^2} + \gamma^2.$$

This distribution allows us to estimate the expected value of a future patient's outcome based on their expression values.

1.3 Simulations

Simulations were conducted to evaluate the performance of our algorithm and to study the effect of outcome inclusion and different parameter values on the resulting clusters. In the first simulation study, we generated 100 datasets under the model without outcome and 100 datasets under the model with outcome. Each set of data consisted of 80 patients and 50 genes arising from 5 clusters. We considered various values of τ^2 to assess the ability of our method to detect the correct cluster structure when cluster variation is low and when cluster variation is high compared to the variation in the residual error. This ratio, τ^2/σ^2, is what we will refer to as the variance ratio. The remaining parameter values were set to $\sigma = 1, \gamma = 1, \beta_1 = -3, \beta_2 = -1, \beta_3 = 2, \beta_4 = 3,$ and $\beta_5 = 5$. We set $\alpha = 1$ for the Dirichlet prior and $K_{max} = 10$. For every dataset, we ran 10,000 iterations and discarded 5,000 as burn-in.

A visual representation of the simulation results is presented in Figure 1.1. Cluster membership is depicted as a heat map that shows the proportion of iterations that every pair of genes is assigned to the same cluster. In the event of label switching, summarizing the output as a heat map aids in visualizing the groups, but even in the absence of label switching, the heat map has the advantage of providing information about the uncertainty surrounding the allocations. We do not assume that ϕ is a fixed value, but rather a parameter with a distribution where some groupings are more likely than others. Heat maps for the models with and without outcome are shown for $\tau^2/\sigma^2 = 4$ and $\tau^2/\sigma^2 = 0.2$. The genes are listed along both axes in the same order, grouped together by their true cluster membership. Concordance is represented as a gradient from white (0%) to red (100%) with 16 discrete shades of color.

As depicted in the heat maps in Figure 1.1, the clustering is very clear when the variance ratio is large regardless if outcome information is used. On the other hand, we observe a weak signal when the variance ratio is small. However, the clusters become more well-defined when outcome is introduced into the model.

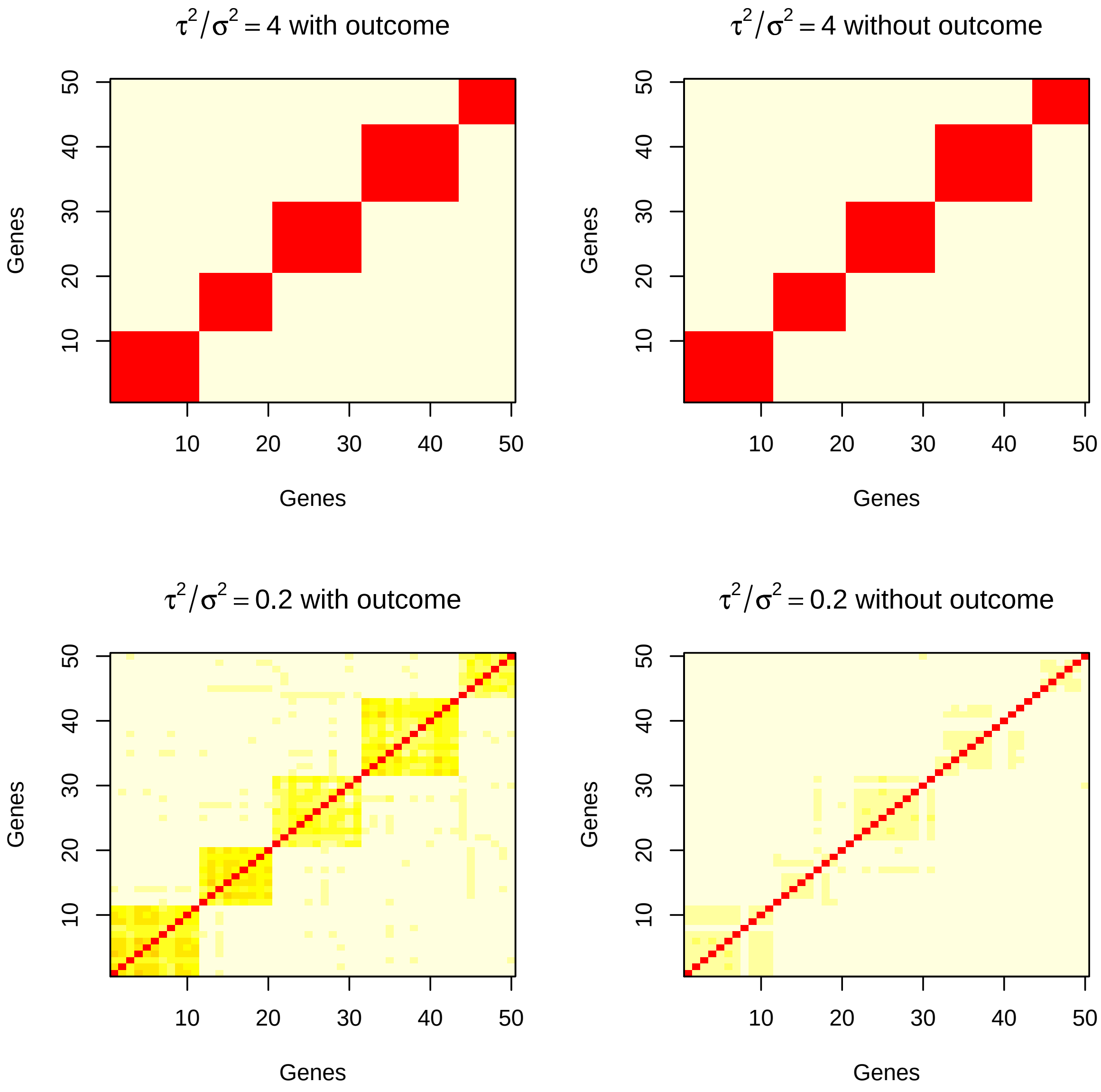

Figure 1.1: Cluster heat maps for simulated data. Concordance varies from 0% (white) to 100% (red).

Table 1.1: Parameter estimates resulting from simulation for the model with and without outcome with N=80 patients, J=50 genes, and 100 replications with 5000 iterations each.

	True value	Mean	SE	Mean of SE
With outcome				
σ	1	0.993	0.012	0.012
τ	2	1.768	0.079	0.065
β_1	-3	-3.824	2.039	1.499
β_2	-1	-1.674	2.623	1.893
β_3	2	1.425	3.107	2.286
β_4	3	2.483	3.189	2.285
β_5	5	4.520	3.389	2.479
Without outcome				
σ	1	0.993	0.012	0.012
τ	2	1.780	0.079	0.065

Table 1.1 displays posterior summary statistics of the parameters that result from the simulation when the variance ratio is large. We see that they are all well estimated by the MCMC algorithm. Note that estimation of the β_k only makes sense when the algorithm has converged to a stable pattern and is conditional on $K = 5$ in the current case. When there is substantial uncertainty in the clustering output, it is not possible to report averaged values for β_k because there are different sets of β_k associated with different values of K. However, when the majority of the genes cluster in the same way across iterations, we can restrict ourselves to those iterations that estimated 5 clusters and expect that the averages are reasonable estimates of the true value.

A second simulation study was conducted to understand the effect of the variance ratio on cluster uncertainty. Uncertainty is defined as the frequency of pairwise clustering inaccuracies as compared to the true cluster pattern. For this we considered a small dataset with 15 patients and 6 genes arising from 3 clusters with every two genes belonging to the same cluster. We varied τ^2 to range from 0.5 to 7, while the other parameters were fixed at $\sigma = 1, \gamma = 1, \beta_1 = -2, \beta_2 = 1$, and $\beta_3 = 3$. For every case, 10,000 iterations were run with half discarded as burn-in. The model was fit both with and without outcome, and uncertainty was calculated for the range of variance ratios. The results are plotted in Figure 1.2. Given the described clustering pattern, the maximum amount of uncertainty that can be attained is 0.33. As τ^2/σ^2 increases, the uncertainty decreases towards zero and we see that clustering with outcome consistently produces less uncertainty than clustering without outcome.

When the clustering signal is strong, that is when τ^2 is large, the clustering parameter tends to converge quickly to the correct answer. Though this is advantageous, the drawback is that the clustering parameter does not mix as well as one would hope. This is because our algorithm only considers moves of one gene at a time, so the likelihood tends to not change enough to accept reallocations of a single gene. Nevertheless, there is no apparent need to over-explore the partition space when there is a strong signal because we still obtain convergence to the right clustering pattern. On the other hand, when the clustering signal is weak or nonexistent, mixing is irrelevant because the algorithm cannot reach convergence anyways due to a weak signal. It is in the case of a moderate signal that we would most want to see good mixing and hope that the frequency of genes clustering together is reflective of the probability of belonging in the same cluster. In the simulation setting, the lack of mixing in any given dataset is circumvented by averaging across all the simulated datasets. This is effectively the same as implementing several chains for every run and averaging across the chains, which is what we proceed to do in the Glue Grant data analysis.

1.4 Application

We applied our methodology to the Inflammation and Host Response to Injury trauma data, a rich dataset that contains information on numerous factors related to the biology of inflammation following severe traumatic injury. There are a total of 167 patients in the trauma dataset, each of whom has their blood leukocyte expression levels mea-

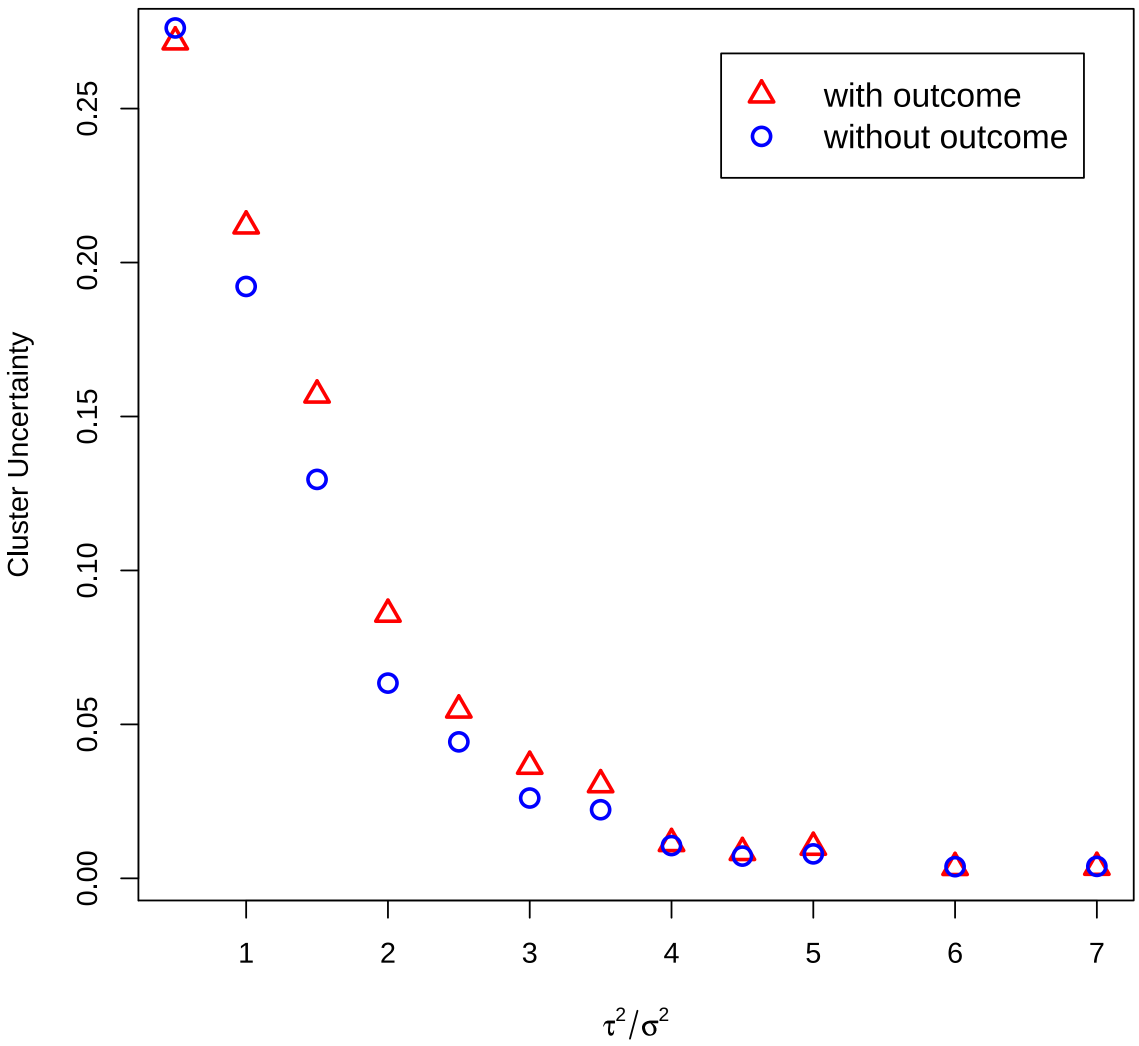

Figure 1.2: Plot of cluster uncertainty with and without outcome for different variance ratios.

sured on an Affymetrix microarray chip consisting of 54,674 probe sets (which we will henceforth call ‘genes’). The full dataset consists of microarrays that have been taken at seven different time points following the patients’ injury, starting from immediately after the injury to up to 28 days later. For our analysis however, we restrict ourselves only to microarray data collected on day four from the 147 patients who are still in the intensive-care unit at that time.

The gene expression values have been pre-processed using dChip, log-transformed and centered prior to analysis. We use a subset of 87 genes for our cluster analysis. These genes were pre-selected by Glue Grant investigators to be those that had significant differential expression with at least a two-fold difference between patients with a clinical outcome of complicated versus uncomplicated recovery. Our objective is to find clusters of genes that are associated with each other as well as associated with a relevant patient outcome. The outcome that we use in our analysis is maximum multiple organ failure (MOF), a continuous score that describes the severity of the patient’s multiple organ failure and is predictive of metabolic recovery. MOF is the cumulative sum of individual scores from the respiratory, renal, hepatic, cardiovascular, and hematologic components, each ranging in value from 0 to 4 for least to most severe. The resulting groups of genes can then be examined for their functional relationships and interdependent roles in the inflammation response pathway.

We ran ten MCMC chains, each starting from a different set of randomly chosen overdispersed starting values. Non-informative priors were specified for all the parameters; hyper-parameters were chosen to be $\alpha = 1$ and $K_{max} = 15$. We evaluated mixing and convergence by assessing the trace plots and observed that convergence occurred fairly quickly. As mentioned previously, we ran multiple chains to simulate good mixing and averaged across all chains. Thus, for each of the MCMC chains, we ran 10,000 iterations with 5,000 discarded as burn-in.

The heat map for clustering combined across chains for the model with and without

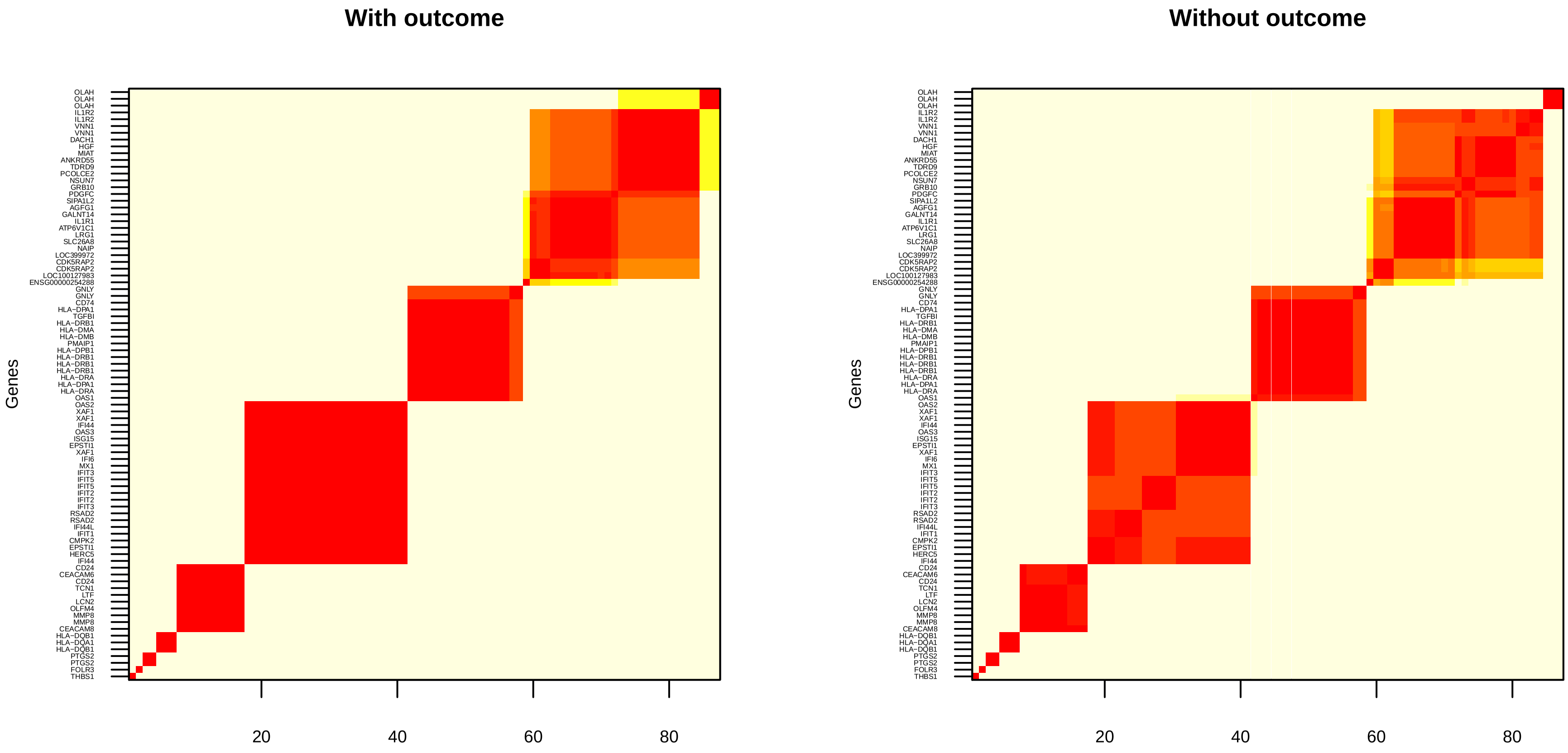

Figure 1.3: Cluster heat maps for Glue Grant trauma data with and without a continuous outcome (maximum MOF score). Concordance varies from 0% (white) to 100% (red).

patient outcome is shown in Figure 1.3. The genes are listed along both axes in the same order for both plots. When the model is fit with outcome, the genes labeled 1-58 fall into seven distinct clusters the majority of the time upon convergence. These seven clusters are clearly distinguished by the red boxes along the diagonal starting from the bottom-left with the exception of some orange overlap between genes labeled 42 to 58 on the plot. It appears that about 20% of the time, genes 57 and 58 form their own cluster of size two, while the remainder of the time they are part of the larger cluster. On the other hand, several breakdown combinations are observed for genes 59-87. They do not group into clear partitions, implying that several partitions have similar posterior probabilities.

When we fit the model without outcome, we obtain a heat map that appears comparable though there is more pronounced uncertainty. The general partition structure remains the same, but now there is more orange and yellow in some groups because the posterior pairwise probabilities are not as high. There are various subsets of genes that form their own clusters on occasion. The only genes for which there is actually less uncertainty are genes 85-87, as they now exclusively cluster together.

In both cases, clusters consisting of only one gene are allowed. Normally we may not wish to have singletons for a dataset that has thousands of genes, but since this is a fairly small subset of genes that was pre-selected to be important, we do not want to be too strict in forcing singletons into larger sized clusters.

A summary of the output is shown in Table 1.2. The mean of the variance ratio is 3.22 for the case with outcome and 3.51 for the case without outcome. The uncertainty surrounding cluster membership is minimal because the estimated variance ratio is relatively large. The coefficient estimates are conditional on $K = 10$, where the clusters are of size 1, 1, 2, 3, 10, 24, 17, 1, 25, and 3 (from left to right in Figure 1.3). Only those iterations for which the genes in each respective cluster exclusively group together are used in calculating the coefficient estimate for that cluster.

Table 1.2: Results of Glue Grant trauma data analysis with and without a continuous outcome (maximum MOF score).

	With outcome		*Without outcome*	
	Mean	95% credible interval	Mean	95% credible interval
σ	0.579	(0.563, 0.592)	0.564	(0.503, 0.587)
τ	1.039	(0.992, 1.085)	1.056	(1.009, 1.142)
β_1	0.239	(-0.625, 0.961)	–	–
β_2	0.179	(-0.499, 0.684)	–	–
β_3	-0.514	(-1.606, 0.030)	–	–
β_4	-0.356	(-1.523, 0.387)	–	–
β_5	0.436	(-3.363, 2.321)	–	–
β_6	-0.344	(-8.745, 3.443)	–	–
β_7	-0.653	(-6.512, 3.036)	–	–
β_8	0.629	(-0.363, 1.329)	–	–
β_9	-0.173	(-8.713, 5.110)	–	–
β_{10}	0.582	(-0.997, 1.273)	–	–

Figure 1.4 displays a heat map of the gene expression data that has been sorted according to the clustering results. Every row is a patient, where the patients are sorted by increasing MOF, and every column is a gene, where the genes are sorted by the mean of the regression coefficient of their respective clusters. Gene expression values have been centered at zero in both directions; red represents under-expression and green represents over-expression. The cluster groupings are denoted by the brackets along the bottom of the figure. As expected, different values of MOF are associated with different gene expression patterns, and genes in the same cluster have similar expression patterns, Furthermore, clusters with a positive coefficient have an opposite pattern from clusters with a negative coefficient. Cluster nine is the one exception. Even though the mean of β_9 is negative yet the pattern implies the opposite, its value is very close to zero and suffers from high variability. Substantial cluster uncertainty surrounding cluster nine accounts for its high coefficient variability and expected instability.

Lastly, the average prediction error was calculated by 5-fold cross-validation. The mean-squared error (MSE) for our method is 4.78, while the MSE from partial least

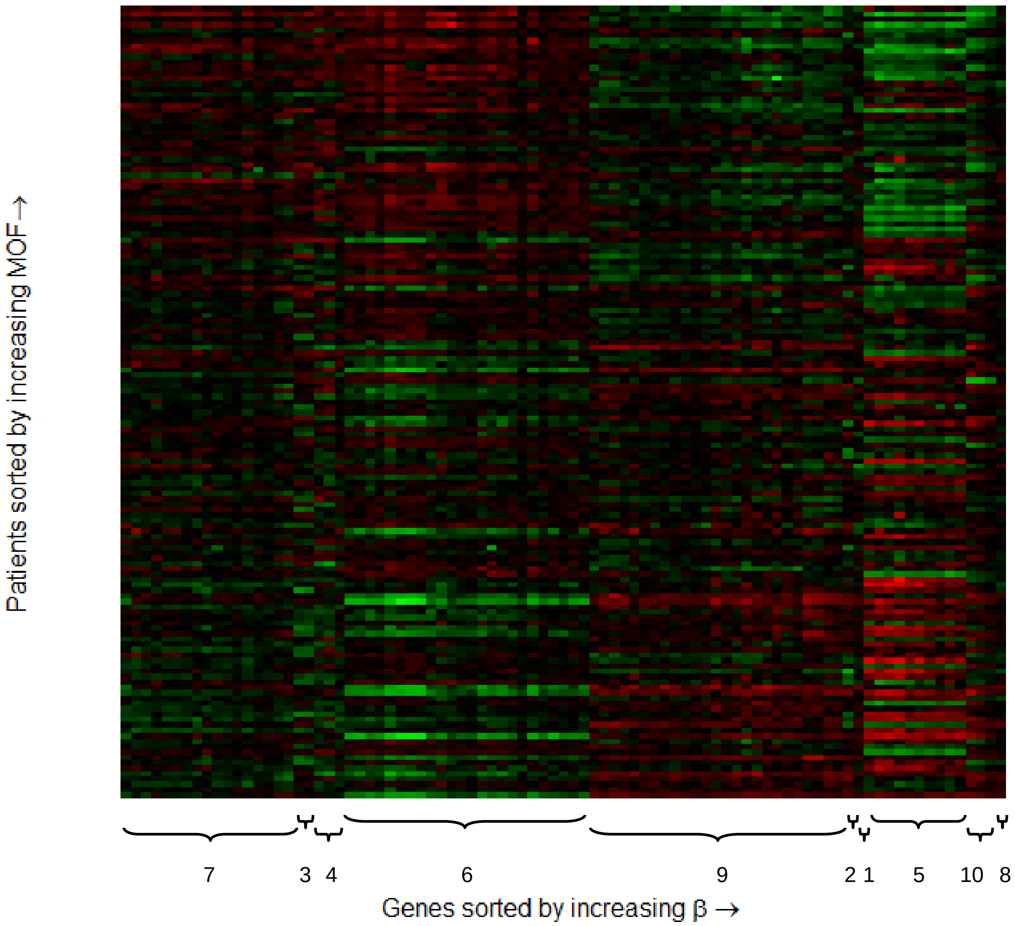

Figure 1.4: Heat map of sorted gene expression data. Red represents under-expression and green represents over-expression. The numbers correspond to the cluster labels in Table 1.2.

squares regression is 9.61 conditional on K=10. In addition to having a lower MSE, our method fulfills the additional purpose of providing interpretable gene clusters.

1.5 Discussion

We have proposed Bayesian methodology for the informative clustering of genes. Our model accounts for correlation between genes in the same cluster and jointly relates the gene expression values to a continuous patient outcome such that this additional information helps drive the clustering of the genes.

It would be worthwhile to consider relaxing some of the assumptions of our model. For example, a heterogeneous covariance structure where a different τ_k is specified for every cluster would allow for more flexibility. A non-linear relationship between the clusters and outcome could also be modeled. We mentioned some solutions to deal with the mixing problem, but the best way to handle this issue would be to incorporate global moves such as splitting or combining clusters. Though this would allow the partition space to be explored more fully, it would add extra computational complexity.

Our model can be extended to accommodate categorical outcomes using a probit or logistic model, or time to event outcomes using semi-parametric models. Additionally, the model can be extended to the longitudinal microarray setting where it is assumed that groups of genes cluster together in their patterns over time.

Latent Variable Methods for Clustering Genes Using Binary and Failure Time Outcome Data

Jessie J. Hsu, Dianne M. Finkelstein, and David A. Schoenfeld

Department of Biostatistics, Harvard School of Public Health
Biostatistics Center, Massachusetts General Hospital

2.1 Introduction

The relationship between microarray data and binary patient outcomes is generally framed as a classification problem (Ring and Ross, 2002; Quackenbush, 2006). Gene expression data can be highly predictive of clinical outcomes such as disease type, as demonstrated in Golub *et al.* (1999). Microarrays can also be used to distinguish between different diseases. For example, Dudoit *et al.* (2002) compares nearest-neighbor classifiers, linear discriminant analysis, and classification trees for discriminating malignant versus normal tissue based on gene expression data in cancer patients. On the other hand, modeling the relationship between gene expression and binary outcomes using logistic regression can also lead to models that are highly descriptive of outcome.

The association between microarray data and survival outcomes can be studied in a similar manner. Typically, the objective of these studies is to determine the hazard of experiencing the event that is associated with the observed expression measurements. Jung *et al.* (2005) and Gui and Li (2005) developed methods for identifying a subset of genes that are biologically important and predictive markers of survival. In an approach suggested by Bair and Tibshirani (2004), each gene is given a Cox score based on the proportional hazards partial likelihood and the top ranked ones are included in a multivariate Cox model. A comprehensive overview of predicting patient survival from microarray data is presented in Bøvelstad *et al.* (2007) and Wieringen *et al.* (2009).

Due to the high-dimensionality of microarray experiments, dimension reduction is often a necessary step in data analysis. Common methods for dimension reduction include principal component analysis and partial least squares, both of which reduce the expression values to fewer dimensions based on correlation. If outcome information is available, a supervised approach is generally preferred. Though these methods have been extended to both binary and survival settings (Nguyen and Rocke, 2002; Park *et al.*, 2002), the drawback is they do not give interpretable components.

We use clustering as a form of dimension reduction and as a way to gain insight into underlying gene expression patterns. A standard cluster analysis involves a two-step process where clustering is performed by a method such as K-means and the cluster averages are used as covariates to model outcome. However, ideally we would like to find clusters and simultaneously predict outcome such that each part is influenced by the other. To this end, the Bayesian approach for model fitting is a natural way to allow for outcome-driven clustering of gene expression data due to its iterative approach. Previous Bayesian contributions in the realm of clustering multivariate data include Booth *et al.* (2008) and Tadesse *et al.* (2005), both of whom use Markov chain Monte Carlo (MCMC) methods for clustering data. For our contribution, we seek to use patient outcome to inform the discovery of gene clusters with the hope that the resulting clusters provide a more coherent depiction of the underlying biological mechanism.

We propose a joint model that relates clusters of gene expression measurements to binary and event time outcomes. Our model for gene expression adds complexity to standard mixture models (McLachlan *et al.*, 2002) by incorporating cluster and subject specific random effects. These random effects account for correlation between genes in the same cluster and allow us to extend the mixture-model construction to the setting where non-independent patient features are being clustered. For a binary outcome, the probability of experiencing the event is related to the clusters via the introduction of latent continuous variables into the model. The latent variables are then linearly related to the cluster random effects which transforms the model into the standard linear regression formulation. Conditional on the continuous latent response, the methodology for estimating the posterior distribution of the parameters is equivalent to clustering using a continuous outcome as described in Chapter 1. Obtaining these latent parameters is readily achieved by adding an extra step into the MCMC sampling scheme (Albert and Chib, 1993). An example of this type of Bayesian latent variable approach is described in Sha *et al.* (2004). In their paper, they augment a probit model with continuous latent variables to accommodate multinomial response variables for the purpose of high-dimensional variable selection. We extend our binary model to accommodate survival data by treating time-to-recovery as a series of binary observations at a fixed number of discrete time points where the outcome at every time point is evaluated as a binary response. Again, we augment the data by assuming that the hazard of the event at any given time depends on a latent continuous variable. Then, a negative binomial model with a constant hazard of recovery is assumed for describing the amount of time that a patient is at risk.

Our method is applied to trauma data from the Inflammation and Host Response to Injury Program. Also known as the Glue Grant, this research project is an interdisciplinary study of the biological changes that a patient goes through after experiencing severe trauma injury. The data consists of expression values measured on thousands of genes, as well as various clinical measurements and recovery endpoints for every patient. Utilizing patient recovery to drive the process of clustering genes can potentially result in groups that more thoroughly capture the relationship between genes.

We proceed with a detailed description of the methodology in Section 2.2. The results of simulations are shown in Section 2.3. An analysis of microarray data from the Glue Grant is presented in Section 2.4, and we end with a discussion in Section 2.5.

2.2 Methods

2.2.1 Clustering Genes Using a Binary Outcome

For every subject i, $i = 1, \dots, N$, we observe (Y_i, X_i), where Y_i is a vector of gene expression values and X_i is a single binary outcome. Our goal is to group the genes into several clusters based on similarities in their expression values and their association to the binary response. The genes should cluster in such a way that genes in the same cluster are highly correlated with each other, while genes in different clusters are mutually independent. Genes in the same cluster should also share a similar relationship to the response variable. In order to obtain clusters with these properties, we will fit a joint model that relates the gene expression values to the binary outcome.

The first part of the joint model describes the observed gene expression data. We assume that the dependence among genes in the same cluster is induced by subject and cluster specific random effects. Thus, for gene j belonging in cluster k, the model for gene expression is formulated as follows:

$$Y_{ij} = c_{ik(j)} + \epsilon_{ij}, c_{ik(j)} \sim N(0, \tau^2), \epsilon_{ij} \sim N(0, \sigma^2) \tag{2.1}$$

It can be shown that the presence of patient-cluster specific random effects in the model, represented by $c_{ik(j)}$, results in a covariance of τ^2 between genes in the same cluster and a covariance of zero between genes in different clusters. We assume that c_{ik} and $c_{ik'}$ are independent for $k \neq k'$. We also note that the reason that both the random effects and the error terms have mean zero is because we assume the data have been centered at zero for every patient and gene prior to analysis.

Though the random effects provide information about the relationship between genes in different clusters, they provide no indication of the clustering pattern itself. It is therefore necessary to introduce additional parameters into the model to represent the unknown cluster membership. We use indicator variables ϕ_{jk} to denote the membership of gene j in cluster k. We assume the vector of indicators associated with each gene has a multinomial distribution with probabilities ω_k, $k = 1, \ldots, K$, where $\omega_k > 0$ $\forall k$ and $\sum\limits_k \omega_k = 1$. The entire clustering pattern can then be obtained directly from the matrix of cluster indicators.

The second part of the joint model describes the observed binary response. For every patient i, X_i is a Bernoulli(p) distributed random variable that equals one if the patient experienced the event. However, fitting a model that has a Bernoulli distributed random variable greatly increases the difficulty of implementing a MCMC because none of the posterior distributions are tractable. Therefore we facilitate the Bayesian model fitting procedure by introducing a normally distributed latent variable Z_i that will be simulated by MCMC and assume that the probability associated with X_i depends on Z_i. Known as the data augmentation approach (Tanner and Wong, 1987), Z_i can be thought of as an unmeasured underlying process that directly determines the value of the observed binary response X_i. Augmenting the data to include Z_i requires the specification of a function that links the relationship between X_i and Z_i (Albert and Chib, 1993). In the context of data augmentation, the probit link is most commonly used:

$$P(X_i = 1) = \Phi(Z_i) \tag{2.2}$$

where Φ is the cumulative density function of the standard normal distribution. The dependence of X_i on Z_i is straightforward; a smaller value of Z_i implies that X_i is more likely to be zero and a larger value of Z_i implies that X_i is more likely to be one.

Up to this point, we have proposed separate models for the clusters of gene expression data and for the binary outcome. The final layer of the model is to connect these two components together. The gene clusters are related to the binary outcome by a linear relationship between the cluster random effects and Z_i, the continuous latent representation of the binary response:

$$Z_i = \mu + \sum_{k=1}^{K} \beta_k c_{ik(j)}. \tag{2.3}$$

This is essentially a linear regression model where the β_k act as coefficients that describe the effect of the cluster centers on the continuous latent outcome.

We noticed that when β was unconstrained, it tended to increase without bound. The reason this model may not converge is because we only observe X_i and have fewer degrees of freedom than provided by the normal model. We found that convergence occurs when β is constrained to lie on the unit sphere such that $\beta^T \beta = 1$. A convenient distribution for points on a sphere is the von Mises-Fisher distribution, which we detail in Section 2.2.2.

Prior Distributions

Non-informative prior distributions are specified for every parameter. The priors for the hierarchical standard deviation components σ and τ are uniform densities on a wide range. This is approximately equivalent to specifying an Inverse-χ^2 prior distribution on σ^2 and τ^2. For the vector of cluster probabilities ω, we specify a conjugate symmetric Dirichlet$(\alpha, \ldots, \alpha)$ prior. Smaller values of α reflect a prior belief that there should be fewer clusters and larger values drive the clustering towards more clusters. The cluster membership variable has a conjugate multinomial prior that depends on the weights, $P(\phi_{jk} = 1) = \omega_k$. The intercept term μ is given a non-informative uniform prior, and a von Mises-Fisher prior distribution with concentration parameter $\lambda = 0$ is specified for the vector β. This parameter setting is non-informative and is equivalent to uniformity on a K-dimensional unit sphere.

2.2.2 MCMC Clustering Algorithm

The MCMC algorithm iterates between draws from the full conditional posterior distributions $f(Z_i|Y_i, X_i, \Theta)$ for every patient i and $f(\Theta|Y, X, Z)$, where Θ denotes the entire set of parameters $\{\sigma, \tau, \mu, \beta, \phi, \omega\}$. We exclude the random effects c from the parameter set because we will only be working with distributions that have the random effects integrated out so that they no longer appear in the likelihood. Carrying out this mathematical detail greatly reduces the dimension of the parameter space and increases the stability of the algorithm.

To sample Z_i, write

$$\begin{aligned}
&f(Z_i|Y_i, X_i, \Theta) \\
&\propto f(X_i|Z_i, Y_i, \Theta) f(Z_i|Y_i, \Theta) \\
&= [\Phi(Z_i) I(X_i = 1) + \Phi(-Z_i) I(X_i = 0)] \phi(\mu_{Z_i|Y_i,\Theta}, \sigma^2_{Z_i|Y_i,\Theta}). \qquad (2.4)
\end{aligned}$$

Since both Z_i and Y_i are normally distributed, $f(Z_i|Y_i, \Theta)$ is readily available as a conditional normal distribution with mean $\mu_{Z_i|Y_i,\Theta} = \mu + \tau^2 \sum_{k=1}^{K}(\beta_k/(\sigma^2 + n_k\tau^2))(\sum_{j\in S_k} Y_{ij})$ and $\sigma^2_{Z_i|Y_i,\Theta} = \tau^2\sigma^2 \sum_{k=1}^{K} \beta_k^2/(\sigma^2 + n_k\tau^2)$.

Due to the difficulty of drawing Z_i directly from (2.4), we utilize the acceptance-rejection algorithm for sampling Z_i. The acceptance-rejection algorithm for simulating random variables with density $f(\cdot)$ operates by finding a density $g(\cdot)$ from which it is easy simulate, along with a constant M such that $f(\theta)/g(\theta) \leq M \; \forall\theta$. The algorithm proceeds by simulating values θ^* from $f(\cdot)$ and accepting these values with probability $f(\theta^*)/(Mg(\theta^*))$. As a result, the elements in the set of values that are accepted will be random variables from $f(\cdot)$.

In our case, $f(\cdot)$ is the expression in (2.4). If we let $g(\cdot) = \phi(\mu_{Z_i|Y_i,\Theta}, \sigma^2_{Z_i|Y_i,\Theta})$, then $M = 1$ is an upper bound. Since $Z_1, \ldots, Z_N$ are independent random variables, the steps in the algorithm are as follows:

1. Generate $Z_i^* \sim \phi(\mu_{Z_i|Y_i,\Theta}, \sigma^2_{Z_i|Y_i,\Theta})$.
2. Generate $U \sim Uniform(0, 1)$.
3. If $U < [\Phi(Z_i)I(X_i = 1) + \Phi(-Z_i)I(X_i = 0)]$, then accept Z_i^*; otherwise, reject Z_i^*.

Next, we need to sample from $f(\Theta|Y, X, Z)$. Conditional on Z and Y, it is not necessary to also condition on X because X gives no additional information given Z. To simulate any single parameter θ from the set Θ, we write the full conditional posterior distribution for θ:

$$f(\theta|Y, Z, \Theta_{-\theta}) \propto f(Y, Z|\Theta)f(\theta). \tag{2.5}$$

Note that $f(Y, Z|\Theta)$ is a product across independent patients i, where $f(Y_i, Z_i|\Theta)$ is multivariate normal with mean $(0, \ldots, 0, \mu)'$ and covariance Σ, a symmetric $(J+1)(J+$

$1)$ matrix that is block diagonal in all but the last row and column:

$$\Sigma_{u,v} = \sigma^2 I(u = v) + \tau^2 \sum_{k=1}^{K} I(u, v \in S_k)$$
$$\Sigma_{u,J+1} = \tau^2 \sum_{k=1}^{K} I(u \in S_k)\beta_k$$
$$\Sigma_{J+1,J+1} = \tau^2 \sum_{k=1}^{K} \beta_k^2$$

where $u = (1, \ldots, J)$ and $v = (1, \ldots, J)$ index the matrix elements, and S_k denotes the k^{th} cluster set. The expression for the multivariate normal distribution simplifies substantially because a closed form expression exists for both the inverse and the determinant of Σ.

If the distribution represented by (2.5) is available in closed form for any given parameter, basic Gibbs sampling is used and samples are drawn directly from the closed form distribution. If the full conditional posterior cannot be sampled from directly, we utilize the Metropolis-Hastings algorithm, where candidate values are drawn from a proposal distribution and accepted with probability proportional to the ratio of the posterior density evaluated at the current value to the posterior density evaluated at the new value. More explicitly, supposing that θ' is the current parameter value and θ^* is the candidate value, samples are accepted with probability

$$min(1, \frac{P(\theta^*|Y, Z)/Q(\theta^*|\theta')}{P(\theta'|Y, Z)/Q(\theta'|\theta^*)})$$

where Q is the proposal density and P is the posterior likelihood.

Using the theory presented above, we continue describing the details of sampling each parameter in Θ. To simulate the variance parameters σ^2 and τ^2, the Metropolis-Hastings algorithm is used. We draw candidate values from an inverse gamma proposal distribution with shape parameter s and scale parameter s/θ. These tuning parameters are determined experimentally during initial runs to accept proposed samples at the recommended rate of $40\% - 45\%$ (Gelman, 2006).

The probabilities associated with belonging in each cluster are sampled via a Gibbs step. The full conditional distribution of ω is Dirichlet$(\alpha{+}n_1, ..., \alpha{+}n_K)$, where n_k is the number of genes in the k^{th} cluster at the current iteration. We found that setting $\alpha = 1$ provides a reasonable result. The cluster membership of each gene, ϕ_j, is sampled from a multinomial distribution with probabilities proportional to the weighted likelihood given the current parameter values. The clustering space is explored in a stochastic search where each gene is moved into every cluster and the likelihood of belonging in each of the K clusters is calculated. The value of the likelihood weighted by the current value of ω then becomes the updated multinomial sampling probabilities.

The intercept term μ is sampled using Metropolis-Hastings. Candidate values are drawn from a normal proposal distribution with variance equal to one. The vector of coefficients, β, is obtained by Metropolis-Hastings sampling as well. As mentioned earlier, we constrain β to exist on the K-sphere such that the sum of squares equals one. The von Mises-Fisher (vMF) distribution for a unit vector of dimension K has probability density function $f(x) = C_K(\lambda)exp(\lambda\mu^T x)$ and is suitable for drawing candidate values with the desired constraint. Here, C is a constant, $\lambda \geq 0$ is the concentration parameter, and μ is the mean direction. Following the steps described in Wood (1994) on how to sample from the vMF distribution, the result is a K-dimensional unit vector with modal direction $(0, \ldots, 0, 1)^T$ and concentration parameter λ. Applying QR decomposition rotates the vector such that the modal direction is located at the proposed value of β.

2.2.3 Extension to Failure Time Outcome

Our proposed hierarchical model can be extended to time-to-event outcomes in a straightforward manner. We represent time-to-event as an indicator variable that is a function of time, $X_i(t)$. If patient i experiences an event at time t, then $X_i(t) = 1$; otherwise if the patient has not yet had the event by time t, then $X_i(t) = 0$. Rather than observing a single binary endpoint, we now observe a vector of binary responses for every patient with one response at every time point. The responses are recorded at fixed discrete time points until the patient is no longer in the risk set.

As in the case with a binary outcome, we utilize the data augmentation approach and introduce latent variables Z_i into the model, where Z_i is normally distributed and modeled as shown in (2.3). Let L_i be the number of times that patient i is evaluated for having the event. The hazard of the event at any particular time $t_l, l = 1, \ldots, L_i$, depends on Z_i as follows:

$$P(X_i(t_l) = 1 | X_i(t_{l-1}) = 0) = \Phi(Z_i). \tag{2.6}$$

For the purposes of illustrating our method, we assume that each patient has a constant underlying hazard of experiencing the event. However, this assumption can be relaxed to accommodate non-constant hazards with the inclusion of an additional parameter per time point. To model the amount of time that a patient is in the risk set, we assume a negative binomial distribution where the probability of success is simply the hazard of recovery as shown in (2.6). The probability of recovering at the l^{th} time point then becomes $\Phi(-Z_i)^{l-1}\Phi(Z_i)$.

The MCMC for fitting the survival model follows almost exactly the same steps as in the case of the binary outcome. The only difference occurs in step three of the acceptance-rejection algorithm because the full conditional posterior distribution of Z_i is now a product across the time points:

$$\begin{aligned} & f(Z_i | Y_i, X_i, \Theta) \\ & \propto f(X_i | Z_i, Y_i, \Theta) f(Z_i | Y_i, \Theta) \\ & = \prod_{l=1}^{L_i} [\Phi(Z_i) I(X_i(t_l) = 1) + \Phi(-Z_i) I(X_i(t_l) = 0)] \phi(\mu_{Z_i|Y_i,\Theta}, \sigma^2_{Z_i|Y_i,\Theta}). \end{aligned} \tag{2.7}$$

Therefore we have the same function $g(\cdot)$ and the same constant $M = 1$, but acceptance of Z_i^* is now based on a comparison of the uniform random variable to $\prod_{l=1}^{L_i} [\Phi(Z_i) I(X_i(t_l) = 1) + \Phi(-Z_i) I(X_i(t_l) = 0)]$.

2.3 Simulations

We conducted simulations to compare the effect of using an informative outcome against a non-informative outcome. For both, expression data was generated under the proposed model with fixed parameter values. In particular, we point out that we set $\tau = 0.5$ and $\sigma = 1$, which represents a fairly small amount of variability between the clusters compared to the residual variance. Non-related event outcomes were obtained by generating outcomes at random for every patient. We simulated 50 datasets for both informative and non-informative binary and failure time outcomes. Each set of data consisted of 80 patients and 50 genes arising from 5 clusters. For every dataset, we ran 5000 iterations and discarded 2000 as burn-in.

The cluster heat maps for the simulated data are presented in Figures 2.1 and 2.2. In both figures, having a non-informative outcome produces more uncertainty, where uncertainty is defined as the frequency of pairwise clustering inaccuracies as compared to the true cluster pattern. Given the described clustering pattern, random noise produces an uncertainty of 0.206. The uncertainties for an informative and non-informative binary outcome are 0.042 and 0.068, respectively. The uncertainties for an informative and non-informative survival outcome are 0.043 and 0.047, respectively.

2.4 Application

The Glue Grant dataset contains information on numerous factors related to the biology of inflammation following severe traumatic injury. Data on 147 subjects are included in the analysis, each of whom has their blood leukocyte expression levels measured on an Affymetrix microarray chip consisting of 54,674 probe sets (which we henceforth call ‘genes’). Arrays collected on day 4 following trauma will be used for the analysis, the reason being that allowing a few days to pass after the event will give expression levels that are more differentiated and thus more predictive of outcome.

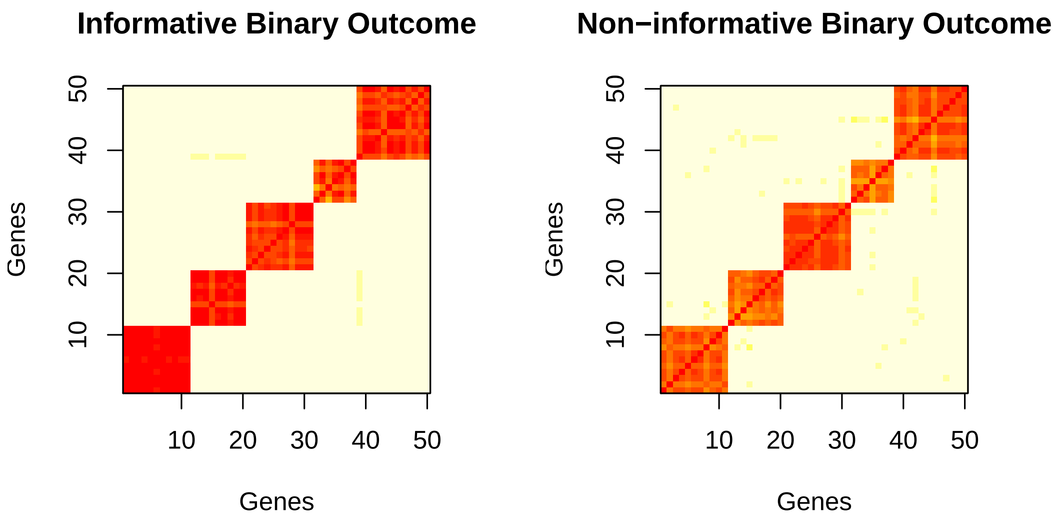

Figure 2.1: Cluster heat maps for simulated data with a non-informative and informative binary outcome.

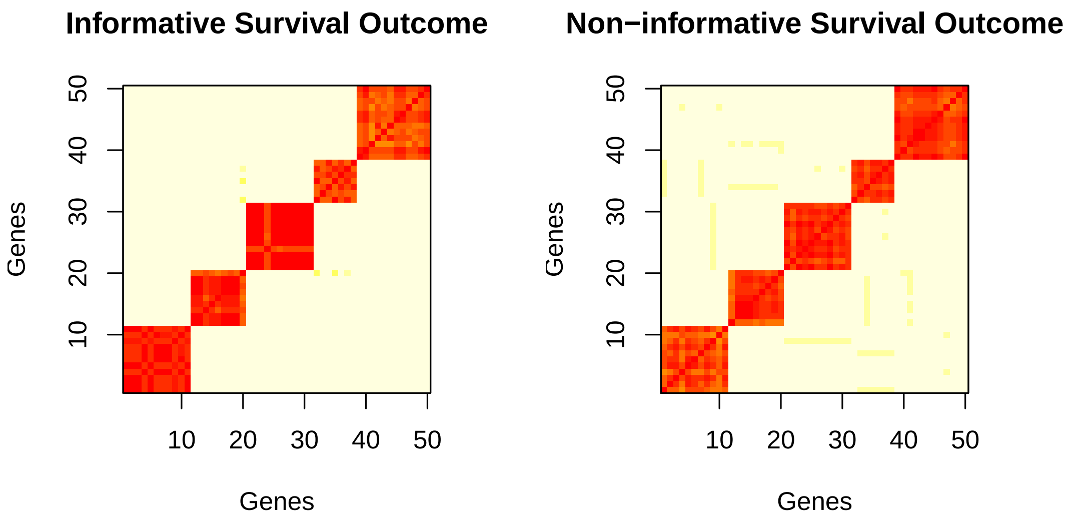

Figure 2.2: Cluster heat maps for simulated data with a non-informative and informative survival outcome.

Table 2.1: Results of Glue Grant trauma data analysis with a binary outcome (complicated vs. uncomplicated recovery) and survival outcome (time to recovery).

	Binary outcome		*Survival outcome*	
	Mean	95% credible interval	Mean	95% credible interval
σ	0.372	(0.356, 0.388)	0.381	(0.368, 0.395)
τ	0.687	(0.660, 0.716)	0.690	(0.661, 0.723)
μ	0.929	(0.569, 1.306)	-1.382	(-1.583, -1.184)

The gene expression values have been pre-processed using dChip, log-transformed and centered prior to analysis. We use a subset of 87 genes for our cluster analysis that have been pre-selected by Glue investigators to be those that had significant differential expression with at least a two-fold difference between patients with complicated versus uncomplicated recovery. Complicated recovery implies the patient had a time to recovery of more than 14 days, and patients with an uncomplicated status recovered in less than 14 days.

For both the binary and survival analyses, we ran eight chains with over-dispersed starting values. We ran 8000 iterations until convergence and discarded 2000 iterations as burn-in. The maximum number of clusters was set to be 15. For our method, we only need to specify the maximum number of clusters and not the exact number because our algorithm allows for empty clusters when the genes are tested for membership against every cluster. However, since we only make single gene transitions when sampling cluster membership, there is a tendency to under-explore the partition space. Therefore, several chains at different starting values were implemented and subsequently averaged for purposes of inference. Estimates of the parameters are shown in Table 2.1.

We define the event of interest to be complicated versus uncomplicated recovery class for the binary outcome. For the survival outcome, the response measurement is time to recovery from trauma. Patients are followed for 28 days, and time to recovery is calculated as the maximum time to cardiovascular, hematologic, hepatic, renal, or respiratory recovery. We assume that recovery can only occur once for every patient and that once recovery has occurred, the patient is no longer at risk. Recovery is the only absorbing state in the model; once a patient recovers, the patient is considered to have reached the end of the study. If patients do not recover during the course of the study or if they die prior to the last observed day, they are censored on day 28 and have an observed indicator vector that consists of all zeroes. Since only five of the 147 subjects died from their injuries, mortality was not considered an appropriately sensitive variable for informing distinct clusters. In addition to the five patients who died within the first 28 days, seven patients did not recover within the first 28 days. Both of these groups of patients are censored at 28 days since it is evident that none of these patients will recover by day 28.

A visual representation of clustering with and without outcome is presented in Figure 2.3. Cluster membership is depicted as a heat map that shows the proportion of iterations that every pair of genes was assigned to the same cluster. The similarity, or concordance, between two genes is defined as the percentage of iterations that they are assigned to the same cluster. Concordance is depicted by a color gradient and ranges from 0% (white) to 100% (red). By representing the cluster results in a heat map, label switching is accounted for and the allocation frequencies can be visualized clearly. In all three heat maps, the genes are aligned in the same order along both axes. The resulting partitions from using a binary outcome and from using a survival outcome appear similar to each other. Slightly different groups are found when clustering without outcome.

2.5 Discussion

In this chapter, we have developed methodology for using binary and failure time outcomes to inform the clustering of gene expression data. The intention is primarily for exploratory purposes, though the method can also be used for prediction. The clusters

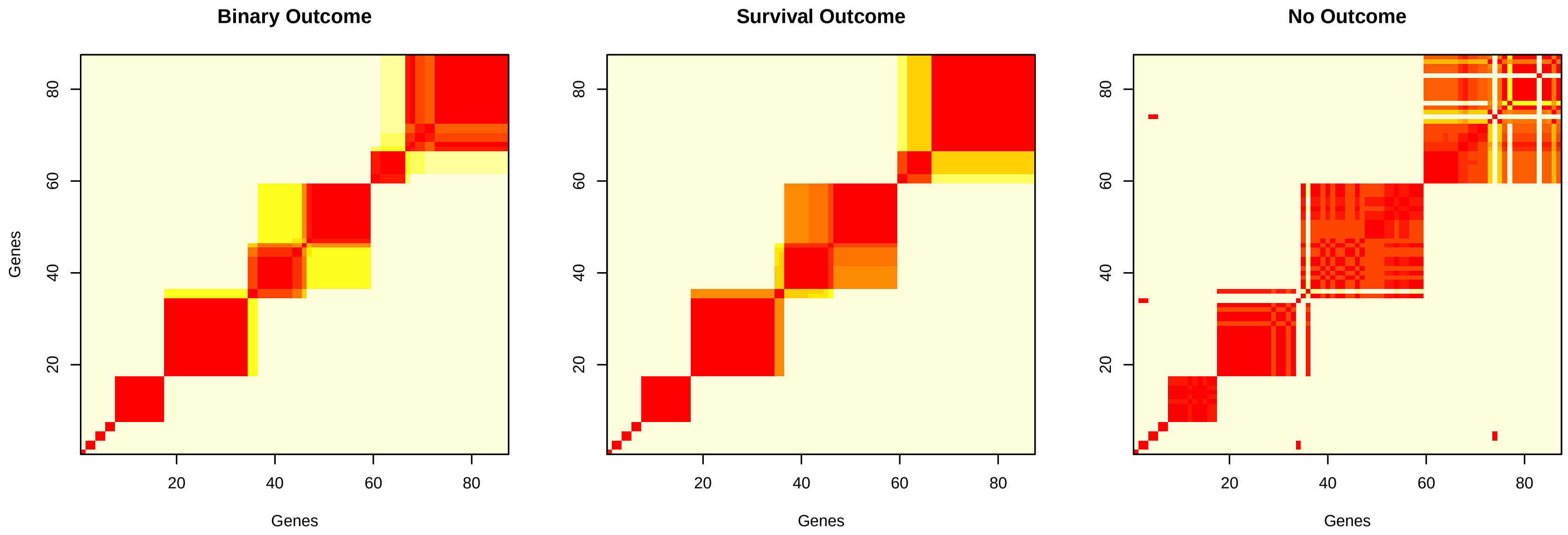

Figure 2.3: Cluster heat maps for Glue Grant trauma data with a binary outcome (complicated vs. uncomplicated recovery), survival outcome (time to recovery), and no outcome.

can act as prognostic markers in predicting recovery among trauma patients, making it possible to determine the posterior predictive probability that a future patient will experience the event. By augmenting the data with latent continuous variables, we are able to utilize the methods developed in Chapter 1. When applied to the Glue Grant data, we can determine the probability that a pair of genes are in the same cluster and identify groups of genes that tend to cluster together. Our approach does not require specification of the exact number of clusters and is also adaptable to situations where there are more genes than subjects.

A limitation we foresee is that a more comprehensive stochastic search may be necessary to fully explore the partition space. Implementing a reversible-jump MCMC, or adding steps where sets of genes are combined and removed, may improve the mixing of cluster membership. Several extensions of our methodology are possible as well. The latent variable approach can be extended to accommodate multinomial outcomes by allowing the response to take on more categories. Currently, our survival model assumes a stationary process where the chance of recovery at every time point is independent of how long the patient has already been in the hospital. Our model can be thought of as a discrete analog to the exponential distribution. We can extend our survival method to have non-constant hazards by specifying a different parameter for each time. For example, indexing the intercept by time would allow for a different hazard at every time point and would provide additional flexibility to the model.

Outcome-Driven Clustering of Longitudinal Gene Expression Data

Jessie J. Hsu, Dianne M. Finkelstein, and David A. Schoenfeld

Department of Biostatistics, Harvard School of Public Health
Biostatistics Center, Massachusetts General Hospital

3.1 Introduction

Conducting microarray experiments has become standard procedure in the biological sciences. Each experiment generates expression data on tens of thousands of genes and is often repeated several times under different experimental conditions, further increasing the dimensionality of the data. Reducing the dimension of longitudinal microarrays through clustering can elucidate underlying disease mechanisms and pathways. We will present a way to cluster high-dimensional longitudinal gene expression data that utilizes clinical outcomes to drive the clustering.

Analyzing the entire time-course gene expression is appealing because gene expression is naturally a temporal process that is constantly regulated and changing. However, methods relating to baseline microarrays do not translate directly to longitudinal settings because they do not capture the time course patterns and the ordered nature of the time index (Bar-Joseph, 2004). The key to analyzing longitudinal genes is to specify a model for the time-dependent gene expression trajectory. Ramoni *et al.* (2002) modeled the correlation between successive measurements of the same gene with autoregressive equations. Luan (2003) modeled the trajectory using splines, suggesting that the flexibility of splines can account for non-linear relationships between genes at different time points and for data measured at unevenly spaced time intervals. More recently, methods have been developed for modeling longitudinal gene expression data using random effects in order to select differentially expressed time-course genes (Storey *et al.*, 2005; Rajicic *et al.*, 2006).

Clustering longitudinal gene expression is based on the idea that similar temporal profiles are involved in similar biological processes. Standard distance-based clustering methods assume observations of each gene are independent and identically distributed. For example, K-means clustering treats the input as a vector of independent samples. This is inappropriate for clustering longitudinal data because it does not account for correlation between successive observations of the same gene nor does it account for correlation between genes in the same cluster. Normal mixture models are also too simplistic for describing time-course dynamics (McLachlan *et al.*, 2002). Booth *et al.* (2008) and Ng *et al.* (2006) presented solutions for clustering correlated gene expression profiles by proposing models with random effects that are shared among correlated measurements of expression on the same gene and among gene expressions from the same cluster.

For our contribution, we seek to use patient outcome to inform the discovery of gene clusters with the hope that the resulting clusters provide a more coherent depiction of the underlying biological mechanism. The scenario of clustering with a normally distributed continuous outcome was considered in Chapter 1. Here, we extend our previous methodology to the longitudinal microarray setting with binary and failure time outcomes. To accommodate longitudinal measurements, we build upon the random-effects model described in Rajicic *et al.* (2006) and assume a linear gene trajectory by including a random intercept and a random slope for time. These random effects induce correlation between genes in the same cluster, and the random slope captures the relationship between repeated observations of the same gene over time.

The other challenge that we will deal with is the issue of specifying the relationship between the non-continuous outcome and the longitudinal expression data. The relationship between binary outcomes and baseline gene expression data has been thoroughly studied and is usually framed as a classification problem (Ring and Ross, 2002; Quackenbush, 2006; Hastie *et al.*, 2001a). The association between baseline microarray data and survival outcomes can be studied in a similar manner and is usually modeled using a Cox proportional hazards model. Typically, the objective of these studies is to determine the risk of experiencing the event that is associated with the observed gene expression measurements (Bøvelstad *et al.*, 2007; Wieringen *et al.*, 2009).

Despite established methods for analyzing binary and survival data, these models become difficult to fit when considered jointly with clustering longitudinal microarrays.

Sha *et al.* (2004) considers a similar situation of relating genes to categorical outcome variables in order to identify differentially expressed genes. They followed the methods developed in Albert and Chib (1993) and augmented the data with a latent continuous representation of the outcome. They specified that the probability of having the outcome depends on the latent variable which is sampled by Markov chain Monte Carlo (MCMC) methods. We adopt this approach as well because having the latent variable simplifies the model fitting procedure and allows us to apply the MCMC clustering algorithm presented in Chapter 1.

In Section 3.2, we begin by summarizing the previous chapters. In Sections 3.2.1 and 3.2.2, we present the model for clustering longitudinal microarrays and the binary response model. All the model parameters are assigned non-informative prior distributions in Sections 3.2.3, and Section 3.2.4 describes how each of the model parameters is sampled and how the clusters are found by the MCMC algorithm. An important aspect of this is the number of clusters which is discussed in Section 3.2.5. We expand our model to accommodate failure time outcomes in Section 3.2.6. In Section 3.3, we illustrate our clustering method using trauma data from the Inflammation and Host Response to Injury Program. We end with a discussion in Section 3.4.

3.2 Methods

We aim to cluster longitudinal gene expression data into several groups based on similarities in their expression trajectories and their association to binary and failure time outcomes. We build upon the methods proposed in Chapters 1 and 2. Chapter 1 allowed for a continuous outcome to drive the clustering of baseline microarray data and modeled gene expression with patient-cluster specific random effects that accounted for correlation between genes in the same cluster. In Chapter 2, we considered binary and failure time outcomes to drive the clustering of baseline arrays. A novel aspect of our previous methodology is that the random effects do not need to be estimated because we integrate over them such that they no longer appear in the likelihood. In this chapter, we continue to work with the marginal likelihood. By not having to sample all the random effects, the MCMC algorithm performs with increased stability.

For the remainder of this section, we will introduce the joint model for clustering longitudinal gene expression data. We will describe in detail the data augmentation approach that relates the gene clusters to binary outcomes and extend the method to failure time outcomes. Then we will describe the Bayesian MCMC procedure for estimating parameters and finding clusters for both types of outcomes.

3.2.1 Model for Clustering Longitudinal Gene Expression Data

In the longitudinal microarray setting, we assume that groups of genes behave similarly in their expression patterns over time. For every subject i, $i = 1, \dots, N$, we observe Y_i, a matrix of gene expression values measured at various time points t. The temporal response of the genes is approximated by a random effects model where we assume a linear change in the gene expressions over time. We allow the dependence among genes in the same cluster to be induced by subject-cluster specific random intercepts, c_{ik}. Correlation between genes at different times is induced by subject-cluster specific random slopes for the time effect, d_{ik}. Therefore, for gene j belonging in cluster k, the model for gene expression is

$$Y_{ijt} = c_{ik(j)} + d_{ik(j)}t + \epsilon_{ijt}. \tag{3.1}$$

Supposing that all the data has been centered at zero, we specify that $c_{ik} \sim N(0, \tau^2)$, $d_{ik} \sim N(0, \nu^2)$, and $\epsilon_{ijt} \sim N(0, \sigma^2)$. Also, we assume independence between c_{ik} and $c_{ik'}$ and between d_{ik} and $d_{ik'}$ for $k \neq k'$. This implies that genes in different clusters have covariance equal to zero. We use indicator variables, ϕ_{jk}, to denote the membership of gene j in cluster k and assume the vector of indicators associated with each gene has a multinomial distribution with probabilities ω_k. While these are latent, we will later show how they are used to find the clusters.

For purposes of inference, we will use sufficient statistics for c_{ik} and d_{ik} instead of all the expression data. Let h index the times when arrays are collected. Assuming that patient i has arrays for T_i time points, the sufficient statistics are $\sum_{h=1}^{T_i} Y_{ijt_h}$ and $\sum_{h=1}^{T_i} t_h Y_{ijt_h}$. For the remainder of the chapter, let Y_i denote the expression data in terms of sufficient statistics.

3.2.2 Model for Clustering Genes using a Binary Outcome

For every patient i, let X_i be a Bernoulli(p) distributed random variable that indicates whether or not the patient experienced the event. We assume that p, the probability of experiencing the event, depends on a normally distributed latent variable Z_i. Known as the data augmentation approach (Tanner and Wong, 1987; Albert and Chib, 1993), Z_i can be thought of as an unmeasured underlying process that is directly related to the value of the observed binary response. Introducing Z_i into the model facilitates the Bayesian model fitting procedure because once we estimate Z_i, we no longer need X_i and the likelihood becomes a simple multivariate normal distribution. Obtaining these latent parameters is readily achieved by adding an extra step into the MCMC sampling scheme, the details of which are presented in Section 3.2.4. Augmenting the data to include Z_i requires the specification of a function that links the relationship between X_i and Z_i. We use the probit link to describe the relationship, where Φ is the cumulative density function of the standard normal distribution:

$$P(X_i = 1) = \Phi(Z_i) \tag{3.2}$$

The dependence of X_i on Z_i is straightforward; a smaller value of Z_i implies that X_i is more likely to be zero and a larger value of Z_i implies that X_i is more likely to be one.

So far, separate models have been proposed for the clusters of gene expression data and for the binary outcome. Now we describe the relationship between the two components. The gene clusters are related to the binary outcome by a linear relationship between the cluster random effects and Z_i, the continuous latent representation of the binary response:

$$Z_i = \mu + \sum_{k=1}^{K} \beta_k (c_{ik(j)} + d_{ik(j)}). \tag{3.3}$$

This is the standard linear regression formulation where β acts as coefficients that describe the effect of the cluster centers and the cluster slopes on the continuous latent outcome. For every cluster, the intercept and slope are combined into an average value for every cluster and are associated with one β. We note that it would have been possible to have different values of β_k for c_{ik} and d_{ik} to allow for different effects of intercept and slope on outcome. Also, we noticed that when β was unconstrained, it tended to increase without bound. The reason this model may not converge is because the probit model is much coarser than the normal model and causes the unconstrained coefficients to be non-identifiable. We found that convergence occurs when β is constrained to lie on the unit sphere. A convenient distribution for points on a sphere is the von Mises-Fisher distribution. We will describe how to generate random variables from this distribution in Section 3.2.4.

Once we have obtained values for Z_i, it is not necessary to account for X_i anymore because X_i gives no additional information about Z_i. This is a very useful property because the joint distribution of Y_i and Z_i is multivariate normal and is much easier to work with as compared to the joint distribution of Y_i and X_i. We detail the joint distribution of Y_i and Z_i here because all of the distributions that are used in the MCMC sampling algorithm are based on this density. Recall that the random effects have been integrated out and that $Y_i = (\sum_{h=1}^{T_i} Y_{ijt_h}, \sum_{h=1}^{T_i} t_h Y_{ijt_h})$ are the sufficient statistics.

The distribution of $f(Y_i, Z_i | \sigma, \tau, \nu, \phi, \omega, \mu, \beta)$ is multivariate normal with mean $(0, \ldots, 0, \mu)'$ and covariance matrix Σ, where

$$\Sigma = \begin{bmatrix} \Sigma_{YY} & \Sigma_{YZ} \\ \Sigma_{ZY} & \Sigma_{ZZ} \end{bmatrix}.$$

Define the following elements for Σ:

$$\begin{aligned}
\Sigma_{YY} &= \begin{bmatrix} A & B \\ B & C \end{bmatrix} \\
\Sigma'_{YZ} &= \Sigma_{ZY} = \begin{bmatrix} D & E \end{bmatrix} \\
\Sigma_{ZZ} &= (\tau^2 + \nu^2) \sum_{k=1}^{K} \beta_k^2
\end{aligned}$$

Suppose we let $u = (1, \ldots, J)$ and $v = (1, \ldots, J)$ index the gene elements in each submatrix and let S_k denote the k^{th} cluster set. Let T_i denote the number of arrays that are available for patient i. Then,

$$\begin{aligned}
A_{u,v} &= \sigma^2 T_i I(u = v) + (\tau^2 T_i^2 + \nu^2 (\sum_{h=1}^{T_i} t_{ih})^2) \sum_{k=1}^{K} I(u, v \in S_k) \\
B_{u,v} &= \sigma^2 (\sum_{h=1}^{T_i} t_{ih}) I(u = v) + (\tau^2 T_i (\sum_{h=1}^{T_i} t_{ih}) + \nu^2 (\sum_{h=1}^{T_i} t_{ih})(\sum_{h=1}^{T_i} t_{ih}^2)) \sum_{k=1}^{K} I(u, v \in S_k) \\
C_{u,v} &= \sigma^2 (\sum_{h=1}^{T_i} t_{ih}^2) I(u = v) + (\tau^2 (\sum_{h=1}^{T_i} t_{ih})^2 + \nu^2 (\sum_{h=1}^{T_i} t_{ih}^2)^2) \sum_{k=1}^{K} I(u, v \in S_k) \\
D_{1,u} &= (\tau^2 T_i + \nu^2 \sum_{h=1}^{T_i} t_{ih}) \sum_{k=1}^{K} I(u \in S_k) \beta_k \\
E_{1,u} &= (\tau^2 \sum_{h=1}^{T_i} t_{ih} + \nu^2 \sum_{h=1}^{T_i} t_{ih}^2) \sum_{k=1}^{K} I(u \in S_k) \beta_k
\end{aligned}$$

3.2.3 Prior Distributions for Model Parameters

Non-informative prior distributions are specified for every parameter. The priors for the hierarchical standard deviation components σ, τ, and ν are uniform densities on a wide range. This is approximately equivalent to specifying an Inverse-χ^2 prior distribution for σ^2, τ^2, and ν^2. The vector of cluster probabilities, ω, has a conjugate symmetric Dirichlet$(\alpha, \ldots, \alpha)$ prior, where smaller values of α reflect a prior belief that there should be fewer clusters and larger values suggest the opposite. The cluster membership variable has a conjugate multinomial prior that depends on the weights, $P(\phi_{jk} = 1) = \omega_k$. The intercept term μ is given a non-informative uniform prior, and a von Mises-Fisher prior distribution that is uniform on the K-dimensional unit sphere is specified for the vector β.

3.2.4 MCMC Algorithm for Clustering Genes using a Binary Outcome

MCMC methods are iterative procedures that allow one to sample from a variety of probability distributions. We wish to know the posterior distribution of Z and all the model parameters. In order to accomplish this, the MCMC algorithm iterates between draws from the full conditional posterior distributions of Z, σ, τ, ν, ϕ, ω, μ, and β. As a result, the collection of samples that are obtained is representative of the posterior density. The steps are detailed below.

1. Sample the latent variable, Z_i.

   Write

$$\begin{aligned}
&f(Z_i|Y_i, X_i, \sigma, \tau, \nu, \phi, \omega, \mu, \beta)\\
&\propto f(X_i|Z_i, Y_i, \sigma, \tau, \nu, \phi, \omega, \mu, \beta) f(Z_i|Y_i, \sigma, \tau, \nu, \phi, \omega, \mu, \beta)\\
&= [\Phi(Z_i)I(X_i = 1) + \Phi(-Z_i)I(X_i = 0)] f(Z_i|Y_i, \sigma, \tau, \nu, \phi, \omega, \mu, \beta). \qquad (3.4)
\end{aligned}$$

   Since both Z_i and Y_i are normally distributed, $f(Z_i|Y_i, \sigma, \tau, \phi, \omega, \mu, \beta)$ is also normally distributed with mean

$$\mu + \Sigma_{YZ}\Sigma_{ZZ}^{-1}(Y_i)$$

   and covariance

$$\Sigma_{ZZ} - \Sigma_{ZY}\Sigma_{YY}^{-1}\Sigma_{YZ}$$

   where $\Sigma_{YY}, \Sigma_{YZ}, \Sigma_{ZY}$, and Σ_{ZZ} were previously defined.

   Due to the difficulty of drawing Z_i directly, we implement an acceptance-rejection algorithm for sampling Z_i. Candidate values of Z_i are first drawn from $f(Z_i|Y_i, \sigma, \tau, \nu, \phi, \omega, \mu, \beta)$. The candidate values are accepted with probability $\Phi(Z_i)$ if $X_i = 1$ and with probability $\Phi(-Z_i)$ if $X_i = 0$. We demonstrated in Chapter 2 that this generates values of Z_i with the correct distribution conditional on the outcome X_i and the model parameters.

2. Sample the variance parameters: σ^2, τ^2, ν^2.

   To simulate the variance parameters, the Metropolis-Hastings algorithm is used. We draw candidate values from an inverse gamma proposal distribution with shape parameter s and scale parameter s/θ. This is approximately equivalent to specifying an Inverse-χ^2 prior distribution. These tuning parameters are determined experimentally during initial runs to accept proposed samples at the recommended rate of $40\% - 45\%$ (Gelman *et al.*, 2004). The proposed values are accepted with probability proportional to the ratio of the posterior density evaluated at the current value to the posterior density evaluated at the candidate value.

3. Sample the cluster weights, ω.

   The probability weights associated with belonging in each cluster are sampled via a Gibbs sampling step. For Gibbs sampling, we draw directly from the full conditional posterior distribution of ω:

   $$f(\omega|\sigma,\tau,\nu,\phi,\mu,\beta,Y,Z) \propto f(Y,Z|\omega,\sigma,\tau,\nu,\phi,\mu,\beta)f(\omega)$$

   The full conditional distribution of ω is Dirichlet$(\alpha + n_1, ..., \alpha + n_K)$, where n_k is the number of genes in the k^{th} cluster at the current iteration. We found that $\alpha = 1$ provides a reasonable result.

4. Sample the cluster memberships, ϕ.

   The cluster membership parameter is also obtained by Gibbs sampling. The clustering space is explored in a stochastic search where each gene is moved into every cluster and the likelihood of belonging in each of the K clusters is calculated. The likelihood of each rearrangement weighted by the current value of ω then becomes the updated multinomial sampling probabilities. Thus, the cluster membership of gene j is sampled from a multinomial distribution with probabilities proportional to the weighted likelihood given the current parameter values.

The full conditional distribution of ϕ is given by

$$f(\phi_j|Y, Z, \sigma, \tau, \nu, \mu, \beta, \omega) \propto \prod_{k=1}^{K}(f(Y, Z|\sigma, \tau, \nu, \mu, \beta, \omega) * \omega_k)^{\phi_{jk}}$$

and we sample directly from this multinomial distribution.

5. Sample the regression intercept, μ.

   The intercept term is sampled using Metropolis-Hastings. A candidate value of μ is drawn from a normal proposal distribution with variance one. Again, the candidate value is accepted with probability proportional to the ratio of the posterior density evaluated at the current value against the posterior density evaluated at the proposed value.

6. Sample the regression coefficients, β.

   The vector of coefficients β is obtained by Metropolis-Hastings sampling as well. As mentioned earlier, we constrain β to exist on the K-sphere such that the sum of squares equals one. The von Mises-Fisher distribution for a unit vector of dimension K is suitable for drawing candidate values with the desired constraint. Following the steps described in Wood (1994) on how to sample from this distribution, the result is a K-dimensional unit vector with mean direction $(0, \ldots, 0, 1)'$. Applying QR decomposition rotates the vector so that the proposed value of β becomes the mean direction.

3.2.5 Determining the Number of Clusters

We have been using K to denote the numbers of clusters. To be more specific, K is actually the number of empty and non-empty clusters. For our algorithm, only the maximum value of K needs to specified in advance. However, the number of non-empty clusters does not need to be specified nor does it need to be sampled separately as a parameter in the algorithm because it is allowed to change with every iteration when cluster membership is sampled. Since ω_k is positive for all k regardless of cluster size, the multinomial probabilities of belonging in each cluster are always non-zero. Therefore, there is always a chance that a cluster will end up with no genes or that an empty cluster will become filled at any given iteration due to the probabilistic nature of the allocation.

3.2.6 Extension to Clustering Genes using a Failure Time Outcome

In this section, we extend our model to accommodate failure time data by treating time-to-event as a series of binary observations at a fixed number of discrete time points that indicate whether or not the event of interest has occurred yet. This vector of indicator variables is denoted by $X_i(t)$. If patient i experiences the event at time t, then $X_i(t) = 1$; otherwise if the patient has not yet had the event by time t, then $X_i(t) = 0$.

As in the case with a binary outcome, we augment the data by assuming that the hazard of recovery at any given time depends on a latent continuous variable, $Z_i(t)$. The difference here is that there is a value of $Z_i(t)$ for every time point that the event is evaluated. The values of t do not necessarily need to correspond to the times that the longitudinal genes are measured. $Z_i(t)$ is modeled as

$$Z_i(t) = \mu + \sum_{k=1}^{K} \beta_k (c_{ik(j)} + d_{ik(j)} t) + \xi_{it}. \tag{3.5}$$

We assume $\xi_{it} \sim N(0, 1)$. Here, we use the actual value of t in the model. Again, more complex models are possible. We could have specified separate effects for c_{ik} and $d_{ik}t$ with a different β_k for each. Let $l = 1, \ldots, L_i$ index the times that the event is evaluated. We assume a proportional hazards model where the hazard of the event at any particular time depends on $Z_i(t_l)$ as follows:

$$P(X_i(t_l) = 1 | X_i(t_{l-1}) = 0) = \Phi(Z_i(t_l)). \tag{3.6}$$

The hazard of experiencing the event is different at every time t_l, and as $Z_i(t_l)$ increases, the hazard of the event increases. To model the amount of time it takes for a subject to have an event, we assume a negative binomial distribution where the probability of the event is the hazard of recovery as shown in (3.6). Assuming that patient i is evaluated for the event at L_i time points, the probability of recovering at the l^{th} time point then becomes $\prod_{l=1}^{L_i-1} \Phi(-Z_i(t_l))\Phi(Z_i(t_{L_i}))$.

The MCMC for fitting the survival model follows almost exactly the same steps as before, except that here the full conditional posterior distribution of the vector Z_i is a product across all the time points:

$$\begin{aligned} & f(Z_i|Y_i, X_i, \sigma, \tau, \nu, \phi, \omega, \mu, \beta) \\ & = \prod_{l=1}^{L_i} [\Phi(Z_i(t_l))I(X_i(t_l) = 1) + \Phi(-Z_i(t_l))I(X_i(t_l) = 0)] f(Z_i|Y_i, \sigma, \tau, \nu, \phi, \omega, \mu, \beta). \end{aligned} \tag{3.7}$$

The conditional distribution $f(Z_i|Y_i, \sigma, \tau, \nu, \phi, \omega, \mu, \beta)$ is normal with mean

$$\mu + \Sigma_{YZ}\Sigma_{ZZ}^{-1}(Y_i)$$

and covariance

$$\Sigma_{ZZ} - \Sigma_{ZY}\Sigma_{YY}^{-1}\Sigma_{YZ}$$

where we redefine

$$\begin{aligned} \Sigma'_{YZ} &= \Sigma_{ZY} = \begin{bmatrix} D^* & E^* \end{bmatrix} \\ \Sigma_{ZZ} &= F \end{aligned}$$

Again, let $u = (1, \dots, J)$ and $v = (1, \dots, J)$ index the gene elements in each submatrix and let $l = (1, \dots, L_i)$ index the event time. Then,

$$\begin{aligned} D^*_{l,u} &= (\tau^2 T_i + \nu^2 t_l \sum_{h=1}^{T_i} t_{ih}) \sum_{k=1}^{K} I(u \in S_k)\beta_k \\ E^*_{l,u} &= (\tau^2 \sum_{h=1}^{T_i} t_{ih} + \nu^2 t_l \sum_{h=1}^{T_i} t_{ih}^2) \sum_{k=1}^{K} I(u \in S_k)\beta_k \\ F_{l,l'} &= (\tau^2 + \nu^2 t_l t_{l'}) \sum_{k=1}^{K} \beta_k^2 \end{aligned}$$

Candidate values of Z_i are drawn from $f(Z_i|Y_i, \sigma, \tau, \nu, \phi, \omega, \mu, \beta)$ and are accepted with probability $\prod_{l=1}^{L_l} [\Phi(Z_i(t_l))I(X_i(t_l) = 1) + \Phi(-Z_i(t_l))I(X_i(t_l) = 0)]$.

3.3 Application

The Inflammation and Host Response to Injury research program, also known as the Glue Grant, is an interdisciplinary study of the genomic changes that occur after a patient experiences a traumatic injury. The data consists of longitudinal expression values measured on thousands of genes as well as various clinical measurements and recovery endpoints for every patient. Patients are followed for 28 days and microarrays are available for up to seven time points that are taken at scheduled study visits following injury. The number of arrays that are available at each time point is presented in Table 3.1. For our analysis, we use arrays collected starting on day 4. This is because our model assumes linearity in the time effect and the expression trajectories appear highly non-linear for the first few days after injury.

Table 3.1: Glue Grant array count on various days following injury.

Day	0	1	4	7	14	21	28
Arrays	167	159	147	135	86	53	30

The gene expression values are measured from blood leukocyte cells and have been pre-processed using dChip, log-transformed and centered prior to analysis. We use a subset of 87 genes for our cluster analysis that have been pre-selected by Glue Grant investigators to be those that had significant differential expression with at least a two-fold difference between patients with a clinical outcome of complicated versus uncomplicated recovery. Complicated recovery implies the patient had a time to recovery of more than 14 days, and uncomplicated patients recovered in less than 14 days. We define the event of interest to be complicated versus uncomplicated recovery class for the binary outcome.

For the survival outcome, the response measurement is time to recovery. Time to recovery is calculated as the maximum time of cardiovascular, hematologic, hepatic, renal, or respiratory recovery. The median time to recovery was 7 days. We observe a vector of binary responses for every patient with one response at every time point. The responses are recorded at fixed discrete time points that are the same for every patient. For the Glue Grant analysis, we use a recovery indicator that is recorded every four days from day 0 to day 28. We assume that recovery can only occur once for every patient and that once recovery has occurred, the patient is no longer at risk. Recovery is the only absorbing state in the model; once a patient recovers, the patient is considered to have reached the end of the study. If patients do not recover during the course of the study or if they die prior to the last observed day, they are censored on day 28 and have an observed indicator vector that consists of all zeroes. Since very few subjects died from their injuries, mortality was not considered an appropriately sensitive variable for informing distinct clusters. In addition to the five patients who died within the first 28 days, seven patients did not recover within the first 28 days. Both of these groups of patients are censored at 28 days, since obviously none of these patients will recover by the end of the study.

For both the binary and survival analyses, we ran eight chains with over-dispersed starting values. The maximum number of clusters K was set to be 10. We ran 2000 iterations until convergence and discarded 500 iterations as burn-in. Since we only make single gene transitions when sampling cluster membership, there is a tendency to under-explore the partition space. Therefore, several chains at different starting values were implemented and subsequently averaged for purposes of inference. Estimates of the parameters are shown in Table 3.2.

A visual representation of the clustering is presented in Figure 3.1. Cluster membership is depicted as a heat map that shows the proportion of iterations that every pair of genes is assigned to the same cluster. In the event of label switching, summarizing the output as a heat map aids in visualizing the groups and prevents us from having to follow the movement of the genes at every iteration. Furthermore, heat map visualization has the advantage of providing information about the uncertainty surrounding the allocations. The Bayesian approach does not assume there is only one correct parti-

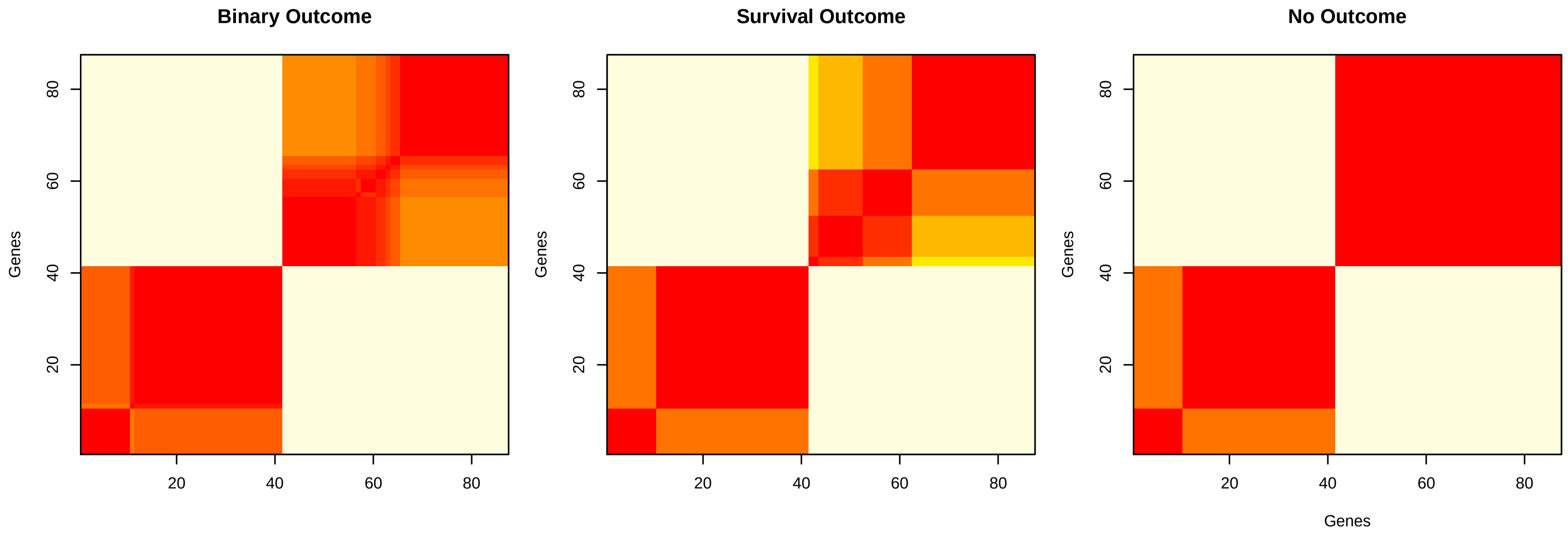

Figure 3.1: Cluster heat maps for longitudinal Glue Grant trauma data with a binary outcome (complicated vs. uncomplicated recovery), survival outcome (time to recovery), and no outcome.

Table 3.2: Results of longitudinal Glue Grant trauma data analysis with a binary outcome (complicated vs. uncomplicated recovery) and survival outcome (time to recovery).

	Binary outcome		*Survival outcome*	
	Mean	95% credible interval	Mean	95% credible interval
σ	2.247	(1.575, 2.878)	1.887	(0.923, 2.667)
τ	2.127	(0.924, 2.813)	1.823	(0.826, 2.655)
ν	1.839	(0.582, 2.531)	1.437	(0.428, 3.009)
μ	0.731	(-3.523, 3.057)	-0.240	(-0.764, 0.010)

tion, but rather that there is a distribution of partitions where some are more likely than others. The similarity, or concordance, between two genes is measured by the percentage of iterations that they are assigned to the same cluster. Concordance is represented by a color gradient and ranges from 0% (white) to 100% (red). In all three cases, there are two large non-overlapping groups. When outcome is included however, genes 42-87 have several breakdown combinations, implying that several partitions have similar posterior probabilities. On the other hand, when outcome is not used, genes 42-87 cluster together all the times and do not break down into smaller groups. It is possible that including outcome can produce more clusters if the relationship to gene expression is strong.

A plot of the expression trajectories of four representative genes is shown in Figure 3.2. The gene numbers correspond to the axes labels in the heat map. According to the heat maps, genes 1 and 20 cluster together about half the time, and genes 45 and 80 cluster together about half the time. However, the pairs never cluster together. One can see in the plot that there are differences in their trajectories that correspond to the way they cluster. Though not entirely linear, genes 1 and 20 have a general downward trend, and genes 45 and 80 have a general upward trend.

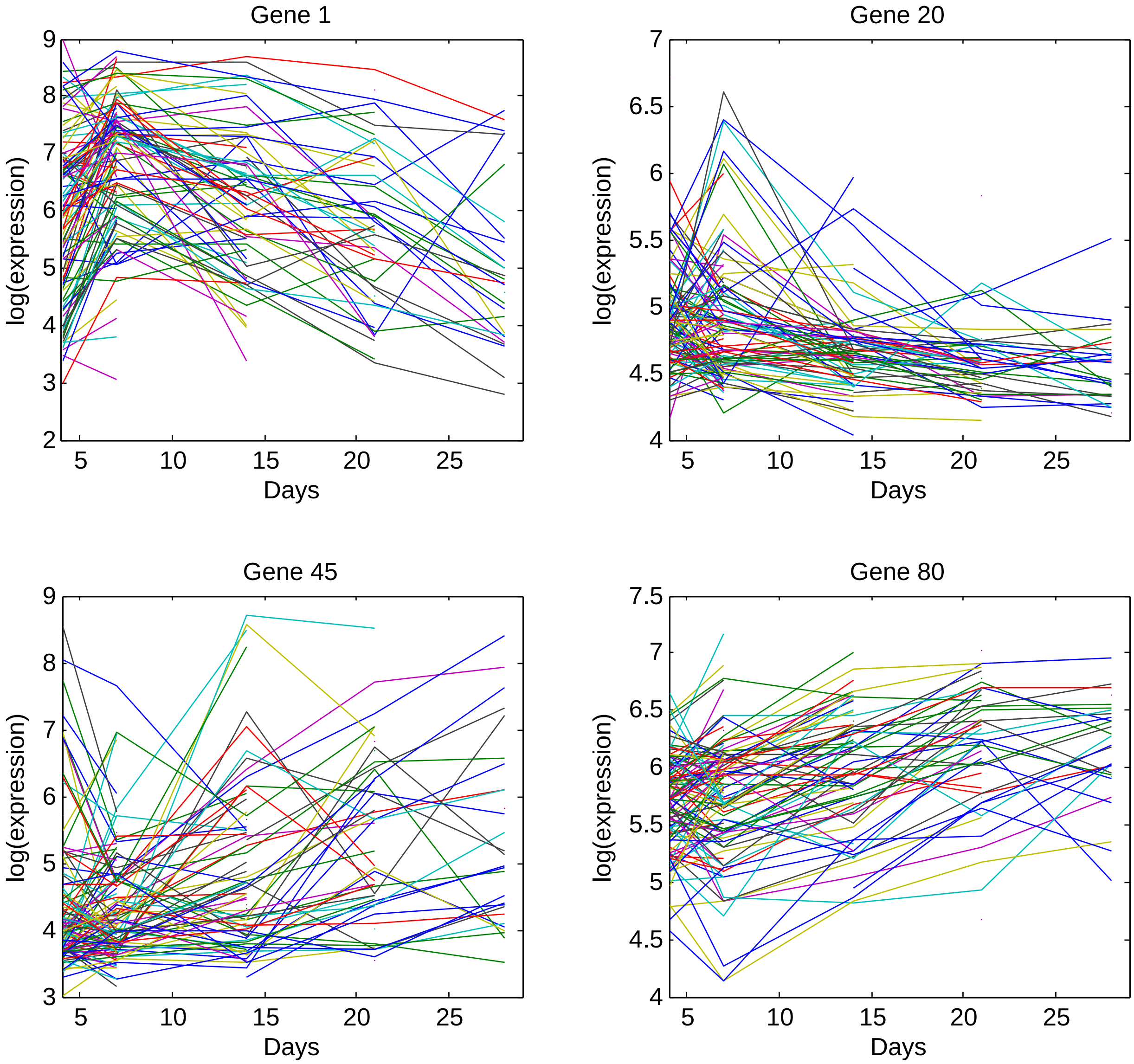

Figure 3.2: Gene expression trajectories of representative genes from different clusters.

3.4 Discussion

In this chapter, we have provided a Bayesian approach for clustering longitudinal microarray data that uses patient outcome to inform the partitions. Our method is primarily intended for exploratory purposes and pattern discovery. However, the clustering model can also be used for prediction where the clusters act as prognostic markers in predicting the outcome.

An advantage of the Bayesian framework is that all the parameters, including those for cluster membership, are associated with a probability distribution. This implies that genes do not necessarily interact with the same group of genes all the time, but may in fact interact with several different networks, a perspective that seems quite reasonable. Our model allows for correlation between genes in the same cluster and between repeated measurements of the same gene. We do not need to specify the exact number of clusters but only need to specify the maximum number of clusters. Additionally, the longitudinal arrays do not need to be measured at the same time as the evaluation of recovery status.

Our model can be extended to more complex settings with the inclusion of additional parameters. Though this would lead to fewer assumptions, the additional parameters would need to be sampled in extra MCMC steps. For example, there can be separate coefficients for the intercept and slope effects in the model for outcome. Furthermore, a non-linear relationship with time can be specified for the gene trajectory. Splines can also be used to model the time-course expression to provide the most amount of flexibility.

The information encoded in microarray data has the potential to contain new insights about the human genome that could eventually lead to new developments in medicine. Uncovering the underlying cluster structure of gene expression data and determining the functional properties of the gene clusters will help us understand the biological basis of events following traumatic injury. Developing a reliable method of predicting patient recovery can save valuable resources that are required for careful monitoring of every patient. If we can successfully accomplish these objectives, we can develop intervention strategies that have the potential of putting more patients on the road to recovery.

References

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

Bair, E. and Tibshirani, R. (2004). Semi-supervised methods to predict patient survival from gene expression data. *PLoS Biol*, **2**, e108.

Bar-Joseph, Z. (2004). Analyzing time series gene expression data. *Bioinformatics*, **20**(16), 2493–2503.

Booth, J. G., Casella, G., and Hobert, J. P. (2008). Clustering using objective functions and stochastic search. *Journal of the Royal Statistical Society: Series B*, **70**(1), 119–139.

Bøvelstad, H., Nygård, S., Størvold, H., Aldrin, M., Borgan, O., Frigessi, A., and Lingjaerde, O. (2007). Predicting survival from microarray data - a comparative study. *Bioinformatics*, **23**(16), 2080–2087.

Dettling, M. and Bühlmann, P. (2002). Supervised clustering of genes. *Genome Biology*, **3**(12), research0069.1–research0069.15.

Dudoit, S., Fridlyand, J., and Speed, T. P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association*, **97**(457), 77–87.

Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*, **95**(25), 14863–14868.

Fraley, C. and Raftery, A. E. (1998). How many clusters? Which clustering method? Answers via model-based cluster analysis. *The Computer Journal*, **41**(8), 578–588.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, **1**, 1–19.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian Data Analysis*. New York: Chapman & Hall/CRC, 2nd edition.

Ghosh, D. and Chinnaiyan, A. M. (2002). Mixture modelling of gene expression data from microarray experiments. *Bioinformatics*, **18**(2), 275–286.

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., and Bloomfield, C. D. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.

Gui, J. and Li, H. (2005). Penalized cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. *Bioinformatics*, **21**(13), 3001–3008.

Hartigan, J. A. (1975). *Clustering algorithms*. New York: Wiley.

Hastie, T., Tibshirani, R., and Friedman, J. H. (2001a). *The elements of statistical learning: data mining, inference, and prediction*. New York: Springer-Verlag.

Hastie, T., Tibshirani, R., Botstein, D., and Brown, P. (2001b). Supervised harvesting of expression trees. *Genome Biology*, **2**(1), research0003.1–research0003.12.

Jung, S.-H., Owzar, K., and George, S. L. (2005). A multiple testing procedure to associate gene expression levels with survival. *Statistics in Medicine*, **24**(20), 3077–3088.

Luan, Y. (2003). Clustering of time-course gene expression data using a mixed-effects model with B-splines. *Bioinformatics*, **19**(4), 474.

McLachlan, G. J., Bean, R., and Peel, D. (2002). A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics*, **18**(3), 413–422.

Ng, S. K., McLachlan, G. J., Wang, K., Ben-Tovim Jones, L., and Ng, S.-W. (2006). A mixture model with random-effects components for clustering correlated gene-expression profiles. *Bioinformatics*, **22**(14), 1745–1752.

Nguyen, D. V. and Rocke, D. M. (2002). Partial least squares proportional hazard regression for application to DNA microarray survival data. *Bioinformatics*, **18**(12), 1625–1632.

Park, P. J., Tian, L., and Kohane, I. S. (2002). Linking gene expression data with patient survival times using partial least squares. *Bioinformatics*, **18**(suppl 1), S120–S127.

Quackenbush, J. (2006). Microarray analysis and tumor classification. *New England Journal of Medicine*, **354**(23), 2463–2472.

Rajicic, N., Finkelstein, D. M., and Schoenfeld, D. A. (2006). Survival analysis of longitudinal microarrays. *Bioinformatics*, **22**(21), 2643–2649.

Ramoni, M. F., Sebastiani, P., and Kohane, I. S. (2002). Cluster analysis of gene expression dynamics. *Proceedings of the National Academy of Sciences*, **99**(14), 9121–9126.

Ring, B. and Ross, D. (2002). Microarrays and molecular markers for tumor classification. *Genome Biology*, **3**(5), comment2005.1–comment2005.6.

Sha, N., Vannucci, M., Tadesse, M. G., Brown, P. J., Dragoni, I., Davies, N., Roberts, T. C., Contestabile, A., Salmon, M., Buckley, C., and Falciani, F. (2004). Bayesian variable selection in multinomial probit models to identify molecular signatures of disease stage. *Biometrics*, **60**(3), 812–819.

Storey, J. D., Xiao, W., Leek, J. T., Tompkins, R. G., and Davis, R. W. (2005). Significance analysis of time course microarray experiments. *Proceedings of the National Academy of Sciences of the United States of America*, **102**(36), 12837–12842.

Tadesse, M. G., Sha, N., and Vannucci, M. (2005). Bayesian variable selection in clustering high-dimensional data. *Journal of the American Statistical Association*, **100**(470), 602–617.

Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, **82**(398), 528–540.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, **58**, 267–288.

Vogl, C., Sanchez-Cabo, F., Stocker, G., Hubbard, S., Wolkenhauer, O., and Trajanoski, Z. (2005). A fully Bayesian model to cluster gene-expression profiles. *Bioinformatics*, **21**(suppl 2), ii130–ii136.

Wieringen, W. N. V., Kun, D., Hampel, R., and Boulesteix, A.-l. (2009). Survival prediction using gene expression data: a review and comparison. *Multiple Sclerosis*, **53**(5), 1590–1603.

Wood, A. T. A. (1994). Simulation of the von mises fisher distribution. *Communications in Statistics Simulation and Computation*, **23**(1), 157–164.