Published in final edited form as:

Commun Stat Case Stud Data Anal Appl. 2018 ; 4(1): 18–27. doi:10.1080/23737484.2018.1455542.

Informatively clustering longitudinal microarrays using binary or survival outcome data

Jessie J. Hsu¹, Dianne M. Finkelstein^{2,*}, and David A. Schoenfeld²

¹Genentech, San Francisco, CA, USA

²Massachusetts General Hospital and Harvard University, Boston MA, USA

Abstract

The goal of this research is to discover what groups of genes are associated with the disease process. We use binary and failure time outcomes to inform the clustering of longitudinally-collected microarray data. We propose a linear model with normally distributed cluster-specific random effects for the longitudinal gene expression trajectory. The random effects are linearly related to a latent continuous representation of the outcome, where the probability or hazard of the outcome depends on these latent variables. We apply our method to microarray data collected from trauma patients in the Inflammation and Host Response to Injury project.

Keywords

Clustering; Bayesian; Microarray; Gene expression; Trauma

1 Introduction

The information contained in microarray data has the potential to provide new insights about the human genome that could eventually lead to deeper understanding of the disease process and potential therapeutic interventions. With the rapid development of high-throughput technology in recent years, the conduct of microarray experiments has become ubiquitous. Each experiment generates expression data on tens of thousands of genes and is often repeated several times under different controlled conditions, further increasing the dimensionality of the data. Reducing the dimension of longitudinal microarrays through clustering can elucidate underlying biological processes as well as disease mechanisms and pathways.

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 single time-point 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. In earlier work, Ramoni et al. [2002] modeled

*Corresponding Author: dfinkelstein@mgh.harvard.edu.

the correlation between successive measurements of the same gene with autoregressive equations while 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]).

Once we have characterized a longitudinal course of the genes, the next step would be to determine what genes have similar temporal profiles, as they may also have similar biological process. A clustering algorithm could be used to identify these groups of genes. For example, despite the fact that K -means clustering was designed to cluster N independent observations, each with p correlated features that have K different mean vectors, it is often used by investigators in genomics to cluster the features themselves rather than the observations. Standard distance-based clustering methods assume observations of each gene are independent and identically distributed. 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]). More recently, Booth et al. [2008] and Ng et al. [2006] extended the clustering methods so that they provide solutions for clustering correlated gene expression profiles by proposing models with random effects that are shared among correlated expression measurements on the same gene and among genes from the same cluster.

For our contribution, we seek to use longitudinally collected microarray data as well as patient outcome to additionally inform the discovery of gene clusters with the hope that the resulting clusters provide a more coherent depiction of the underlying disease mechanism. For this we will build on our previous work in Hsu et al. [2015], where we developed a method of model-based clustering of genes into groups that were highly correlated and had a similar effect on a normally-distributed outcome variable. We used a random effects model, where the genes within each cluster could be expressed as the sum of a random effect, specific to the observation and the cluster and an independent error term. The outcome variable was expressed as a linear combination of the random effects of each cluster. The model was fit using an MCMC algorithm. Simulation studies showed that this method which included outcomes in the clustering discovery improved on clustering based only on the genes. To accommodate longitudinal measurements of the genes, we will 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.

For the extension of our clustering algorithm to include information from binary or event-driven outcomes, we need to specify the relationship between the binary or failure 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. [2001]). 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. The methods described in Sha et al. [2004] consider 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 will adopt this approach as well because having the latent variable simplifies the model fitting procedure and allows us to apply a straightforward MCMC clustering algorithm.

In Sections 2.1 and 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 2.3, and Section 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 2.5. We expand our model to accommodate failure time outcomes in Section 2.6. In Section 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 4.

2 Methods

We aim to cluster longitudinal gene expression data into several groups based on similarities in their expression trajectories and their association to the outcome. The basic model we propose is a random effects model where there is a patient-cluster specific random slope and intercept which effects each gene in the cluster. We assume genes in different clusters are independent. In addition, we assume that the outcome (which could be binary or a failure-time outcome) is modeled as a function of the same random slope and intercept. This is appropriate when genes are observed to decrease (or increase) linearly with time. This model allows us to cluster genes with the same trajectory. We build upon the methods proposed in Hsu et al. [2015], which allowed for a continuous outcome to drive the clustering of baseline microarray data and modeled the gene expression with patient-cluster specific random effects. 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. 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.

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} . Suppose gene j belongs to cluster k . Then the model for the gene expression is

$$Y_{ijt} = c_{ik} + d_{ik}t + \varepsilon_{ijt}. \quad (1)$$

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 $\varepsilon_{ijt} \sim N(0, \sigma^2)$. We assume that the trajectory of gene j is independent between subjects. By centering the values of each gene across times and subjects, the average slope and intercept of each gene across patients is zero. 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. We need to specify the maximum number of clusters as we discuss later in section 2.5.

In order to reduce the dimension of the data space, we summarize the data from patient i for gene j measured at times t_{ih} (where h indexes time from 1 to T_i) by

$$\left(\sum_{h=1}^{T_i} Y_{ijt_{ih}}, \sum_{h=1}^{T_i} t_{ih} Y_{ijt_{ih}} \right), \quad (2)$$

which together could be used to estimate c_{ik} and d_{ik} , where k is the cluster to which gene j belongs. For the remainder of the article, let Y_i denote the expression data in terms of the statistics in (2), as a single vector with the pair of sums in (2) as entries for each gene.

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. This allows us to apply methods we developed for normally distributed outcomes in Hsu et al. [2015]. 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 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 this relationship, where Φ is the cumulative density function of the standard normal distribution:

$$P(X_i = 1) = \Phi(Z_i) \quad (3)$$

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:

$$Z_i = \mu + \sum_{k=1}^K \beta_k (c_{ik} + d_{ik}). \quad (4)$$

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. This allows us to link the path of a patient's gene and their outcome. For every cluster, the intercept and slope are combined into an average value 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 but we choose this model for simplicity of presentation and discuss generalizations later. 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 is not a problem 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 2.4 and will discuss the consequences of this assumption in the Discussion.

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.

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

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

Define the following elements for Σ :

$$\Sigma_{YY} = \begin{bmatrix} A & B \\ B & C \end{bmatrix}$$

$$\Sigma'_{YZ} = \Sigma_{ZY} = [D \ E]$$

$$\Sigma_{ZZ} = (\tau^2 + \nu^2) \sum_{k=1}^K \beta_k^2$$

Suppose we let $u = (1, \dots, J)$ and $v = (1, \dots, J)$ 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,

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

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($\alpha, \dots, \alpha$) 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 .

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

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

(5)

Since both Z_i and Y_i are normally distributed, $f(Z_i | Y_i, \sigma, \tau, \nu, \phi, \omega, \mu, \beta)$ 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 Σ_{YY} , Σ_{YZ} , 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 $X_i = 0$. 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 . 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 :

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

The full conditional distribution of ω is Dirichlet($\alpha + n_1, \dots, \alpha + n_K$), 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

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

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 $(0, \dots, 0, 1)'$. Applying QR decomposition rotates the vector so that the proposed value of β becomes the mean direction.

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 be 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.

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

$$Z_i(t) = \mu + \sum_{k=1}^K \beta_k(c_{ik} + d_{ik}t) + \xi_{it}. \quad (6)$$

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 $l = 1, \dots, L_i$ 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:

$$P(X_i(t_l) = 1 \mid X_i(t_{l-1}) = 0) = \Phi(Z_i(t_l)). \quad (7)$$

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 that the probability of the event at each time, assuming it has not yet happened, is as shown in (7). 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:

$$f(Z_i | Y_i, X_i, \sigma, \tau, \nu, \phi, \omega, \mu, \beta) = \prod_{l=1}^{L_i} [\Phi(Z_i(t_l))I(X_i(t_l) = 1) + \Phi(-Z_i(t_l))I(X_i(t_l) = 0)] f(Z_i | Y_i, \sigma, \tau, \nu, \phi, \omega, \mu, \beta). \quad (8)$$

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

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

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 $\prod_{l=1}^{L_i} [\Phi(Z_i(t_l))I(X_i(t_l) = 1) + \Phi(-Z_i(t_l))I(X_i(t_l) = 0)]$.

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 (Storey et al. [2005]). 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 were 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 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.

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. This is more fully described in Rajicic et al. [2006]. 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 2.

A visual representation of the clustering is presented in Figure 1. Three panels represent the inclusion of the binary outcome, the event time (survival) outcome, and no outcome to illustrate the impact of the outcome on the clustering. 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. Genes are sorted according to the value of the mean β for the cluster the gene is in. Sorting the genes by the β associated with their cluster and showing how often genes are clustered together avoids the use of labels and thus the problem of label-switching. Summarizing the output as a heat map aids in visualizing the groups and provides information about the uncertainty surrounding the allocations. The Bayesian approach does not assume there is only one correct partition, 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. However, when outcome is included, 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 time 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. Thus including outcome has affected the clustering in a meaningful way.

A plot of the expression trajectories of four representative genes is shown in Figure 2. The gene numbers correspond to the axes labels on the heat maps. 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. The biological interpretation for these results is given in Hsu et al. [2015].

4 Discussion

In this paper, we have developed a Bayesian approach for clustering longitudinal microarray data that uses patient outcome to inform the clusters. Our method is primarily intended for exploratory purposes and pattern discovery, though the clustering algorithm 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 of genes. Our model allows for correlation between genes in the same cluster and between repeated measurements of the same gene. Additionally, we do not need to specify the exact number of clusters, only the maximum number of clusters. Our methodology is adaptable to situations where there are more genes than patients, and the longitudinal arrays do not need to be measured at the same time as the evaluation of recovery status. Upon applying these methods to the Glue Grant data, we are able to determine the probability that a pair of genes are in the same cluster and identify groups of genes that tend to cluster together. Comparing results of our methods with and without outcome illustrates that outcome has clearly impacted the clustering. It is also possible to find the posterior predictive probability that a future patient will experience the event. The next step would be to address the biological interpretation of the gene clusters, but this is beyond the scope of this paper.

A limitation that we foresee is that a more comprehensive stochastic search may be needed in order 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.

The proposed model was selected to simplify description and computation for the paper, but could be generalized in several ways. First, we assume linearity of the longitudinal microarray data but it is important to assess whether this linearity is appropriate, and if not, to apply transformations that would make it linear. Similarly, for convenience we assumed the β were the same for both the intercept and slope, but it would be possible to have separate coefficients for the intercept and slope effects in the model for outcome. However, more parameters would be added at the expense of complexity since the additional parameters would need to be sampled in extra MCMC steps, and it again would be necessary to check for goodness of fit. We also note that to ensure convergence, we constrained β to lie on a unit sphere. This effects the magnitude of the β , but does not impact either the direction or the relative effect of clusters on the outcome. Also, currently, our survival model assumes a stationary process where the chance of recovery at the next visit is independent of how long the patient has already been in the intensive care unit but we could extend our survival method to have non-constant hazards by specifying a different intercept for each time point. Lastly, the latent variable approach can be extended to accommodate a multinomial outcome by allowing the response to take on more categories.

Acknowledgments

This work was supported by the Joint Interdisciplinary Training Program in Biostatistics and Computational Biology [T32GM074897] and Inflammation and the Host Response to Injury Large-Scale Collaborative Project [2-U54-GM062119] from the National Institute of General Medical Sciences.

References

- Albert, James H; Chib, Siddhartha. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*. 88(422):669–679.1993;
- Bar-Joseph, Ziv. Analyzing time series gene expression data. *Bioinformatics*. 20(16):2493–2503.2004; [PubMed: 15130923]

- Booth, James G; Casella, George; Hobert, James P. Clustering using objective functions and stochastic search. *Journal of the Royal Statistical Society: Series B.* 70(1):119–139.2008; 1467–9868.
- Bøvelstad HM, Nygård S, Størvold HL, Aldrin M, Borgan Ø, Frigessi A, Lingjaerde OC. Predicting survival from microarray data - a comparative study. *Bioinformatics.* 23(16):2080–2087.Aug; 2007 1367–4803 DOI: 10.1093/bioinformatics/btm305 DOI: 10.1093/bioinformatics/btm305 [PubMed: 17553857]
- Gelman, Andrew; Carlin, John B, Stern, Hal S, Rubin, Donald B. *Bayesian Data Analysis.* 2. New York: Chapman & Hall/CRC; 2004.
- Hastie, Trevor; Tibshirani, Robert; Friedman, JH. *The elements of statistical learning: data mining, inference, and prediction.* New York: Springer-Verlag; 2001.
- Hsu, Jessie J; Finkelstein, Dianne M; Schoenfeld, David A. Outcome-driven clustering of microarray data. *PLoS ONE.* 10(11):e0141874.2015; [PubMed: 26562156]
- Luan Y. Clustering of time-course gene expression data using a mixed-effects model with B-splines. *Bioinformatics.* 19(4):474.2003; doi: 10.1093/bioinformatics/btg014doi: 10.1093/bioinformatics/btg014 [PubMed: 12611802]
- McLachlan, Geoffrey J; Bean, Richard; Peel, David. A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics.* 18(3):413–422.2002; [PubMed: 11934740]
- Ng SK, McLachlan GJ, Wang K, Ben-Tovim Jones L, Ng S-W. A mixture model with random-effects components for clustering correlated gene-expression profiles. *Bioinformatics.* 22(14):1745–1752.2006; [PubMed: 16675467]
- Quackenbush, John. Microarray analysis and tumor classification. *New England Journal of Medicine.* 354(23):2463–2472.2006; DOI: 10.1056/NEJMra042342 [PubMed: 16760446]
- Rajicic N, Finkelstein Dianne M, Schoenfeld David A. Survival analysis of longitudinal microarrays. *Bioinformatics.* 22(21):2643–2649.2006; [PubMed: 17032680]
- Ramoni MF, Sebastiani P, Kohane IS. Cluster analysis of gene expression dynamics. *Proceedings of the National Academy of Sciences.* 99(14):9121–9126.Jul; 2002 0027-8424 DOI: 10.1073/pnas.132656399 DOI: 10.1073/pnas.132656399
- Ring, Brian; Ross, Douglas. Microarrays and molecular markers for tumor classification. *Genome Biology.* 3(5):comment2005.1–comment2005.6.2002; 1465-6906 DOI: 10.1186/gb-2002-3-5-comment2005 [PubMed: 12049658]
- Sha N, Vannucci M, Tadesse MG, Brown PJ, Dragoni I, Davies N, Roberts TC, Contestabile A, Salmon M, Buckley C, Falciani F. Bayesian variable selection in multinomial probit models to identify molecular signatures of disease stage. *Biometrics.* 60(3):812–819.2004; [PubMed: 15339306]
- Storey, John D; Xiao, Wenzhong; Leek, Jeffrey T; Tompkins, Ronald G; Davis, Ronald W. Significance analysis of time course microarray experiments. *Proceedings of the National Academy of Sciences of the United States of America.* 102(36):12837–12842.Sep; 2005 0027-8424 DOI: 10.1073/pnas.0504609102 DOI: 10.1073/pnas.0504609102 [PubMed: 16141318]
- Tanner, Martin A; Wong, Wing H. The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association.* 82(398):528–540.1987; 01621459 DOI: 10.2307/2289457 DOI: 10.2307/2289457
- VanWieringen, Wessel N; Kun, David; Hampel, Regina; Boulesteix, Anne-laure. Survival prediction using gene expression data: a review and comparison. *Multiple Sclerosis.* 53(5):1590–1603.2009;
- Andrew T, Wood A. Simulation of the von mises fisher distribution. *Communications in Statistics Simulation and Computation.* 23(1):157–164.1994;

Figure 1.

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

Figure 2.
Gene expression trajectories of representative genes from different clusters.

Table 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

Table 2

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

	<i>Binary outcome</i>		<i>Survival outcome</i>	
	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)