

Clustering with multiple distance metrics - mixture models with profile transformations.

Rebecka Jörnsten

Department of Statistics and Biostatistics

Rutgers University

110 Frelinghuysen Road

Piscataway, NJ, 08854

April 23, 2007

SUMMARY. Clustering methods often require the selection of a distance metric; how do we define data objects as 'close' enough to be grouped together, or 'far' enough apart to be separated? Choosing an appropriate distance metric is not always easy. We consider high-dimensional gene expression data as an example. The shape of a gene's expression profile across experimental conditions is often considered to be the most informative, which translates to choosing correlation as a similarity metric. However, when genes with a similar expression profile exhibit expression differences on a scale of two-fold to ten-fold, correlation comparisons do not suffice, implying that a Euclidean distance metric is more appropriate. We propose a model-based clustering approach, $\mathcal{MIX}_{\mathcal{T}}$ (**MIX**tured modeling with profile **T**ransformations), which incorporates multiple distance metrics simultaneously. The modeling framework constitutes a between-cluster parameterization, allowing for direct and objective cluster comparisons. With this more efficient parameterization, we detect clusters that a standard model-based clustering approach may miss. We demonstrate the utility of the $\mathcal{MIX}_{\mathcal{T}}$ model via the analysis of a time-course gene expression data set, with two experimental factors, and discuss the biological relevance of the gene clusters identified.

KEY WORDS: Clustering; Gene Expression; Mixture Model; Model Selection; Profile EM

1. Introduction

Clustering is a common approach for dimension reduction, and has been vastly popular for the analysis of high-throughput biological data, such as gene expression microarrays. Algorithmic clustering methods require that a distance or similarity metric is specified (e.g. PAM (Kaufman and Rousseeuw (1990))). It is not always easy to choose one single distance metric of interest (e.g. euclidean, 1-(correlation), or 1-(absolute correlation)), since multiple groupings of data objects may offer insight into the data structure.

In this paper we introduce a model-based clustering method which allows for clustering under several distance metrics simultaneously. Moreover, the modeling structure provides an efficient parameterization of the data, by incorporating both within- and between-cluster subset model selection into model-based clustering.

While the proposed methodology is generally applicable to clustering of high-dimensional data, we will illustrate the method on a time-course multi-factor gene expression study, examining the effects of trauma on spinal cord in rats. The experiment involved exposing anesthetized rats to mild and moderate levels of spinal cord trauma. Spinal cord tissue was then removed at 4h, 24h, 48h, 7 days, and 28 days after injury, and mRNA samples extracted and analyzed with the Affymetrix GeneChip Rat Genome U34 arrays (GEO series GSE2270, De Biase et al. (2005)). The data was normalized against samples taken at 4h after a "Sham" injury, in which the rats were anesthetized and the spinal cord exposed, but no trauma applied. A preliminary clustering of the data set revealed that many groups of genes exhibited a similar expression profile *shape* (over time) after injury, but at different levels of response. In addition, the expression across injury levels also seemed to share a common profile. These observations motivate the need to develop a clustering methodology that formalizes such comparisons; (1) are cluster profiles similar across injury levels, and can the time-course profiles be efficiently represented (within cluster parameterization)?; (2) are cluster profiles similar in terms shape, but representing stronger/weaker responses (between cluster parameterization)? We propose the \mathcal{MLX}_T modeling approach to address these questions.

We begin with a brief review of cluster subset model selection and cluster parameterization.

In regular model-based clustering, data is assumed to be drawn from a mixture model (usually Gaussian) with a cluster specific mean and covariance. That is, data objects \mathbf{x}_g , $g = 1, \dots, G$ are distributed as

$$\mathbf{x}_g \sim \sum_{k=1}^K \pi_k N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k),$$

where $\boldsymbol{\mu}_k = \{\mu_k(j), j = 1, \dots, J\}$ is the J -dimensional cluster mean, and $\boldsymbol{\Sigma}_k$ is the $J \times J$ dimensional cluster covariance. However, in complex data structures, the distinct cluster means may be more efficiently described with an appropriate parameterization. Recently, Raftery and Dean (2006) developed a Bayesian model selection scheme, where it is assumed that $\mu_k(j)$ is equal to a the global mean $\mu(j)$ for a subset of data dimensions $j \in \{1, \dots, J\}$. Hoff (2006) proposed a model where $\mu_k(j) = \mu(j) + \delta_k(j)$, where $\delta_k(j)$ are cluster and dimension specific contrasts, and $\delta_k(j) \neq 0$ for only a subset of clusters $k \in \{1, \dots, K\}$. In Jornsten (2007) and Jornsten and Keles (2006) we proposed a regression parameterization of the cluster mean, $\boldsymbol{\mu}_k = W\boldsymbol{\theta}_k$, where W is a design matrix chosen to reflect the scientific question at hand. This parameterization allows for cluster specific subset models that are more flexible than those proposed to-date. As an example, consider a two-factor experiment with "time" and "treatment". The regression formulation, $\boldsymbol{\mu}_k = W\boldsymbol{\theta}_k$, can identify clusters whose means are distinguished by a time and/or treatment effect, by setting a subset of parameters $\boldsymbol{\theta}_k$ to 0. Such subset models do not necessarily correspond to a subset of data dimensions. The regression formulation thus constitutes an efficient within-cluster parameterization.

In this paper, we focus on efficient *between-cluster* parameterizations. Thus, we allow for parameters to be shared across clusters. The importance of an efficient cluster model parameterization is two-fold; (1) if a sparse representation of the clusters can be provided, this alleviates the reliance on subjective interpretation of the clustering outcome (e.g. that a cluster "appears to" represent a particular shape, or two clusters "seem to" represent similar shapes); (2) if we avoid spending parameters where they are not needed (such as distinguishing between clusters of similar shape) we can instead allocate the model complexity to discover a greater number of distinct clusters.

The paper is organized as follows. In section 2 we review the regression formulation of the cluster mean, and introduce the $\mathcal{MIX}_{\mathcal{T}}$ model that incorporates multiple distance metrics. We outline an

EM algorithm to fit the $\mathcal{MIX}_{\mathcal{T}}$ model to data. We also present an extension to multi-factor experiments. In section 3 we apply the mixture model with profile transformations to a gene expression data set. We study the $\mathcal{MIX}_{\mathcal{T}}$ model in a controlled setting via simulation studies in section 4, and conclude with a discussion in section 5.

2. $\mathcal{MIX}_{\mathcal{T}}$ - model based clustering with profile transformations.

Model-based clustering is a well-structured approach to summarizing high-dimensional data. In contrast, many non-parametric approaches to clustering put implicit constraints on cluster representatives and scales. As an example, k-means clustering assumes an equal, spherical cluster scale, although this is not explicitly stated in the clustering algorithm cost function. In this sense, model-based clustering is more "honest", since the cluster representative (mean) and scale (covariance) are explicitly defined. However, in practise it is often desirable to use a flexible notion of distance. Model-based clustering (with a Gaussian component distribution) uses the Mahalanobis distance, while often a correlation based metric may be more intuitive from a biological standpoint. It would be beneficial if we could stay within the model-based clustering framework while incorporating different distance metrics. In that setting, we would still have control over representative and scale descriptions (mean and covariance).

We propose $\mathcal{MIX}_{\mathcal{T}}$ to address this issue. The basic parameters of the mixture model is a set of cluster profile *shapes*, $\boldsymbol{\mu}_k$, and covariances, $\boldsymbol{\Sigma}_k$, $k = 1, \dots, K$. These parameters are then transformed into a set of *sub-cluster* parameters, $(\boldsymbol{\mu}_{kl}, \boldsymbol{\Sigma}_{kl})$, $l = 1, \dots, L_k, k = 1, \dots, K$. To stay within the Gaussian mixture model, transform functions are restricted to affine transformations, such that $\boldsymbol{\mu}_{kl} = \mathbf{A}_{kl}(\boldsymbol{\mu}_k + \mathbf{b}_{kl})$, $\boldsymbol{\Sigma}_{kl} = \mathbf{A}_{kl}\boldsymbol{\Sigma}_k\mathbf{A}'_{kl}$. The number of free parameters in $(\mathbf{A}_{kl}, \mathbf{b}_{kl})$ must be limited, such that the model is not over-parameterized.

Let us first introduce the basic notation. We assume that G data objects, $\mathbf{x}_g, g = 1, \dots, G$, are generated from a J -dimensional Gaussian mixture. Thus,

$$\mathbf{x}_g \sim \sum_{k=1}^K \sum_{l=1}^{L_k} \pi_{kl} N(\mathbf{A}_{kl}(\boldsymbol{\mu}_k + \mathbf{b}_{kl}), \mathbf{A}_{kl}\boldsymbol{\Sigma}_k\mathbf{A}'_{kl}).$$

Given cluster membership indicators R_g and U_g ,

$$\mathbf{x}_g | R_g = k, U_g = l \sim N(\mathbf{A}_{kl}(\boldsymbol{\mu}_k + \mathbf{b}_{kl}), \mathbf{A}_{kl} \boldsymbol{\Sigma}_k \mathbf{A}'_{kl}),$$

where R_g is a cluster membership indicator that classifies a data object (gene) g as having a particular profile shape $\boldsymbol{\mu}_k$. The second indicator U_g identifies the sub-cluster membership.

In the following section, we discuss some simple transform functions that are practically relevant, and intuitive for modeling high-dimensional biological data.

2.1 *Sign-flip and Scale transformation*

Consider the case where genes (data objects) g are frequently following a similar expression profile, with the exception of a sign difference. From a biological standpoint, such genes may constitute activators or repressors of a biological process, respectively. Similarly, regulatory microRNAs (miRNAs) are believed to act as switches in biological processes, degrading or preventing translation of certain mRNAs (effectively acting as repressors). Thus, miRNA and their targets could be identified by expression profiles that are in all sense similar with the exception of a sign difference (Goff et al. (2007)), i.e. via a mixture model that incorporates sign-flip transformations.

For each profile cluster k , there are two possibilities; a single sub-cluster ($L_k = 1$) with a positive sign, or two sub-clusters ($L_k = 2$), one of each sign. The sign-flip transformation function has no free parameters; $\mathbf{A}_{k1} = I$ (the identity matrix) for the positive sign sub-cluster, and $\mathbf{A}_{k2} = -I$ for the negative sign-flip sub-cluster, and $\mathbf{b}_{kl} = \mathbf{0}, \forall k, l$. The cost of generating a sign-flip cluster for cluster k (i.e. setting $L_k = 2$) is only one parameter; the mean profile $\boldsymbol{\mu}_k$ and the covariance $\boldsymbol{\Sigma}_k$ are shared for sign-flip clusters, so only the cluster proportion π_{kl} needs to be specified.

Another transformation function of interest is the scale transformation, where $\mathbf{A}_{kl} = \beta_{kl}I$, and $\mathbf{b}_{kl} = \alpha_{kl}1_J$ (1_J is a $J \times 1$ vector of ones). We restrict $\mu_k(j = 1) = 0, \forall k$, and $\beta_{k1} = 1, \forall k$, such that the model is not over-parameterized for the special case when $L_k = 1$. Marginally, we thus assume

$$\mathbf{x}_g \sim \sum_{k=1}^K \sum_{l=1}^{L_k} \pi_{kl} N(\beta_{kl}(\boldsymbol{\mu}_k + \alpha_{kl}1_J), \beta_{kl}^2 \boldsymbol{\Sigma}_k).$$

With this model formulation, an increased level of expression changes (large β_{kl}) is also associated with an increased heterogeneity of expression ($\boldsymbol{\Sigma}_{kl} = \beta_{kl}^2 \boldsymbol{\Sigma}_k$). The scale transformation can be used

to identify genes which exhibit a similar expression profile shape, but differs in terms of the level of the response (β_{kl}), or baseline response level (α_{kl}). Take our spinal cord injury data as an example. We wish to identify genes that respond to spinal cord injury, but we also want to examine whether the level of response is associated with different forms of repair mechanisms. Identifying the genetic pathways that are activated after injury is a preliminary step toward developing better therapies to reduce the risk of permanent loss of function due to injury (e.g. Pan et al. (2004)).

Note, the scale transformation (which incorporates the sign-flip transformation as a special case) corresponds to clustering with three different distance metrics simultaneously. We can choose to draw inferences using the sub-cluster memberships, (R_g, U_g) , which translates to using the Mahalanobis distance as a distance metric. If we aggregate cluster memberships, we can focus on the profile cluster shapes (R_g membership indicators only). Sub-clusters l such that $\beta_{kl} > 0$ all have a positive association with cluster shape $\boldsymbol{\mu}_k$. This translates to clustering with a distance metric similar to $1 - \text{correlation}$. If we ignore the sign of β_{kl} , sub-clusters $l = 1, \dots, L_k$ have a positive or negative association with cluster shape $\boldsymbol{\mu}_k$, which translates to using a distance metric similar to $1 - |\text{correlation}|$.

2.2 Fitting the $\mathcal{MIX}_{\mathcal{T}}$ model to data.

We derive an EM algorithm to fit the scale (and/or sign-flip) transformation mixture model. We can write the complete data log likelihood as

$$\begin{aligned} \log P(X, R, U) = & \sum_{g=1}^G \sum_{k=1}^K \sum_{l=1}^{L_k} I\{R_g = k, U_g = l\} \log \Phi(\mathbf{x}_g | \mathbf{A}_{kl}(\boldsymbol{\mu}_k + \mathbf{b}_{kl}), \mathbf{A}_{kl} \boldsymbol{\Sigma}_k \mathbf{A}'_{kl}) + \\ & - I\{R_g = k, U_g = l\} \log P(R_g = k, U_g = l), \end{aligned}$$

where $P(R_g = k, U_g = l) = \pi_{kl}$.

The E-step is simply applied at the sub-cluster level, since profile shape cluster memberships, $\eta_{gk} = E[I\{R_g = k\}]$, can be inferred by aggregating across sub-clusters. Thus,

$$\eta_{gkl} = \frac{\Phi(x_g; \mathbf{A}_{kl}(\boldsymbol{\mu}_k + \mathbf{b}_{kl}), \mathbf{A}_{kl} \boldsymbol{\Sigma}_k \mathbf{A}'_{kl}) * \pi_{kl}}{\sum_{k', l'} \Phi(x_g; \mathbf{A}_{k'l'}(\boldsymbol{\mu}_{k'} + \mathbf{b}_{k'l'}), \mathbf{A}_{k'l'} \boldsymbol{\Sigma}_{k'} \mathbf{A}'_{k'l'}) * \pi_{k'l'}}.$$

The M-step involves a constrained maximization problem. We can write the conditional likeli-

hood as

$$Q(\Theta) = \sum_{g,k,l} -\frac{1}{2} \eta_{gkl} [(\mathbf{x}_g - \mathbf{A}_{kl}(\boldsymbol{\mu}_k + \mathbf{b}_{kl}))' (\mathbf{A}_{kl} \boldsymbol{\Sigma}_k \mathbf{A}_{kl}')^{-1} (\mathbf{x}_g - \mathbf{A}_{kl}(\boldsymbol{\mu}_k + \mathbf{b}_{kl})) + \log |(\mathbf{A}_{kl} \boldsymbol{\Sigma}_k \mathbf{A}_{kl}')^{-1}|],$$

where $\Theta = \{\pi_{kl}, \mathbf{A}_{kl}, \mathbf{b}_{kl}, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, l = 1, \dots, L_k \text{ and } k = 1, \dots, K\}$ is the full set of parameters. The cluster profile shapes, $\boldsymbol{\mu}_k$, are restricted via a within-cluster parameterization: $\boldsymbol{\mu}_k = W\boldsymbol{\theta}_k$, where only a subset of cluster shape parameters $\boldsymbol{\theta}_k$ are non-zero.

This M-step optimization problem is greatly simplified if we partition the task into a sub-cluster estimation problem, and an "aggregation", or profile cluster shape, estimation problem.

Given $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$, we take a profile likelihood approach to estimating the sub-cluster parameters $(\mathbf{A}_{kl}, \mathbf{b}_{kl})$. In the sign-flip transformation case, there are no free parameters. The scale-transformation model involves two free parameters, $(\alpha_{kl}, \beta_{kl})$. We thus consider the following problem:

$$\max_{\alpha_{kl}, \beta_{kl}} \sum_g \eta_{gkl} \left[-\frac{1}{2} \frac{(\mathbf{x}_g - \beta_{kl} \boldsymbol{\mu}_k - \alpha'_{kl} \mathbf{1}_J)' \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_g - \beta_{kl} \boldsymbol{\mu}_k - \alpha'_{kl} \mathbf{1}_J)}{\beta_{kl}^2} - \frac{J}{2} \log(\beta_{kl}^2) + c \right],$$

where $\alpha'_{kl} = \beta_{kl} \alpha_{kl}$, c is a constant. Thus, taking a derivative with respect to α'_{kl} , we get

$$\sum_g \eta_{gkl} \frac{\mathbf{1}'_J \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_g - \beta_{kl} \boldsymbol{\mu}_k - \alpha'_{kl} \mathbf{1}_J)}{\beta_{kl}^2} = 0.$$

Solving for α'_{kl} ,

$$\alpha'_{kl} = \frac{\sum_g \eta_{gkl} \mathbf{1}'_J \boldsymbol{\Sigma}_k^{-1} \mathbf{x}_g}{\sum_g \eta_{gkl} \mathbf{1}'_J \boldsymbol{\Sigma}_k^{-1} \mathbf{1}_J} - \beta_{kl} \frac{\mathbf{1}'_J \boldsymbol{\Sigma}_k \boldsymbol{\mu}_k}{\mathbf{1}'_J \boldsymbol{\Sigma}_k^{-1} \mathbf{1}_J}. \quad (1)$$

Taking a derivative with respect to β_{kl} yields the following quadratic equation:

$$\beta_{kl}^2 - A\beta_{kl} - B = 0, \quad (2)$$

where

$$A = \frac{1}{J \sum_g \eta_{gkl}} \left[\frac{\mathbf{1}'_J \boldsymbol{\Sigma}_k^{-1} \boldsymbol{\mu}_k \sum_g \eta_{gkl} \mathbf{1}'_J \boldsymbol{\Sigma}_k^{-1} \mathbf{x}_g}{\mathbf{1}'_J \boldsymbol{\Sigma}_k^{-1} \mathbf{1}_J} - \sum_g \eta_{gkl} \mathbf{x}'_g \boldsymbol{\Sigma}_k^{-1} \boldsymbol{\mu}_k \right],$$

and

$$B = \frac{1}{J \sum_g \eta_{gkl}} \left[\sum_g \eta_{gkl} \mathbf{x}'_g \boldsymbol{\Sigma}_k^{-1} \mathbf{x}_g - \frac{(\sum_g \eta_{gkl} \mathbf{1}'_J \boldsymbol{\Sigma}_k^{-1} \mathbf{x}_g)^2}{\sum_g \eta_{gkl} \mathbf{1}'_J \boldsymbol{\Sigma}_k^{-1} \mathbf{1}_J} \right].$$

We solve for

$$\beta_{kl} = (A/2) \pm \sqrt{A^2/4 + B}. \quad (3)$$

To incorporate both sign-flip and scale into the $\mathcal{MIX}_{\mathcal{T}}$ model, we evaluate the conditional likelihood at the two solutions, and choose the β_{kl} (positive or negative) that maximizes it. Given appropriate starting values, β_{kl} tends to take on the same sign from iteration to iteration in the EM algorithm. If we wish to restrict the model to scale transformations only, we use the positive solution to equation (3), where the starting values of the transform parameters have been restricted to $\beta_{kl} > 0$ (see section 2.4.1.). Finally, we obtain $\alpha_{kl} = \alpha'_{kl}/\beta_{kl}$.

To avoid numerical instabilities when $\sum_g \eta_{gkl}$ is small, we regularize the β estimator as follows:

$$\beta_{kl}^{reg} = \frac{\nu + n_{kl}\beta_{kl}}{\nu + n_{kl}}, \quad (4)$$

where $n_{kl} = \sum_g \eta_{gkl}$ and ν is a regularization factor. This effectively shrinks individual estimates β_{kl} toward the baseline $\beta = 1$. We find that $\nu \simeq 5$ works well in practice. For the $\mathcal{MIX}_{\mathcal{T}}$ model with sign-flips and scale transformation, the regularization can in fact change the sign of the β_{kl} estimate. Thus, if a spurious sign-flip cluster was generated at the initialization step (section 2.4.1.), the EM procedure may eliminate this cluster if there is little support for it in the data. One could also consider a regularization scheme for α_{kl} (shrinking toward the baseline 0, or the global mean), but this was not found to be necessary in practice.

To solve for $\boldsymbol{\theta}_k$, we condition on the sub-cluster parameters, $(\mathbf{A}_{kl}, \mathbf{b}_{kl})$, and on the previous estimate of the cluster covariance, $\boldsymbol{\Sigma}_k$. To constrain a set of $\theta_k(j) = 0$, we reduce the design matrix W to a cluster specific W_k , eliminating columns j of the matrix corresponding to the parameter constraints. Given $\boldsymbol{\Sigma}_k$, we have thus reduced the estimation problem of $\boldsymbol{\theta}_k$ to the following;

$$\forall k : \frac{dQ(\boldsymbol{\theta})}{d\boldsymbol{\theta}_k} = \sum_{g,l} \eta_{gkl} W'_k \boldsymbol{\Sigma}_k^{-1} \mathbf{A}_{kl}^{-1} (\mathbf{x}_g - \mathbf{A}_{kl}(W_k \boldsymbol{\theta}_k + \mathbf{b}_{kl})) = \mathbf{0},$$

where we assume \mathbf{A}_{kl} is restricted to be invertible. We can solve for $\boldsymbol{\theta}_k$ as

$$\boldsymbol{\theta}_k = \sum_{g,l} \eta_{gkl} (W'_k \boldsymbol{\Sigma}_k^{-1} W_k)^{-1} W'_k \boldsymbol{\Sigma}_k^{-1} [\mathbf{A}_{kl}^{-1} (\mathbf{x}_g - \mathbf{b}_{kl})] / \sum_{g,l} \eta_{gkl}. \quad (5)$$

The shape parameter estimates, $\boldsymbol{\theta}_k$, are thus obtained via weighted Generalized Least Squares, where each data object contributes to the profile shape parameters in multiple ways through the inverse transformation $\mathbf{A}_{kl}^{-1}(\mathbf{x}_g - \mathbf{b}_{kl})$, aggregated across sub-clusters and moderated by the posterior

cluster membership probabilities, η_{gkl} . For example, with the sign-flip transformation, each data object contributes to $\boldsymbol{\theta}_k$ both via a positive and negative association. Similarly, the estimate of $\boldsymbol{\Sigma}_k$, given $\boldsymbol{\mu}_k = W_k \boldsymbol{\theta}_k$, is obtained as

$$\tilde{\boldsymbol{\Sigma}}_k = \sum_{g,l} \eta_{gkl} ([\mathbf{A}_{kl}^{-1}(\mathbf{x}_g - \mathbf{b}_{kl})] - \boldsymbol{\mu}_k) ([\mathbf{A}_{kl}^{-1}(\mathbf{x}_g - \mathbf{b}_{kl})] - \boldsymbol{\mu}_k)' / \sum_{g,l} \eta_{gkl}. \quad (6)$$

Since this estimator can lead to degenerative distribution estimates, we regularize the cluster covariance estimate (Raftery (2004)). That is, we use

$$\boldsymbol{\Sigma}_k = \frac{\boldsymbol{\Delta} + n_k \tilde{\boldsymbol{\Sigma}}_k}{\nu + n_k}, \quad (7)$$

where ν is the smallest value such that $\boldsymbol{\Sigma}_k$ is non-singular, $n_k = \sum_{k,s} \eta_{gks}$, and $\boldsymbol{\Delta} = Cov(\mathbf{x})^{2/M}$, where $M = \sum_k L_k$ is the total number of clusters. In the following sections, we assume that this regularization is always applied to the covariance estimates. The cluster proportions are easily obtained as $\pi_{kl} = \sum_g \eta_{gkl} / G$.

2.3 Multi-factor experiments

Multi-factor experiments present a challenge for model-based clustering. We would clearly want to take the experimental design explicitly into account, and this requires some careful consideration of the cluster model parameterization. Take our three-factor experiment as an example; we are observing gene expression developments across time for two levels of spinal cord injury (mild and moderate). We may want to determine the common expression profile patterns across levels of injury, but we are also interested in the injury-level specific regulation of expression. We extend the $\mathcal{MIX}_{\mathcal{T}}$ model to this setting.

Let us denote by \mathbf{x}_g and \mathbf{y}_g the time-course expression of mild (\mathbf{x}) and moderate (\mathbf{y}) injury levels. We assume that

$$(\mathbf{x}_g, \mathbf{y}_g) | R_g = k, U_g = l \sim N((\beta_{kl}(\boldsymbol{\mu}_k^X + \alpha_{kl} \mathbf{1}_J), \delta_{kl}(\boldsymbol{\mu}_k^Y + \gamma_{kl} \mathbf{1}_J)), \boldsymbol{\Sigma}_{kl}),$$

where

$$\boldsymbol{\Sigma}_{kl} = \begin{bmatrix} \beta_{kl}^2 \boldsymbol{\Sigma}_k^X & \beta_{kl} \delta_{kl} \boldsymbol{\Sigma}_k^{XY} \\ \beta_{kl} \delta_{kl} \boldsymbol{\Sigma}_k^{YX} & \delta_{kl}^2 \boldsymbol{\Sigma}_k^Y \end{bmatrix}.$$

Here, $\mu_k^X(j=1) = 0, \forall k$, and $\mu_k^Y(j=1) = 0, \forall k$. We also fix $\beta_{k1} = \delta_{k1} = 1, \forall k$.

We derive an EM algorithm to fit this model, partitioning the M-step into a sub-cluster estimation problem and an aggregation estimation problem. The joint maximization with respect to scale transformation parameters $(\alpha_{kl}, \beta_{kl}, \gamma_{kl}, \delta_{kl})$, given the profile shape parameters $(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, is complex. We approximate this optimization by de-coupling the problem in terms of the components (\mathbf{x}, \mathbf{y}) , essentially assuming that $\boldsymbol{\Sigma}_k$ is block-diagonal. We thus solve for scale transformation parameters $(\alpha_{kl}, \beta_{kl})$ and $(\gamma_{kl}, \delta_{kl})$ separately (see equations (1)-(4)). In practice, the correlation within each level of the factor tends to exceed the correlation between levels of the factor, and the approximation is thus intuitively motivated. Furthermore, this allows for simple extensions to experiments with more than two factor levels, since the transformation parameter estimation is undertaken separately for each factor level.

Given the scale transformation parameters $(\alpha_{kl}, \beta_{kl}, \gamma_{kl}, \delta_{kl})$, we apply the inverse transformations to the components \mathbf{x}_g and \mathbf{y}_g , respectively. We define $\tilde{\mathbf{x}}_g = \mathbf{x}_g / \beta_{kl} - \alpha_{kl}$, and $\tilde{\mathbf{y}}_g = \mathbf{y}_g / \delta_{kl} - \gamma_{kl}$. Then,

$$(\tilde{\mathbf{x}}_g, \tilde{\mathbf{y}}_g) | R_g = k, U_g = l \sim N((\boldsymbol{\mu}_k^{IX}, \boldsymbol{\mu}_k^{IY})', \boldsymbol{\Sigma}_k),$$

where

$$\boldsymbol{\Sigma}_k = \begin{bmatrix} \boldsymbol{\Sigma}_k^X & \boldsymbol{\Sigma}_k^{XY} \\ \boldsymbol{\Sigma}_k^{YX} & \boldsymbol{\Sigma}_k^Y \end{bmatrix}.$$

We can now estimate mean parameters $\boldsymbol{\theta}_k$, where $\boldsymbol{\mu}_k = (\boldsymbol{\mu}_k^X, \boldsymbol{\mu}_k^Y) = W_k \boldsymbol{\theta}_k$, via weighted GLS as before, and similarly for the covariance matrix $\boldsymbol{\Sigma}_k$ (see equations (5) through (7)). In the multi-factor model, the parameterization W should include the factor level comparisons. We can thus consider parameterizations where $\boldsymbol{\mu}_k^Y = \boldsymbol{\mu}_k^X + \Delta_k = W_k^X \boldsymbol{\theta}_k^X + \Delta_k$, where Δ_k is a vector of cluster contrast parameters between the two levels of injury. This parameterization provides a sparse representation of $\boldsymbol{\mu}_k$ if the cluster profiles $\boldsymbol{\mu}_k^X \simeq \boldsymbol{\mu}_k^Y$. If the cluster profiles are not similar, it is generally more efficient to model each factor level component separately, i.e. let $\boldsymbol{\mu}_k^X = W_k^X \boldsymbol{\theta}_k^X, \boldsymbol{\mu}_k^Y = W_k^Y \boldsymbol{\theta}_k^Y$.

2.4 The $\mathcal{MIX}_{\mathcal{T}}$ algorithm

To fit the $\mathcal{MIX}_{\mathcal{T}}$ model to data, the E- and M-steps previously described form the body of the algorithm. However, there are several important computational details that must be addressed;

(i) the choice of initial starting values; (ii) the selection of the number of profile shape clusters K , and sub-clusters $L_k, k = 1, \dots, K$; and (iii) the within-cluster parameterization $\boldsymbol{\mu}_k = W_k \boldsymbol{\theta}_k$. In the following sections we discuss these algorithmic details.

2.4.1 Initial values.

The algorithm requires initial values of π_{kl} , $\boldsymbol{\mu}_{kl}$ and $\boldsymbol{\Sigma}_{kl}$, for any constellation of profile shape clusters $k = 1, \dots, K$, and sub-clusters $L_k, k = 1, \dots, K$. The search space over possible sub-clusters is very large, and in practice we find that a simple pruning strategy works well. Thus, we parameterize the model via two complexity parameters (M, K) , where $M = \sum_k L_k$ is the total number of clusters fit to the data. To initialize the algorithm for a given (M, K) pair we proceed as follows;

(I) We cluster the data into M clusters via kmeans.

(II) For each transformation model, we reduce the M clusters to K profile shape clusters using one of the strategies (a)-(d).

(a) **Sign-flip:** We compute the distances between the M centroids $c_m, m = 1, \dots, M$ and their sign-flips $-c_m, m = 1, \dots, M$. Given K , we need to construct $\lfloor M/K \rfloor$ pairs of sign-flip clusters. We construct sign-flip pairs in a sequential fashion, starting from (m, m') corresponding to the smallest c_m to $-c_{m'}$ distance.

$\boldsymbol{\mu}_k$ are simply obtained as the cluster mean of the positive and negative sub-clusters. $\boldsymbol{\Sigma}_k$ is obtained as the (regularized) cluster covariance.

(b) **Scale transformation:** We join M clusters into K profile clusters using the Partition-around-medoids (PAM) algorithm, with $1 - \rho(C)$ as the distance metric, where ρ is the correlation, and C refers to the set of centroids $c_m, m = 1, \dots, M$.

For each K profile cluster, the medoid c_k is used to initialize transform parameters $(\alpha_{kl}, \beta_{kl})$ as follows; We standardize the medoid c_k to have standard deviation 1, and first dimension value $c_k(1) = 0$. Since $\boldsymbol{\mu}_{kl} = \alpha'_{kl} + \beta_{kl} \boldsymbol{\mu}_k$, we find $(\alpha'_{kl}/\tau, \tau \beta_{kl})$ via a

regression of c_m on the standardized c_k , for centroids m in top-level cluster k (allocated via PAM). τ is a proportionality constant that is removed by enforcing $\beta_{k1} = 1$, and scaling the remaining $(\alpha_{kl}, \beta_{kl})$ accordingly.

$\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ are initiated from the weighted average of the mean and covariance of the m cluster components within each profile cluster k (after the inverse scale-transform has been applied to each cluster component m).

- (c) **Sign-flip and Scale:** We proceed as in (b) but use PAM with $1 - |\rho(C)|$ as the distance metric.
- (d) **2-factor model:** We join the M kmeans centroids into K top-level clusters using PAM with distance $1 - (\rho(C_1) + \rho(C_2))/2$, where C_1 and C_2 refer to the centroids within the two levels of the factor. Scale transform parameters $(\alpha_{kl}, \beta_{kl}, \gamma_{kl}, \delta_{kl})$ are estimated separately for the two components using the methods in (b).

(II) Each M -to- K profile cluster formation corresponds to a sub-cluster constellation $L_k, k = 1, \dots, K$, where L_k is the number of centroid components c_m allocated to profile cluster k .

(III) Sub-cluster probabilities π_{kl} are obtained from the proportions of each of the M kmeans cluster components.

2.4.2 Model selection

Selecting the number of clusters. As discussed above, the model-space $(K, \{L_k, k = 1, \dots, K\})$ is very large. We simplified the initialization by letting the data propose the sub-cluster structure $L_k, \forall k$, only pre-specifying the total number of clusters, $M = \sum_k L_k$.

To select the number of clusters and sub-clusters we perform a forward search over $M = \sum_k L_k$, and for each M we perform a backward search over $K = M, \dots, 1$. We evaluate each model using the BIC criterion.

For the scale-transformation model, the mixture log-likelihood is denoted

$$l(\Theta(K, \mathbf{L}_K)) = \sum_{g=1}^G \log \left(\sum_{k=1}^K \sum_{l=1}^{L_k} \pi_{kl} \Phi(\mathbf{x}_g; \boldsymbol{\mu}_{kl}, \boldsymbol{\Sigma}_{kl}) \right),$$

where

$$\boldsymbol{\mu}_{kl} = \beta_{kl}(W_k \boldsymbol{\theta}_k + \alpha_{kl} \mathbf{1}_J), \text{ and } \boldsymbol{\Sigma}_{kl} = \beta_{kl}^2 \boldsymbol{\Sigma}_k.$$

Note, $\beta_{k1} = 1$, and $\mu_k(j = 1) = 0$, for all k . The number of parameters in this model is thus

$$p(\Theta(K, \mathbf{L}_K)) = K * \frac{J(J+1)}{2} + \sum_k \text{col}(W_k) + \sum_k (2 * L_k - 1) + \left(\sum_k L_k \right) - 1,$$

where the first term refers to the number of covariance parameters $\boldsymbol{\Sigma}_k$, the second term refers to the number of non-zero cluster-mean parameters $\boldsymbol{\theta}_k$ (or columns of the design matrix W_k), the third term is the number of estimated scale-transformation parameters $(\alpha_{kl}, \beta_{kl})$, and the last term is the number of cluster probability parameters π_{kl} , restricted to add to 1. We compute the BIC value as

$$BIC(K, \mathbf{L}_K) = -2l(\Theta(K, \mathbf{L}_K)) + p(\Theta(K, \mathbf{L}_K)) \log(G). \quad (8)$$

Similarly, for the sign-flip mixture model, $\boldsymbol{\mu}_{kl} = \beta_{kl} \boldsymbol{\mu}_k$, where $\beta_{k1} = 1, \forall k$, and if $L_k = 2$, $\beta_{k2} = -1$, and $\boldsymbol{\Sigma}_{kl} = \boldsymbol{\Sigma}_k$. Thus,

$$p(\Theta(K, \mathbf{L}_K)) = K * \frac{J(J+1)}{2} + \sum_k \text{col}(W_k) + \left(\sum_k L_k \right) - 1.$$

For the multi-factor model we have

$$l(\Theta(K, \mathbf{L}_K))_{m-fac} = \sum_{g=1}^G \log \left(\sum_{k=1}^K \sum_{l=1}^{L_k} \pi_{kl} \Phi(\mathbf{x}_g, \mathbf{y}_g; (\boldsymbol{\mu}_{kl}^X, \boldsymbol{\mu}_{kl}^Y), \boldsymbol{\Sigma}_{kl}) \right),$$

$$p(\Theta(K, \mathbf{L}_K))_{m-fac} = K * \frac{2J(2J+1)}{2} + \sum_k \text{col}(W_k) + \sum_k 2(2 * L_k - 1) + \left(\sum_k L_k \right) - 1,$$

and so the BIC value for is obtained as

$$BIC(K, \mathbf{L}_K)_{m-fac} = -2l(\Theta(K, \mathbf{L}_K))_{m-fac} + p(\Theta(K, \mathbf{L}_K))_{m-fac} \log(G).$$

Selecting the cluster subset model. Since we have formulated the cluster mean $\boldsymbol{\mu}$ in terms of regression, we can be quite flexible in terms of the parameterization W . For the spinal cord injury time-course data set, a preliminary analysis detected cluster profiles with plateaus. We thus

parameterize the time course as $\mu_k(j) = \mu_k(j-1) + \Delta_k(j)$, i.e. via contrasts between consecutive time points. For the multi-factor data set, we found that the shapes of the cluster profiles for the two factor levels were sufficiently different, that the most sparse representations were obtained using a separate parameterization for each factor level. Thus, we model $\mu_k^X(j) = \mu_k^X(j-1) + \Delta_k^X(j)$, and $\mu_k^Y(j) = \mu_k^Y(j-1) + \Delta_k^Y(j)$.

To select a sparse model for each cluster is a complex problem, since cluster subset model selection and cluster allocation are not separable; if we change the cluster model $\boldsymbol{\mu}_k$, then this will clearly change the posterior probabilities η_{gkl} . In Jornsten and Keles (2006) we used a simple two-step approach. We employ an extension of this approach here. Thus, we run the EM algorithm with a full parameterization for each cluster k : $W_k = W$. Once the EM has converged, we perform a hard allocation to the clusters, forming separate data sets

$$C(k, l) = \{g : \operatorname{argmax}_{(k', l')} \eta_{gk'l'} = (k, l)\}. \quad (9)$$

For each cluster k , we then apply a backward selection strategy, dropping columns of the design matrix W_k one at a time (keeping $\boldsymbol{\Sigma}_k$ and $(\alpha_{kl}, \beta_{kl})$ fixed). Each subset model is validated via the "local" BIC (using only cluster k data $\{g \in C(k, l), l = 1, \dots, L_k\}$):

$$BIC_k(W_k) = \sum_{l=1}^{L_k} \sum_{g \in C(k, l)} -2 * \log \Phi(\mathbf{x}_g; \beta_{kl}(W_k \boldsymbol{\theta}_k + \alpha_{kl} \mathbf{1}_J), \beta_{kl}^2 \boldsymbol{\Sigma}_k) + \operatorname{col}(W_k) \log(n_k), \quad (10)$$

where $n_k = \sum_l |C(k, l)|$ is the number of observations allocated to profile shape k and its sub-clusters. We select the model W_k which minimizes $BIC_k(W_k)$. Finally, after a subset model has been identified for all clusters k , we update all model parameters by re-running the EM algorithm under the parameter constraints, represented by W_k .

We provide a detailed outline of the $\mathcal{MLX}_{\mathcal{T}}$ algorithm in the appendix.

3. Application to Data

We analyzed a time-course gene expression data set, where gene expression levels were observed at 4, 24, and 48 hours, and 7 and 28 days after spinal cord injury (De Biase et al. (2005)).

The animal model was rats. RNA samples were surgically extracted from the site of injury.

Response to injury was investigated for two levels of injury (mild and moderate). The 48 hour time point was not included in the mild injury study.

We used data from the affymetrix RG-U34A chip (with full length and annotated genes). The data set was first filtered down to a subset of differentially expressed genes between time points and/or injury levels. This reduced the set of 8799 genes on the chip to 1275 genes. The filtering mechanism was a linear model (F-test), and p-values were adjusted using the Benjamini-Hochberg correction. We used a conservative false discovery rate cutoff at 0.01 percent to focus our study on genes with well-defined expression profiles. Replicate observations were averaged to create a 1275 by 5 dimensional data set for the moderate injury level, and a 1275 by 4 dimensional data set for the mild injury level. The data sets were normalized against the expression level for sham injury (surgery only) at 4 hours.

3.1 *Sign-flip and Scale*

[Figure 1 about here.]

We fit a standard mixture model to the moderate injury level data. We then used the $\mathcal{MIX}_{\mathcal{T}}$ algorithm to fit a sign-flip model, a scale model, or a model incorporating both transforms. All clustering methods were run from 10 different starting values. The results are shown in Figure 1. The BIC curve for the standard mixture model (denoted K in the figure) attains a minimum BIC value for 5 clusters, and the BIC values increase rapidly beyond 5 clusters. The BIC curves for the $\mathcal{MIX}_{\mathcal{T}}$ models, parameterized via scale transforms and/or sign-flips (denoted B and F in the figure), demonstrate the advantages of an efficient cluster model parameterization. For illustration purposes, we include two sub-optimal transformation models in the figure, indicated by the BIC values at 3 and 4 clusters exceeding the BIC curve for the standard mixture model. The $\mathcal{MIX}_{\mathcal{T}}$ model with $M = 3$ and $K = 2$ clusters forces one of the 3 clusters to be either a sign-flip or scale transformation of one of the other clusters. However, the set of $M = 3$ clusters is not rich enough to support this constraint, i.e. all 3 unique shapes are needed to describe the data. Thus, $\mathcal{MIX}_{\mathcal{T}}$ actually picks $M = K$, when M is small. However, once M is larger than 4, $\mathcal{MIX}_{\mathcal{T}}$ provides much

more efficient parameterizations of the data by sharing parameters across clusters, as indicated by the low BIC values over a range from $M = 5$ to 10 clusters. $\mathcal{MIX}_{\mathcal{T}}$ selects $M = 9$ clusters, with $K = 2$ profile shape clusters, for this data set.

[Figure 2 about here.]

In Figure 2 (a) we show the $M = 5$ cluster mean profiles obtained with the standard mixture model, and in Figure 2 (b) we depict the $M = 9$ cluster profiles obtained with $\mathcal{MIX}_{\mathcal{T}}$. (Note, $\mathcal{MIX}_{\mathcal{T}}$ with $M = 5$ clusters has a lower BIC value than the winning standard mixture model outcome with $M = 5$ clusters (see Web supplementary Figure 2).) The $\mathcal{MIX}_{\mathcal{T}}$ model with $M = 9$ corresponds to $K = 2$ profile shape clusters. The sub-cluster structure is $\{L_k\} = (L_1 = 7, L_2 = 2)$. The first profile shape has 7 sub-clusters, of which clusters 1, 3 and 4 are sign-flips. The second profile shape has 2 sub-clusters (see Figure 2 (b)).

For comparison, we cluster the data using the PAM (partition around medoids) algorithm, with $1 - correlation$ as the distance metric. We select the number of clusters via the silhouette width (Kaufman and Rousseeuw (1990)). Here, the silhouette is monotone decreasing and thus only $M = 2$ clusters are detected. We compare the two PAM medoids to the $\mathcal{MIX}_{\mathcal{T}}$ profile shapes $(\mu_k, k = 1, 2)$. The first cluster component in PAM largely overlaps with clusters 2, 5, 6 and 7 (defined as sub-clusters by $\mathcal{MIX}_{\mathcal{T}}$), but the sign-flip clusters (1, 3 and 4) are allocated with clusters 8 and 9 in the second PAM cluster component. If we apply PAM with $1 - |correlation|$ as the distance metric, the silhouette method again selects only two clusters. These cluster profiles closely resemble the $\mathcal{MIX}_{\mathcal{T}}$ profiles (see Web supplementary Figure 3). Note, however, that with the $\mathcal{MIX}_{\mathcal{T}}$ approach we have the option to analyze the data at the profile shape cluster level as here (two clusters), or use the sub-cluster structure (9 clusters). This is a clear advantage over both standard mixture modeling, and (absolute) correlation based clustering with PAM. Moreover, $\mathcal{MIX}_{\mathcal{T}}$ allows for flexible covariance structures for the profile shape clusters (parameterized by Σ_k), while PAM tends to produce equal scale and size clusters (Jornsten (2004)).

[Table 1 about here.]

Returning now to model-based clustering, in Table 1 (left panel) we summarize the selection results. We used a parameterization $\boldsymbol{\mu}_k = W\boldsymbol{\theta}_k$, where W is a design matrix such that $\mu_k(j) = \mu_k(j-1) + \Delta_k(j)$. A sparse model thus corresponds to a profile with plateaus, or a profile that levels off. In the standard mixture model (denoted "Standard"), the winning model constrains 4 parameters in the set $\boldsymbol{\theta}$ to 0. In Figure 2 (a) the sparsity is illustrated. Cluster 3 has a plateau in time interval 3-4, cluster 4 levels off in time interval 4-5, and clusters 2 and 5 have $\mu_k(1) = 0$.

Table 1 (left panel) also summarizes the sub-cluster structure of the $\mathcal{MIX}_{\mathcal{T}}$ model. Across $M = 3$ to 9 clusters, $\mathcal{MIX}_{\mathcal{T}}$ selects $K = 2$ to 3 profile shape clusters, at considerable reduction in the total number of cluster parameters. The profile shape $\boldsymbol{\mu}_k$ now has to serve multiple sub-clusters, and this often prevents individual shape parameters $\boldsymbol{\theta}_k$ from being set to 0.

The $M = 9$ clusters we identified with $\mathcal{MIX}_{\mathcal{T}}$ can be grouped into 2 main expression profile shapes (cluster 1-7 vs. clusters 8-9), and within the first shape we have a further division into sign-flip (clusters 1,3 and 4 vs. clusters 2, 5, 6 and 7). We briefly discuss the annotation of the genes in each cluster. Over-represented GO categories are summarized in Web supplementary Tables 1 and 2. We used the DAVID (Dennis et al. (2003)) functional annotation tool to extract GO categories, and identify genes in known pathways.

Clusters 2, 5, 6 and 7: These clusters have profile shapes that peak at 7 days after injury. Clusters 7, 2 and 6 all start near 0 expression (corresponding to sham at 4 hours), meaning their $\alpha_{kl} \simeq 0$. Clusters 2 and 6 correspond to "stronger" responses compared with cluster 7 (β_{kl} is larger). Cluster 5 has a higher expression at the onset ($\alpha_{kl} = .5$), and similar strength of response as cluster 7 ($\beta_{kl} \simeq 1$ in both cases).

Genes in cluster 7 include IL1R (Interleukin receptor), TNF α , and c-fos, which are all part of the inflammatory response. This inflammatory response to spinal cord injury leads to swelling and subsequent cell death, and is the leading cause of secondary damage, and permanent loss of function due to spinal cord injury. In this cluster we also find Nur77 (promoting cell death), Gadd45 (associated with DNA repair), and TGF β (a neuroprotective growth factor). Many of these genes have been identified in gene expression studies of injured spinal cord previously (De Biase et al.

(2005), Song et al. (2001), Carmel et al. (2001)). We also find Decorin in cluster 7, which is known to be associated with central nervous system injury (Carmel et al. (2001), Pan et al. (2004)).

Cluster 5 is almost an offset of cluster 7, with a profile at higher expression levels. In cluster 5 we find several genes related to the production of cytokines (FC ϵ RI β , PKC - a marker for spinal cord), as well as an inflammatory response gene IL-3 (Interleukin) which stimulates T-cell growth. (Interleukines are a particular class of cytokines, which in turn are a group of proteins forming the basis of the immune response, and other signaling and cell communication pathways.) We further find ROCK, which is associated with neurite formation (repair).

Cluster 2 represents a stronger response than cluster 7. In this cluster we find two cell-death inhibitors (BCL2L1, CASP1), and genes associated with recovery; TGF β RI and Smad, which lead to neurite sprouting. Cluster 2 is strongly associated with fatty acid metabolism. Injured tissue requires more energy than healthy tissue, for tasks such as increasing or changing fuel sources under reduced blood supply conditions, or by preparing to start cell division, or by producing extracellular materials in response to injury (Prof. R. Hart, Rutgers, private communication).

Cluster 6 is the cluster corresponding to the strongest response. In this cluster we find some genes associated with cell death (Gadd153, MEK5), but also the protective heat-response HSP72 and C5R1 (an inhibitor of cell death). Another immune response factor IL-4 is also present in this clusters. Moderate levels of some Interleukins can be beneficial for injury repair, but high concentrations can trigger the expression of neurotoxic genes (Song et al. (2001)).

Clusters 1, 3 and 4: These clusters represent sign-flip profiles of clusters 2, 5, 6 and 7. Cluster 1 starts off at high-expression and decreases. Cluster 4 is a moderated response of cluster 1, and cluster 3 is lower still.

Cluster 1 contain genes associated with injury and inflammatory response (VWF, BDKR) (an early response wave around 4 hours). We also find MEK and IGF1 in this cluster. The first stimulates growth, and the latter is neuroprotective. GO categories that are over-represented in cluster 1 include T-cell activation (immune response).

Cluster 4 contains a growth hormone (GNRH1R) and a regulator of the actin cytoskeleton

(CALN). We also find several genes promoting cell-death.

In Cluster 3 we find genes associated with ion binding and transfer (transporter activity). It is known (De Biase et al. (2005), Song et al. (2001)) that spinal cord injury is associated with disruptions of the ion channel signal transduction processes. We also find gene AP-1 in cluster 3, which has previously been found to be associated with spinal cord injury (Song et al. (2001)). Several neuron specific genes (e.g. SNCA) are in this cluster, which could be indicative of delayed cell death at the site of injury.

Clusters 8 and 9: Clusters 8 and 9 attain their lowest expression at 48 hours after injury. Cluster 8 starts off at 0 expression (compared with Sham) and returns to almost the same level after 28 days. Cluster 9 starts off under-expressed compared with Sham.

Cluster 8 contains SNAP and the GABA-receptor, which have previously been reported as neuron specific, and their under-expression can thus be interpreted as a marker of neuron death after injury (Carmel et al. (2001), De Biase et al. (2005), Song et al. (2001)). In this cluster we also find several genes in the JAK-STAT pathway (cytokine response), including cytokineR and IL-10. Since IL-10 is anti-inflammatory, under-expression of IL-10 stops its inhibitory role of the production of cytokines (e.g. IL-3).

Cluster 9 contains many genes associated with cell death. For instance, HSP27 which is an inhibitor of cell-death, is in cluster 9 and is thus under-expressed for moderate levels of injury. We also find several genes whose under-expression could be indicative of ion channel disruption (Calmodulin, Caveolin).

3.2 *A multi-factor experiment*

We analyze the two levels of injury (mild and moderate) and their effect on gene expression at 4 and 24 hours, and 7 and 28 days after injury. The standard mixture model selects only 3 clusters for this data set (see Web supplementary Figure 1). (This is not contradicting the analysis of the moderate level data, since the present data set includes one less time point. In addition, the mild expression profiles are not as distinct.) With the $\mathcal{MIX}_{\mathcal{T}}$ model we detect 7 clusters, depicted in

Figure 3. From Figure 3 we can clearly see that the higher level of injury is associated with stronger responses. In addition, it is interesting to observe that the moderate injury profiles tend to peak earlier than the mild injury profiles (e.g. clusters 5,6 and 7).

[Figure 3 about here.]

The $\mathcal{MI}\mathcal{X}_T$ model with $M = 7$ clusters has $K = 4$ profile clusters. Thus, this representation corresponds to a gain, not only in the number of clusters, but also in the number of distinct shapes (4 instead of 3 with the standard mixture model). No sign-flip clusters are detected. While sign-flip clusters may exist for the moderate injury level data, these clusters do not exhibit a sign-flip structure for the mild level data. For example, clusters (3,4) and (5,6) are almost sign-flip representations of each other for the moderate level injury data. For the mild level injury data however, clusters (5,6) are still exhibiting an increase in expression at 28 days after injury, and cannot be mapped to a sign-flip representation of clusters (3,4) (Figure 3). This suggests that up-regulation of gene expression after injury is injury-level specific, while down-regulation is less so. The down-regulated gene clusters (exhibiting similar shapes across injury levels) are associated with neuron markers and channel disruption (indicating neuron death after injury, see section 3.1). Thus, neuron death seems to follow a similar time progression for both levels of injury, but the moderate level of injury is associated with a higher degree of cell death (scale transform parameters $\delta_{kl} > \beta_{kl}$). Up-regulation on the other hand, which is related to immune response and repair mechanisms (section 3.1), is responding faster for moderate injury level data than mild injury level data.

We detect several scale-transformation clusters ((1,2), (3,4), and (5,6)). The sub-cluster separation for the mild injury level data, $|\beta_{k1} - \beta_{k2}|$, is for all clusters smaller than the sub-cluster separation for the moderate injury level data, $|\delta_{k1} - \delta_{k2}|$. This is again indicating that the level of injury has a significant impact on distinct gene groups.

The winning $\mathcal{MI}\mathcal{X}_T$ model sets two of the shape parameters $\theta_k, k = 1, \dots, 4$ to 0. Thus, clusters 5 and 6 (members of the same profile shape cluster $k = 3$) are flat between time points 2 and 3 (24 hours-7 days) for the mild injury level, and cluster 7 (profile shape cluster $k = 4$) levels

off at 7 days after injury (3rd time point).

We are omitting a discussion on the functional annotation of the genes in these clusters. Many of the results are closely related to those reported in the previous section (e.g. cluster 3 contains many neuron specific genes, and their under-expression is an indication of neuron death after injury, similar to cluster 8 in the previous section). A more detailed biological discussion is outside the scope of this article.

4. Simulations studies

To examine the performance of the $\mathcal{MIX}_{\mathcal{T}}$ method, we simulate data from the selected models in section 3. Model 1 is the standard mixture model with 5 clusters, selected for the moderate level data. However, since these cluster profiles can easily be seen to be well-represented with a scale-transformation structure (see Figure 2 (a)), we permuted the mean value across clusters for the 4th time point. Model 2 is the $\mathcal{MIX}_{\mathcal{T}}$ model with $M = 9$ clusters and $K = 2$ top-level clusters, selected for the moderate level data. Model 3 is the $M = 7$, $K = 4$ $\mathcal{MIX}_{\mathcal{T}}$ model, selected for the multi-factor experiment (mild and moderate).

We generate 50 data sets from each of the 3 models. We search across $M = 2$ to $M = 12$ clusters, allowing for any sub-cluster structure in the $\mathcal{MIX}_{\mathcal{T}}$ fitting procedure. We run the algorithms from one starting value only. (This is somewhat unfavorable for the $\mathcal{MIX}_{\mathcal{T}}$ approach, which depends on not only the M cluster initialization, but also the sub-cluster formation.) In Table 2 we summarize the results.

[Table 2 about here.]

Model 1 is a model for which no sub-cluster exists. From the results in Table 2, we see that both the standard and $\mathcal{MIX}_{\mathcal{T}}$ fits almost always identify the correct model. Moreover, the $\mathcal{MIX}_{\mathcal{T}}$ approach makes only one mistake in terms of generating a sub-cluster.

Model 2 is a scale-transformation and sign-flip model, where the true number of clusters is $M = 9$ with $K = 2$ profile shape clusters. The standard approach cannot find this clustering

structure, and selects only 3 clusters for the data. $\mathcal{MIX}_{\mathcal{T}}$ frequently succeeds at identifying the correct number of profile shape clusters (selecting $K = 2$ in 43 out of 50 simulations). Finding the correct total number of clusters is a more difficult task. Since cluster 1 is relatively small (26 genes), it is sometimes missed, and clusters 2 and 7 are sometimes combined. Thus, we frequently select only $M = 7$ or 8 clusters total.

Model 3 is a scale- and sign-flip transformation model with $M = 7$ clusters total, and $K = 4$ profile shape clusters. Again, the standard approach cannot detect this structure. $\mathcal{MIX}_{\mathcal{T}}$ correctly selects $K = 4$ profile shape clusters in 49 out of 50 simulations. Since cluster 1 is relatively small (46 genes), it is occasionally missed, and thus a total number of $M = 6$ clusters is frequently selected.

In the bottom panel of Table 2 we show the BIC values of the selected models, for each of the 3 simulation scenarios. We calculated, for each run, the mean difference (and the standard error) between the BIC value of the model selected by the standard mixture model, and the BIC value of the model selected with the $\mathcal{MIX}_{\mathcal{T}}$ approach. For model 1, both methods tend to select the correct model, and the BIC values are almost always the same. For models 2 and 3, the BIC of the standard approach always substantially exceeds that of $\mathcal{MIX}_{\mathcal{T}}$.

We can thus conclude that the $\mathcal{MIX}_{\mathcal{T}}$ will not detect sub-clusters where none exists, and can detect sub-clusters when they are present. Across all three simulations, $\mathcal{MIX}_{\mathcal{T}}$ produced clustering results close to the true model. When a transformation structure existed, $\mathcal{MIX}_{\mathcal{T}}$ generated models with far lower BIC values than the standard mixture model.

5. Discussion

The $\mathcal{MIX}_{\mathcal{T}}$ model incorporates multiple distance metrics into model-based clustering. This model formulation represents an efficient parameterization of the cluster model, where clusters sharing a similar shape are simply modeled as transformations of this shape, at a very low cost in the number of parameters. The efficient parameterization can allow for the detection of more distinct clusters in the data, and provides a direct and objective comparison between clusters.

We presented an EM algorithm to fit $\mathcal{MIX}_{\mathcal{T}}$ models, incorporating sign-flips or scale transfor-

mations. Furthermore, we proposed an extension to multi-factor studies.

We demonstrated the utility of the method on a gene expression data set, and discussed the biological relevance. We saw that some of the detected clusters represented varying levels of response to spinal cord injury. These clusters were found to consist of genes within distinct functional categories.

The $\mathcal{MIX}_{\mathcal{T}}$ model formulation allows for many possible extensions. In multi-factor experiments it will be useful to allow for a different number of sub-clusters for the various levels of the factor of interest (e.g. mild vs. moderate). In addition, to make cluster interpretation more objective, subset selection of all parameters (mean shape parameters, and the scale transformation parameters) would be needed. We are currently investigating such extensions.

We focused on a set of simple and practically motivated transformations in this paper; sign-flip and scale. These transforms translated to a model-based clustering with three different distance metrics simultaneously: Mahalanobis distance, $1 - correlation$, and $1 - |correlation|$. However, the $\mathcal{MIX}_{\mathcal{T}}$ framework can be extended to affine transformations which allow for differential offset and scale at different data dimensions (diagonal \mathbf{A}_{kl}), as well as transforms that incorporate delay parameters. In the context of spinal cord injury, it is feasible that there are distinct clusters of genes that only differ in their response level for a particular time interval after injury, and that such gene subsets represent differential targets for drug therapies.

Acknowledgement

We thank Professor Ron Hart of the W.C. Keck Center, Rutgers University, and members of the Hart lab for helping us with a preliminary interpretation of the analysis outcome.

RJ is partially supported by NSF grant DMS0306360. RJ is also supported by the USEPA-funded Environmental Bioinformatics and Computational Toxicology Center (ebCTC), under STAR Grant number GAD R 832721-010. This work has not been reviewed by and does not represent the opinions of the funding agencies.

REFERENCES

- Carmel, J., Galante, A., Soteropoulos, P., Tolia, P., Recce, M., Young, W. and Hart, P. (2001). Expression profiling of acute spinal cord injury reveals spreading inflammatory signals and neuron loss. *Physiological Genomics* **7**, 201–213.
- De Biase, A., Knobloch, S., Di Giovanni, S., Molon, A., Hoffman, P. and Faden, A. (2005). Gene expression profiling of experimental traumatic spinal cord injury as a function of distance from impact site and injury severity. *Physiological Genomics* **22**, 368–381.
- Dennis, G., Sherman, B., Hosack, D., Yang, J., Gao, W., Lane, H. and Lempicki, R. (2003). David: Database for annotation, visualization, and integrated discovery. *Genome Biology* **4**(5), P3. <http://david.abcc.ncifcrf.gov/>.
- Goff, L. A., Davila, J., Jörnsten, R., Keles, S. and Hart, R. P. (2007). Bioinformatic analysis of neural stem cell differentiation. *Journal of Biomolecular Techniques* **18**, 205–212.
- Hoff, P. (2006). Model-based subspace clustering. To appear in Bayesian Analysis.
- Jornsten, R. (2004). Clustering and classification based on the l1 data depth. *Journal of Multivariate Analysis* **90**, 67–89.
- Jornsten, R. (2007). Simultaneous model selection via rate-distortion theory, with applications to clustering and significance analysis of gene expression data. *Technical report 07-01, Rutgers University, Department of Statistics. Submitted to Journal of Computational and Graphical Statistics* .
- Jornsten, R. and Keles, S. (2006). Mixture models with multiple levels, with application to the analysis of multi-factor gene expression data. *Technical report 06-02, Rutgers University, Department of Statistics, Under revision for Biostatistics* .
- Kaufman, L. and Rousseeuw, P. J. (1990). *Finding Groups in Data: An introduction to cluster analysis*. Wiley, New York.
- Pan, J., Jornsten, R. and Hart, R. (2004). Screening anti-inflammatory compounds in injured spinal cord with microarrays: a comparison of bioinformatics approaches. *Physiological Genomics* **17**,

201–214.

Raftery, A. and Dean, N. (2006). Variable selection for model-based clustering. To appear in the *Journal of the American Statistical Association*.

Raftery, C. F. A. (2004). Bayesian regularization for normal mixture estimation and model-based clustering. Technical Report 486, University of Washington.

Song, G., Cechvala, C., Resnick, Dempsey, R. and Rao, V. (2001). Genechip analysis after acute spinal cord injury in rat. *Journal of Neurochemistry* **79**, 804–815.

Appendix

The $ML\mathcal{X}_{\mathcal{T}}$ algorithm.

1. Set $M = M + 1$.
2. Set $K = M + 1$.
 - (a) If $K > 1$, set $K = K - 1$. Otherwise GO TO 1.
 - (b) Apply the $ML\mathcal{X}_{\mathcal{T}}$ EM algorithm (see below).
 - (c) Generate cluster k specific data sets via hard allocation (equation (9)). Perform within-cluster model selection using the local BIC (equation (10)). Output: the cluster specific models (design matrices) W_k .
 - (d) Apply the $ML\mathcal{X}_{\mathcal{T}}$ EM iterations (see below), with cluster specific design matrices W_k .
 - (e) Compute $BIC(M, K)$ via equation (8).

If $K = M$, $BIC(M)^* = BIC(M, K)$.

Otherwise, if $M > K$ and $BIC(M, K) < BIC(M)^*$, set $BIC(M)^* = BIC(M, K)$, $K^*(M) = K$, and GO TO 2(a).

Otherwise, if $M > K$ and $BIC(M, K) \geq BIC(M)^*$, GO TO 3.

3. If $M = \min M$, $BIC^* = BIC(M)^*$.

Otherwise, if $M > \min M$ and $BIC(M)^* < BIC^*$, set $BIC^* = BIC(M)^*$, $M^* = M$, $K^* = K^*(M)$, and GO TO 1.

If $M > \min M$ and $BIC(M)^* \geq BIC^*$, STOP

The $ML\mathcal{X}_T$ EM algorithm.

1. **Initialize:** Initialize the model parameters $\Theta(K, \mathbf{L}_K) = \{\pi_{kl}, (\alpha_{kl}, \beta_{kl}), \boldsymbol{\theta}_k, \boldsymbol{\Sigma}_k, l = 1, \dots, L_k, k = 1, \dots, K\}$ (section 2.4.1).

2. **Iterate:**

E-step

(a) Compute η_{gkl} , $g = 1, \dots, G$, $l = 1, \dots, L_k$, $k = 1, \dots, K$.

M-step

(a) Update π_{kl} , $\forall k, l$.

(b) Given transformation parameters $(\alpha_{kl}, \beta_{kl})$

i. Update $\boldsymbol{\Sigma}_k$, $k = 1, \dots, K$ (equations (6),(7)).

ii. Update $\boldsymbol{\theta}_k$, $k = 1, \dots, K$ via weighted GLS (equation (5)).

(c) Given profile parameters $(\boldsymbol{\mu}_k = W_k \boldsymbol{\theta}_k, \boldsymbol{\Sigma}_k)$

i. Update (β_{kl}) via equation (3).

ii. Regularize β_{kl} via equation (4).

iii. Update α_{kl} via equation (1).

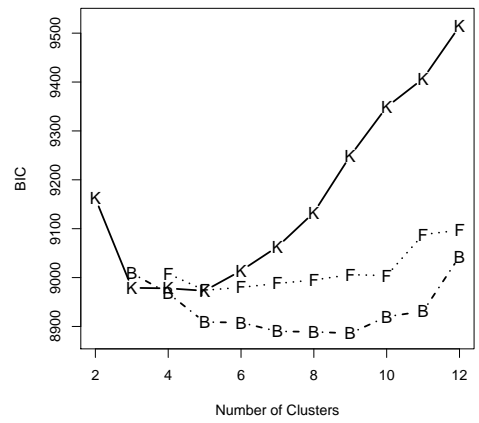


Figure 1. BIC vs. the number clusters for the standard mixture model (K), the sign-flip model (F), and the sign-flip and scale transformation model (B).

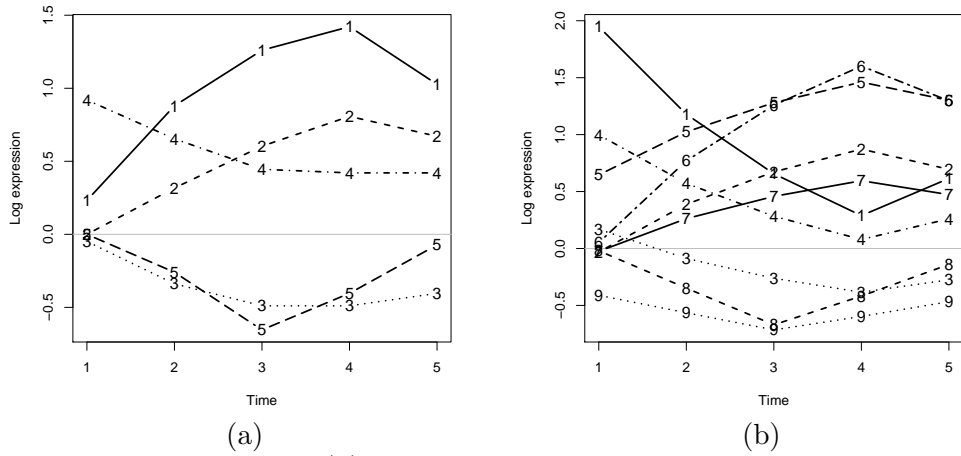


Figure 2. Cluster mean profiles: (a) standard mixture model with $M = 5$ clusters, (b) MIX_T with $M = 9$ clusters.

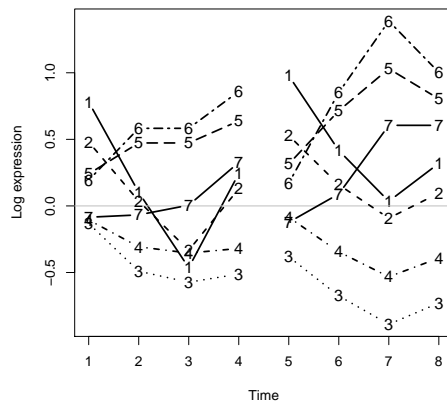


Figure 3. Cluster mean profiles. The left sub-panel of each figure shows the expression for the mild injury data, the right figure sub-panel shows the moderate injury data.

| M | Moderate injury | | | Mild and Moderate injury | | |
|-----|------------------|---------|-------------------------------|--------------------------|-----------|-------------------------------|
| | Standard | | $\mathcal{MIX}_{\mathcal{T}}$ | Standard | | $\mathcal{MIX}_{\mathcal{T}}$ |
| | $\#\theta_k = 0$ | L_k | $\#\theta_k = 0$ | $\#\theta_k = 0$ | L_k | $\#\theta_k = 0$ |
| 3 | 4 | (1,1,1) | 4 | 5 | (2,1) | 0 |
| 4 | 3 | (2,1,1) | 0 | 3 | (2,1,1) | 1 |
| 5 | 4 | (3,2) | 0 | 6 | (3,1,1) | 2 |
| 6 | 2 | (3,1,2) | 1 | 3 | (2,3,1) | 2 |
| 7 | 2 | (2,2,3) | 1 | 7 | (2,2,2,1) | 2 |
| 8 | 1 | (4,3,1) | 0 | 6 | (2,2,3,1) | 1 |
| 9 | 0 | (7,2) | 0 | 6 | (4,4,1) | 0 |

Table 1

Number of coefficients set to 0 by subset selection for the standard mixture model, and $\mathcal{MIX}_{\mathcal{T}}$. Number of sub-clusters selected in the $\mathcal{MIX}_{\mathcal{T}}$ model. Left panel: moderate injury level data. Right panel: Multi-factor study, mild and moderate injury levels.

| | Model 1 | | | Model 2 | | | Model 3 | | |
|------------------|-----------|-------------------------------|---|------------|-------------------------------|---|-----------|-------------------------------|-----------|
| | Std. | $\mathcal{MIX}_{\mathcal{T}}$ | | Std. | $\mathcal{MIX}_{\mathcal{T}}$ | | Std. | $\mathcal{MIX}_{\mathcal{T}}$ | |
| K= | M | 5 ^o | 6 | M | 2* | 3 | M | 3 | 4** |
| M=2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| M=3 | 0 | 0 | 0 | 49 | 0 | 0 | 3 | 0 | 0 |
| M=4 | 0 | 0 | 0 | 1 | 0 | 0 | 46 | 0 | 4 |
| M=5 ^o | 49 | 47 | 0 | 0 | 0 | 0 | 1 | 0 | 5 |
| M=6 | 1 | 1 | 2 | 0 | 8 | 1 | 0 | 1 | 28 |
| M=7** | 0 | 0 | 0 | 0 | 10 | 2 | 0 | 0 | 12 |
| M=8 | 0 | 0 | 0 | 0 | 11 | 2 | 0 | 0 | 0 |
| M=9* | 0 | 0 | 0 | 0 | 9 | 1 | 0 | 0 | 0 |
| M=10 | 0 | 0 | 0 | 0 | 5 | 1 | 0 | 0 | 0 |
| $\Delta(BIC)$ | -0.8(3.4) | | | 117.7(6.4) | | | 93.1(6.8) | | |

Table 2

The number of times each (M, K) cluster structure is selected. Model 1 - the correct model is $M = K = 5$, indicated with a circle (o). Model 2 - the correct model is $M = 9, K = 2$, indicated with *. Model 3 - the correct model is $M = 7, K = 4$, indicated with **. Tabulated $\Delta(BIC) = BIC(\text{Standard}) - BIC(\mathcal{MIX}_{\mathcal{T}})$ values of the winning models (and standard errors).