

Mixture models with multiple levels, with application to the  
analysis of multi-factor gene expression data.

Rebecka Jörnsten\*

Department of Statistics

Rutgers University

501 Hill Center

Piscataway, NJ 08854, USA

Sündüz Keleş

Department of Statistics,

Department of Biostatistics

and Medical Bioinformatics

University of Wisconsin-Madison

1300 University Avenue

Madison, WI 53706, USA

---

\*Corresponding Author: [rebecka@stat.rutgers.edu](mailto:rebecka@stat.rutgers.edu), telephone +1 732 445-3145, fax +1 732 445-3428

## Abstract

Model-based clustering is a popular tool for summarizing high-dimensional data. With the number of high-throughput large-scale gene expression studies still on the rise, the need for effective data summarizing tools has never been greater. By grouping genes according to a common experimental expression profile, we may gain new insight into the biological pathways that steer biological processes of interest. Clustering of gene profiles can also assist in assigning functions to genes that have not yet been functionally annotated.

In this paper we propose two model selection procedures for model-based clustering. Model selection in model-based clustering has to-date focused on the identification of data dimensions that are relevant for clustering. However, in more complex data structures, with multiple experimental factors, such an approach does not provide easily interpreted clustering outcomes.

We propose a mixture model with multiple levels,  $\mathcal{MIX}_{\mathcal{L}}$ , that provides sparse representations both *within* and *between* cluster profiles. We explore various flexible *within-cluster* parameterizations, and discuss how efficient parameterizations can greatly enhance the objective interpretability of the generated clusters. Moreover, we allow for a sparse *between-cluster* representation with a different number of clusters at different levels of an experimental factor of interest. This enhances interpretability of clusters generated in multiple-factor contexts.

Interpretable cluster profiles can assist in detecting biologically relevant groups of genes that may be missed with less efficient parameterizations. We use our multi-level mixture model to mine a proliferating cell-line expression data set for annotational context and regulatory motifs. We also investigate the performance of the multi-level clustering approach on several simulated data sets.

**Keywords:** Clustering, Gene Expression, Mixture Model, Model Selection, Profile Expectation-Maximization

## 1 Introduction

Model-based clustering is frequently used to summarize complex high-dimensional gene expression data. The base model is usually Gaussian, though some robust alternatives have recently been proposed (Banfield and Raftery; 1993). The multivariate Gaussian mixture allows for

clusters of varying shape and volume (Fraley and Raftery; 2002, 2004; Raftery and Dean; 2006). Many non-parametric clustering algorithms have also been proposed for the analyses of genomic data. Non-parametric approaches may seem more flexible than parametric mixture modeling. However, many of the most commonly used non-parametric schemes are in fact very restrictive, in that cluster shapes are implicitly defined by the cost function of the clustering algorithm (Jornsten; 2004). We consider k-means as an example (or any center-based clustering like PAM (Partition Around Medoids) or the k-median (Kaufman and Rousseeuw; 1990; Jornsten et al.; 2002)). By making cluster assignment solely dependent on the cluster center, cluster shape is ignored. Thus, k-means tends to produce spherical and equal size clusters, and is thus more restrictive than a model-based clustering approach where the cluster covariances are parameterized.

In this paper we discuss how to generate interpretable and efficient data representations using multivariate gaussian mixture models. We address the following limitations of model-based clustering; (1) Model-based clustering usually treats all experimental conditions interchangeably, even in the case of multi-factor experiments; (2) Subset model selection for model-based clustering has mainly focused on identifying the *dimensions* that are informative with respect to cluster separation (Law et al.; 2004; Raftery and Dean; 2006; Tadesse et al.; 2005; Hoff; 2006), as opposed to the sparsest representation of each cluster mean. The limitations listed in items (1) and (2) above can result in both an overfit and underfit of the data. Overfitting might be caused by assigning an unnecessary degree of complexity to some clusters, whereas underfitting concerns the number of clusters.

We propose a multi-level mixture modeling approach that generates interpretable clusters in multiple-factor experiments. Throughout the paper, we will focus on an example data set involving proliferating stem cell-lines. Our task is to identify sets of genes that are differentially regulated during neurogenesis and gliogenesis, as indicated by different expression levels in two, divergent neural stem cell (NSC) clones. Upon the withdrawal of a growth factor (FGF) from the medium, one clone (L2.3) becomes predominately glial-like (expressing glial markers GFAP, GalC). The other (L2.2) differentiates primarily into cells expressing neuronal markers (TuJ1) (Goff et al.; 2007). Initially, the cell-lines are virtually indistinguishable, and are believed to exist in a state of "pre-conditioning" or "pre-programming". Thus, sets of neuron-specific and

glia-specific genes are active, and will determine the cell-fate of the clones. The two stem cell-lines are observed over the course of three days. Among the scientific questions of interest raised by the biologists (Goff et al. (2007)) are; (a) How do the time course profiles of the glial-like (L2.3) and neuron-like (L2.2) cell-lines differ?; (b) Are there sets of genes for which the expression converges(diverges) between the glial-like and neuron-like cell populations?; (c) How dominant is the "pre-programming" effect?

For this two-factor experiment (cell-line and time), a two-level mixture model is appropriate. (We will focus on the two-level model in detail in the paper, and briefly discuss the generalization to multiple levels in the conclusion.) We introduce the model notation in the context of the above experiment. We denote expression of gene  $g$  in the glial-like population (L2.3) by  $\mathbf{x}_g$ , and in the neuron-like (L2.2) population by  $\mathbf{y}_g$ . The feature vectors,  $\mathbf{x}_g$  and  $\mathbf{y}_g$ , represent the time-course expression profiles of gene  $g$  in the glia and neuron cell-lines, respectively. Preliminary analysis indicates that groups of genes exhibit a similar time-course expression profile in the glia cell-line, but differ in the neuron cell-line. Furthermore, the glia cell-line is also associated with larger differential expression over time. Thus, if the feature vectors ( $\mathbf{x}_g, \mathbf{y}_g$ ) are clustered together, as a single  $2 \times T$ -dimensional feature, the large expression changes in the glia cell-line may dominate the clustering, and we might miss the more subtle expression patterns in the neuron cell-line. To resolve this issue, and help identify groups of genes whose activity is neuron-specific, we use a two-level mixture model. Thus, we allow for a total of  $K$  clusters at the 1st level, representing the clustering of the glia cell-line data. Within each of the  $k = 1, \dots, K$  clusters we allow for  $L_k$  2nd-level (sub)clusters, representing distinct expression profiles in the neuron cell-line. Let  $R_g$  and  $Z_g$  be two gene specific indicators, denoting the cluster labels at the 1st and 2nd levels. Our model assumes that

$$Pr(\mathbf{x}_g, \mathbf{y}_g \mid R_g = k, Z_g = l) \sim MVN(\mu_{kl}, \Sigma_{kl}),$$

where  $\mu_{kl} = (\mu_k, \mu_{l(k)})$  and  $\Sigma_{kl}$  represent the mean and variance-covariance matrix of the  $l$ th second level cluster within the  $k$  first level cluster.  $E(\mathbf{x}_g \mid R_g = k, Z_g = l) = \mu_k$  is the expression profile for the glia cell-line in cluster  $k$ , common to all sub-clusters  $l(k) = 1, \dots, L_k$ .  $E(\mathbf{y}_g \mid R_g = k, Z_g = l) = \mu_{l(k)}$  is the expression profile for the neuron cell-line in the  $l$ th second level cluster, within the first level cluster  $k$ . To further enhance interpretability of the clusters,

we parameterize the cluster means as  $\mu_{kl} = W\beta_{kl} = W(\alpha_k, \gamma_{kl})'$ , where  $\alpha_k$  denotes the 1st level cluster specific parameters, and  $\gamma_{kl}$  the 2nd level specific parameters.  $W$  is a design matrix for the multi-factor experiment, and reflects a scientific question of interest. We perform subset selection on the *parameters*, not the dimensions, and thus obtain cluster means that are directly interpretable in terms of the between-experimental factors, and within-experimental factor expression. We discuss specific choices of parameterizations in Section 2.

While we focus on a two-factor experiment in this paper, the multi-level cluster model is generally applicable to e.g. experiments involving multiple species, or varying treatment dosages and regiments. In this examples, as in our study, it is of interest to focus particularly on differential effects across levels of an experimental factor of interest (e.g. species, dose).

A few other schemes with a multi-level flavor have been proposed. Li (Li; 2005) introduced a layered mixture model to allow for flexible within-cluster structures. Akin to mixture discriminant analysis (MDA) (Hastie and Tibshirani; 1996) for classification, each cluster (class) is assumed to come from a mixture of normals, and can thus incorporate more complex cluster (class) shapes. The number of clusters is assumed known, and clusters do not share any mixture components with other clusters. Our multi-level mixture model differs from Li’s approach in that an unknown number of clusters may share components and model parameters, and that the levels of the mixture relate to the experimental factors. Yuan and Kendziorski (Yuan and C.Kendziorski; 2006) recently proposed a multi-level approach to gene clustering. Each cluster is assumed to be generated from a mixture of differential expression patterns (over-expressed, under-expressed, and no differential expression). An empirical Bayes strategy is adopted to fit the model. The motivation is that the clustering induces a regularization of the gene effect estimates, and thus power of detection of differential expression is increased. Our multi-level approach allows for a more flexible parametrization of the cluster means across multiple experimental conditions. We identify differential expression patterns both within and between the experimental factors through model subset selection.

The paper is structured as follows. In Section 2 we introduce the multi-level mixture model,  $\mathcal{MLX}_{\mathcal{L}}$ , and propose a method for subset selection and validation of the number of clusters. In Section 3 we apply  $\mathcal{MLX}_{\mathcal{L}}$  to a multi-factor gene expression data set. In Section 4, we illustrate the strengths our approach on several simulated data sets. We conclude this paper

with a discussion.

## 2 The $MI\mathcal{X}_{\mathcal{L}}$ model.

### 2.1 A multi-level parametrization for model-based clustering

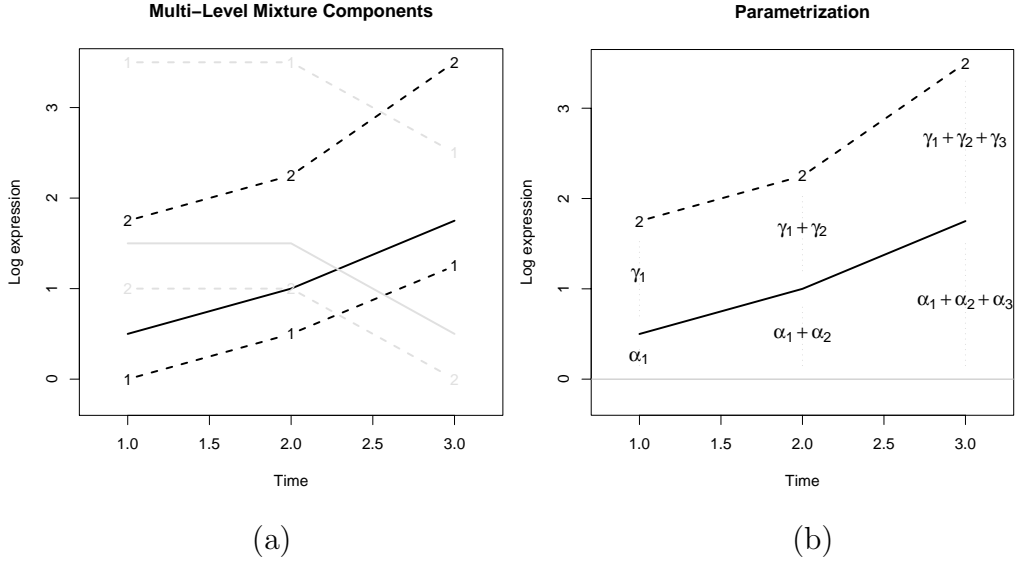


Figure 1: (a) An illustration of a Mixture Model with 2 levels; Two 1st level clusters for cell-line 1 (solid lines). The two sets of dashed lines represent the corresponding sub-clusters (level 2) for cell-line 2. Thus, here  $K = 2$ , and  $L_1 = 2$ ,  $L_2 = 2$ . (b) The "Dynamic DE (differential expression) parametrization". The  $\alpha$  parameters model the time course expression profile for cell-line 1, whereas the  $\gamma$  parameters model the time course of *cell-line differential* expression.

We present the  $MI\mathcal{X}_{\mathcal{L}}$  model in the case of two populations (e.g. cell lines) of interest, and samples from both of these populations are collected across  $T$  time points as in the experiment described in the introduction. (We briefly discuss generalizations in section 5.) Let  $\mathbf{x}_g$  denote the observations across  $T = 3$  time points for gene  $g$  ( $g \in \{1, \dots, G\}$ ) in cell line 1 (glia like), and similarly  $\mathbf{y}_g$  in cell line 2 (neuron like). We denote the total number of clusters at the 1st level (glia) by  $K$ , and the number of 2nd level clusters (neuron) within each 1st level cluster  $k$  by  $L_k$ . Let  $R_g$  and  $Z_g$  be two gene specific cluster indicators denoting the cluster labels at the 1st and 2nd level. Our model assumes that

$$Pr(\mathbf{x}_g, \mathbf{y}_g \mid R_g = k, Z_g = l) \sim MVN(\mu_{kl}, \Sigma_{kl}), \quad (1)$$

where  $\mu_{kl} = (\mu_k, \mu_{l(k)})$  and  $\Sigma_{kl}$  represent the mean and variance-covariance matrix of the  $l$ -th

second level cluster within the  $k$ -th first level cluster. The first  $T$  components of the  $\mu_{kl}$  vector,  $\mu_k$ , correspond to the mean levels of  $\mathbf{x}_g$ . The last  $T$  components,  $\mu_{l(k)}$ , correspond to the mean levels of  $\mathbf{y}_g$  (Fig. 1 (a)). Note, if we let  $L_k = 1$  for all  $k$ , the model formulation in equation (1) coincides with a standard mixture model.

The multi-level framework allows for various interpretable parameterizations at each level. We parameterize the cluster mean as  $\mu_{kl} = W\beta_{kl} = W(\alpha_k, \gamma_{kl})'$ . The  $\alpha$  parameter vector represents 1st level specific parameters, and the  $\gamma$  vector represents the 2nd level model parameters. Next, we will consider the following three parameterizations in detail:

**Parametrization I.** *Mean differential expression.*

$$\begin{aligned}\mu_k &= (\alpha_{k1}, \alpha_{k2}, \alpha_{k3})' \\ \mu_{kl} &= (\alpha_{k1}, \alpha_{k2}, \alpha_{k3}, [\alpha_{k1} + \gamma_{kl1}], [\alpha_{k2} + \gamma_{kl2}], [\alpha_{k3} + \gamma_{kl3}])'\end{aligned}$$

The vector  $(\alpha_{k1}, \alpha_{k2}, \alpha_{k3})$  represents cell-line 1 expression at time points  $(t_1, t_2, t_3)$  in cluster  $k$ . The vector  $(\gamma_{kl1}, \gamma_{kl2}, \gamma_{kl3})$  represents the cell-line differences at each time point in sub-cluster  $l(k)$ . Here, the main scientific question addressed is thus the differential expression between the cell-lines, at any given time point.

**Parametrization II.** *Dynamical differential expression.*

$$\begin{aligned}\mu_k &= (\alpha_{k1}, [\alpha_{k1} + \alpha_{k2}], [\alpha_{k1} + \alpha_{k2} + \alpha_{k3}])' \\ \mu_{kl} &= (\alpha_{k1}, [\alpha_{k1} + \alpha_{k2}], [\alpha_{k1} + \alpha_{k2} + \alpha_{k3}], [\alpha_{k1} + \gamma_{kl1}], [\alpha_{k1} + \alpha_{k2} + \gamma_{kl1} + \gamma_{kl2}], \\ &\quad [\alpha_{k1} + \alpha_{k2} + \alpha_{k3} + \gamma_{kl1} + \gamma_{kl2} + \gamma_{kl3}])'\end{aligned}$$

In the second parametrization (see Fig. 1(b)), the time course profile of the glial-like population is modeled directly, and e.g. flat time profiles are efficiently represented ( $\alpha_{k2} = \alpha_{k3} = 0$ ). The  $\gamma$ -vector represents the differential expression *time-course* of the two cell-lines (e.g. parallel ( $\gamma_{kl2} = \gamma_{kl3} = 0$ ), or divergent ( $\gamma_{kl1} = 0, \gamma_{kl3} \neq 0$ )).

**Parametrization III.** *Pre-programming differential expression.*

$$\begin{aligned}\mu_k &= (\alpha_{k1}, [\alpha_{k1} + \alpha_{k2}], [\alpha_{k1} + \alpha_{k2} + \alpha_{k3}])' \\ \mu_{kl} &= (\alpha_{k1}, [\alpha_{k1} + \alpha_{k2}], [\alpha_{k1} + \alpha_{k2} + \alpha_{k3}], [\alpha_{k1} + \gamma_{kl1}], [\alpha_{k1} + \gamma_{kl1} + \gamma_{kl2}],\end{aligned}$$

$$[\alpha_{k1} + \gamma_{kl1} + \gamma_{kl2} + \gamma_{kl3}]'$$

The third parametrization efficiently models each time course-profile, and a main differential cell-line effect for time point  $t_1$ . Thus, flat glia time profiles are efficiently represented ( $\alpha_{k2} = \alpha_{k3} = 0$ ), and similarly flat time course profiles for the neuron cell-line are obtained if  $\gamma_{kl2} = \gamma_{kl3} = 0$ . The only direct comparison between the cell-lines is at time point  $t_1$ .

Other data sets and experimental structures may require a different set of parameterizations. Ultimately, the choice of parametrization should depend on the biological context, and the scientific questions of interest.

In all parameterizations, the variance-covariance matrix  $\Sigma_{kl}$  also includes parameters specific to the levels of the model:

$$\Sigma_{kl} = \begin{bmatrix} \Sigma_k^X & \Sigma_{kl}^{XY} \\ \Sigma_{kl}^{YX} & \Sigma_{kl}^Y \end{bmatrix}.$$

The variance-covariance structure allows for dependencies between gene expression measurements at all time points and all levels. We further assume that, conditional on the multi-level cluster assignments, the genes are independent of each other. Therefore, we have the following complete data likelihood:

$$\begin{aligned} Pr(\mathbf{X}, \mathbf{Y}, \mathbf{R}, \mathbf{Z} \mid \Psi) &= \prod_{g=1}^G Pr(\mathbf{x}_g, \mathbf{y}_g, R_g, Z_g \mid \Psi) \\ &= \prod_{g=1}^G \prod_{k=1}^K \prod_{l=1}^L [Pr(\mathbf{x}_g, \mathbf{y}_g \mid R_g = k, Z_g = l) \pi_{kl}]^{I(R_g=k, Z_g=l)}, \end{aligned}$$

where  $\Psi = \{\mu_{kl}, \Sigma_{kl}, k = 1, \dots, K, l(k) = 1, \dots, L_k\}$  represents the overall parameter set of the model, and  $\pi_{kl}$  are the mixing proportions. Due to the multi-level parametrization and the general variance-covariance structure, the complete data likelihood does not factorize into terms over which maximization for each parameter can be carried out separately. Therefore, the standard maximization step for the Expectation-Maximization algorithm of the mixture models does not lead to closed form updates. To resolve this issue, we derive a Profile Expectation-Maximization (PEM) algorithm that relies on the factorization of the likelihood into the likelihood of the 1st level of the hierarchy, and the conditional likelihood of the 2nd

level of the hierarchy given the first level. Additionally, each component of the factorized likelihood is maximized by profiling of the corresponding expected complete data log likelihood. Although purely motivated by the factorization of the expected full data likelihood, the proposed PEM algorithm can be classified as a Expected Conditional Maximization algorithm (ECM) proposed by Meng and Rubin (1993).

## 2.2 Profile Expectation-Maximization (PEM) algorithm for fitting the multi-level mixture model

We now describe the profile Expectation-Maximization algorithm for fixed  $K$  and  $L_k, \forall k$ . Here,  $r$  refers to  $r$ -th EM iteration, and we suppress the dependence on  $r$  to ease the notation.

**Initial values.** The algorithm requires initial values of  $\pi_{kl}$  and  $\mu_{kl}$  and  $\Sigma_{kl}$ . These are obtained via a k-means clustering of the data with  $M = \sum_k L_k$  clusters. We then collapse a subset of the  $M$  clusters for the 1st level data ( $\mathbf{x}_g$ , glia cell-line data) to form  $K$  1st level clusters. Within each 1st and 2nd level cluster, we estimate the parameters. Details of the initialization are discussed in the supplementary materials.

**E-step.** This step is a regular E-step in fitting mixtures of multivariate normals. We have posterior class probabilities given by

$$\hat{\eta}_{gkl}^{(r)} = Pr(R_g = k, Z_g = l \mid \mathbf{x}_g, \mathbf{y}_g, \Psi^{(r-1)}) = \frac{MVN(\mathbf{x}_g, \mathbf{y}_g \mid \mu_{kl}^{(r-1)}, \Sigma_{kl}^{(r-1)}) \pi_{kl}^{(r-1)}}{Pr(\mathbf{x}_g, \mathbf{y}_g \mid \Psi^{(r-1)})}.$$

**M-step.** In the M-step, we are dealing with the following maximization problem:

$$\sum_{g=1}^G \sum_{k=1}^K \sum_{l=1}^{L_k} \left( -\frac{1}{2} \hat{\eta}_{gkl} (\mathbf{u}_g - W\beta_{kl})' \Sigma_{kl}^{-1} (\mathbf{u}_g - W\beta_{kl}) - \frac{1}{2} \hat{\eta}_{gkl} \log |\Sigma_{kl}| \right), \quad (2)$$

where  $W$  is the design matrix,  $\beta_{kl} = (\alpha_k, \gamma_{kl})$ , and  $\mathbf{u}_g = (\mathbf{x}_g, \mathbf{y}_g)$ . (We provide explicit forms of the design matrices of the three parameterizations in the supplementary materials.) As discussed in the previous section, the main reason for a nonstandard mixture model M-step is due to the cross-talk between the two levels. The first part of the parameter vector  $\beta_{kl}$ , denoted by  $\alpha_k$ , is the same for all  $l$ , and similarly the left upper diagonal block  $\Sigma_k^X$  of  $\Sigma_{kl}$ .

Hence, the corresponding estimates need to pool information across all second level clusters of the  $k$ th 1st level cluster. We use a regularized profiling method for maximizing the expected complete data log likelihood given in equation 2. Our general iterative scheme is to factorize the joint likelihood of  $\mathbf{x}_g$  and  $\mathbf{y}_g$  as the product of marginal likelihood of  $\mathbf{x}_g$ , and the conditional likelihood of  $\mathbf{y}_g$  given  $\mathbf{x}_g$ . We first maximize the marginal likelihood of  $\mathbf{x}_g$  by profiling to obtain estimates of  $\alpha_k$  and  $\Sigma_k^X$ . Given these estimates, we maximize the conditional likelihood of  $\mathbf{y}_g$  given  $\mathbf{x}_g$ , again by profiling over the mean and the variance-covariance matrix. We thus obtain estimates  $\gamma_{kl}$ , and the 2nd level components of  $\Sigma_{kl}$ . The estimates of  $\alpha_k$  and  $\gamma_{kl}$  are obtained via two weighted generalized least squares procedures. We provide details of these derivations, as well as computational considerations, in Section 2 of the supplementary materials.

### Profile EM algorithm

1. **E-step.** Compute  $\hat{\eta}_{gkl}$ ,  $g = 1, \dots, G$ ,  $k = 1, \dots, K$  and  $l = 1, \dots, L_k$ .
2. **M-step.**
  - (a) Update  $\hat{\pi}_{kl}$ ,  $k = 1, \dots, K$  and  $l = 1, \dots, L_k$ .
  - (b) **M-step-Profile-1.** (marginal  $\mathbf{x}_g$ )
    - i. Update  $\Sigma_k^X$ .
    - ii. Update  $\mu_k$  by reestimating  $\hat{\alpha}_k$  via weighted generalized least squares.
    - iii. Iterate (i) and (ii) till convergence.
  - (c) **M-step-Profile-2.** (conditional  $\mathbf{y}_g|\mathbf{x}_g$ )
    - i. Update conditional covariances and the mean  $\Sigma_{kl}^{YX}$ ,  $\Sigma_{kl}^{Y|X}$ ,  $\Sigma_{kl}^Y$  and  $\mu_{kl}^{Y|X}$ .
    - ii. Update  $\mu_{kl}$  by reestimating  $\gamma_{kl}$  via weighted generalized least squares, and set  $\hat{\beta}_{kl} = (\hat{\alpha}_k, \hat{\gamma}_{kl})'$ .
    - iii. Iterate (i) and (ii) till convergence.

All of the update states in the PEM algorithm are closed form, which makes the implementation of the multi-level mixture free of black-box optimization. Although the profiling steps could in principle benefit from internal iterations (iii above), it is in general more advantageous

to spend the computing time on the outer EM iterations (1 and 2).

## 2.3 Model selection

Model selection in multi-level model-based clustering involves; (1) selection of the appropriate parametrization for each cluster; (2) selection of the number of 1st level clusters  $K$ , and the number of (sub)clusters  $\{L_k, k = 1, \dots, K\}, \forall k$ .

### 2.3.1 Cluster parameterizations and subset selection

Let us first consider the case with fixed  $K$  and  $\{L_k, k = 1, \dots, K\}, \forall k$ . We want to select a sparse representation for each cluster to enhance the objective interpretability of the clustering outcome. For example, is a particular cluster model representing (i) a static cell-line difference, or (ii) a dynamic one, and if so for which time-points do the cell-lines really differ?

Recently, several papers have appeared on the topic of variable selection for model based clustering. These papers focus on the selection of a subset of variables, or dimensions of the feature vector, that can discriminate between cluster components (e.g., Friedman and Meulman (2004), Law et al. (2004), Raftery and Dean (2006), Hoff (2006), Tadesse et al. (2005)).

In our parametrization of the cluster means, we allow for cluster specific descriptions of contrasts between variables *within* a cluster, as well as *between* clusters. For each cluster we allow for a subset of *parameters* to be non-zero. The subset of coefficients that are set to zero do not necessarily correspond to dimensions that are irrelevant for clustering. Take as an example parametrization I; If for cluster  $k$ , a subset of  $\alpha_k$  are set to 0, then these dimensions are unrelated to the clustering; If, however, parameters  $\gamma_{kl}$  are set to 0, this implies that the cluster consists of a set of genes for which there is no cell-line effect.

To perform cluster subset selection we threshold the posterior probabilities  $\eta_{gkl}$  to obtain cluster specific data sets of size  $n_{kl}$  for each sub-cluster  $\{k, l\}$  (or  $n_k$  for a 1st level cluster  $k$ ). For a 1st level cluster,  $k$ ,  $\mathbf{x}_g = W_K \alpha_k + \epsilon$ ,  $\epsilon \sim N(0, \Sigma_K^X)$  for the  $n_k$  genes  $g$  in the cluster.  $W_K$  and  $\Sigma_K^X$  refer to the 1st level specific partition of the design matrix and covariance matrix respectively. After hard thresholding of the posterior probabilities, model selection for each cluster has thus been reduced to a model selection task in regression. We hold  $\Sigma_K$  fixed during the model selection. We select the subset of non-zero  $\alpha_k$  parameters that minimizes the BIC,

and thus obtain a cluster specific model. Model selection for a 2nd level cluster proceeds in a similar fashion. Given the 1st level parameters  $\alpha_k$  and the 1st level data  $\mathbf{x}_g$ , we perform model selection in the regression setting for  $\mathbf{y}_g|\mathbf{x}_g$ . We identify a subset of non-zero 2nd level parameters,  $\gamma_{kl}$ , that minimizes the BIC. A detailed discussion of the model selection can be found in supplementary Section 2.2.1.

### 2.3.2 Selecting the number of clusters.

The selection of the number of clusters is usually based on criteria such as BIC (Bayesian Information Criterion), CIC (Clustering Information Criterion) or MDL (Minimum Description Length) (e.g. Fraley and Raftery (2002), Raftery and Dean (2006)). Here, we will use BIC. Let us consider a multi-level parametrization where the dimensionality of the data vectors at the 1st level is  $Dim(1)$ , and at 2nd level is  $Dim(2)$ . We denote the model coefficients at the 1st level by  $\alpha_k, k = \{1, \dots, K\}$ , and the model coefficients at the 2nd level by  $\gamma_{kl}, l(k) = \{1, \dots, L_k\}$  for all  $k = \{1, \dots, K\}$ . The sub-cluster structure of the model is summarized by the vector  $\mathbf{L}_K = \{L_k, k = 1, \dots, K\}$ .

In the previous section we considered subset model selection for each cluster  $\{k, l\}$ . Thus, the number of non-zero coefficients  $\alpha_k \neq 0$  may be less than  $Dim(1)$ , and similarly for  $\gamma_{kl}$ . We denote the number of non-zero coefficients at each cluster  $\{k, l\}$  by  $(dim(\alpha_k), dim(\gamma_{kl}))$ . We gather all parameters into a set  $\Theta(K, \mathbf{L}_K) = \{\pi_{kl}, \alpha_k, \gamma_{kl}, \Sigma_{kl}, k = \{1, \dots, K\}, l(k) = \{1, \dots, L_k\}\}$ . Then, the total model complexity is given by

$$\begin{aligned}
p(\Theta(K, \mathbf{L}_K)) &= \left[ \sum_{k=1}^K \left( dim(\alpha_k) + \sum_{l=1}^{L_k} dim(\gamma_{kl}) \right) \right]_{(1)} + \\
&+ \left[ \frac{K Dim(1)(Dim(1) - 1)}{2} \right]_{(2)} + \left[ \left( \sum_{k=1}^K L_k \right) - 1 \right]_{(3)} \\
&\left[ \left( \sum_{k=1}^K L_k \right) \left( Dim(1) Dim(2) + \frac{Dim(2)(Dim(2) - 1)}{2} \right) \right]_{(4)},
\end{aligned}$$

where term (1) is the number of mean parameters estimated at the 1st and 2nd levels, term (2) is the 1st-level covariance estimates, term (4) is the 2nd-level covariance estimates and cross-covariance estimates between the 1st and 2nd levels, and term (3) is the number of estimated

cluster proportions. For each given  $K$  and  $\mathbf{L}_K$  we can compute the log-likelihood:

$$l(\Theta(K, \mathbf{L}_K)) = \sum_{g=1}^G \log \left( \sum_{k=1}^K \sum_{l=1}^{L_k} \pi_{kl} \phi(\mathbf{x}_g, \mathbf{y}_g; W\beta_{kl}, \Sigma_{kl}) \right).$$

We then compute the BIC value as

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

The model-space  $\{K, \mathbf{L}_K\}$  is very large, and a complete search across all numbers of clusters and sub-cluster constellations is prohibitively expensive. We explored several different search strategies for identifying the optimal multi-level model. The best performance was obtained using a backward search. We thus searched over a total number of clusters  $M = \sum_k L_k$ , where for each  $M$  we considered a multi-level model with  $K = M, \dots, 1$  1st level clusters. We provide a complete outline of the model search in a flow-chart in supplementary Section 2.2.2.

For both subset selection, and the selection of the number of clusters, we adopt greedy searches. While it is true that such schemes can converge to local optima, a fully exhaustive search is computationally prohibitive. To remedy the problem we run the full algorithm several times while initiating from different starting values.

## 3 Application to Data

### 3.1 The proliferating cell-line data

We apply the  $\mathcal{MIX}_{\mathcal{L}}$  model with subset selection to the data set of proliferating stem cell lines (Goff et al. (2007)) introduced in Section 1 of the paper. mRNA was extracted for array analysis at  $t = 0, 1$  and 3 days after the withdrawal of a growth factor from the medium (to speed up differentiation). The ABI system rat-chips, with 28,000 probes, were used for the array experiments. Of these probes, we studied a subset of 15,111 probes with complete annotation. We refer the reader to Section 3 of the supplementary materials for a description of the pre-processing of these data. Preliminary significance analysis of the expression data identified 780 genes of the 15,111 as being differentially expressed between the cell lines and/or

time points at FDR (False Discovery Rate) 1% (using the Welch F-test and the Benjamini-Hochberg p-value corrections). Similar results were obtained using the limma software of Smyth (2004). For each of the 780 selected genes, we computed the mean gene profile across replicates, and standardized the mean profiles to have standard deviation 1, with a baseline of expression 0 for  $t = 0$  in the glial like population. The final data set to be analyzed is thus of dimension 780 by 5. We denote gene expression in the glial-like population (L2.3) by  $\mathbf{x}$ , where  $Dim(x)$  is 2 (for  $t = 1$  and  $t = 3$ ). We denote the gene expression in the neuron-like population by  $\mathbf{y}$ , where  $Dim(y)$  is 3 ( $t = 0, 1$  and  $t = 3$ ).

### 3.2 Subset selection of cluster model profiles

We first investigate the impact of subset selection on clustering, by fitting single-level models with the three parameterizations,  $W_I$ ,  $W_{II}$ , and  $W_{III}$  described in Section 2.

$K$	$W_I : \sum_k 1\{\beta_k = 0\}$	$W_{II} : \sum_k 1\{\beta_k = 0\}$	$W_{III} : \sum_k 1\{\beta_k = 0\}$
5	2	5	7
6	6	4	7
7	4	4	7
8	2	5	5
9	3	6	6
10	5	7	9
11	5	7	7
12	6	8	10

Table 1: Number of coefficients set to 0 by subset selection for the single-level fits with the three parameterizations.

Figure 2 (c) depicts the BIC curves obtained for various numbers of clusters  $K$  in a single-level fit. The solid line is the BIC curve obtained without subset selection (i.e. a standard gaussian mixture model). The dashed and dotted lines are annotated with "1", "2" and "3", referring to the three parameterizations  $W_I, W_{II}$  and  $W_{III}$  respectively. Across all numbers of clusters, the  $W_{III}$  (cell fate pre-programming) parametrization provides the best fit, as indicated by the lower BIC values. The sparsity of each model is summarized in Table 1. With an efficient parametrization,  $K = 8$  and  $K = 9$  are equally competitive. The  $W_{III}$  parametrization identifies cluster profiles that are static between  $t = 0$  and  $t = 1$ , indicating a later developmental activity in one or both cell lines (e.g. clusters 1, 8) (see Figure 2 (a)). We

Single-level		Multi-level		
$K$	$\sum_k 1\{\beta_k = 0\}$	$(M, K)$	$\sum_{kl} 1\{\beta_{kl} = 0\}$	multi-level constraints
5	7	(5,4)	5	2
6	7	(6,5)	11	2
7	7	(7,6)	2	2
8	5	(8,7)	4	2
9	6	(9,7)	4	4
10	9	(10,8)	4	4
11	7	(11,8)	11	6
12	10	(12,10)	6	4

Table 2: Number of coefficients set to 0 by subset selection for the single- and multi-level fits using parametrization  $W_{III}$ .

focus on the most efficient parametrization ( $W_{III}$ ) in the next section.

### 3.3 Multi-level model-based clustering of the cell-line study.

In Figure 2 (a) we depict the clustering outcome of a single-level fit using parametrization  $W_{III}$ . As can be seen from the figure, the glial like population exhibits larger time differential effects than the neuron like population. Furthermore, for some clusters (e.g. clusters 3 and 4 at the top of the left panel of the figure), the glial like cluster expression profiles almost coincide, whereas the neuron cluster profiles differ substantially. To identify neuron specific variations, we will thus treat the glial like population data as the 1st level in the  $\mathcal{MLX}_{\mathcal{C}}$  model.

Figure 2 (d) illustrates the additional efficiency of multi-level parameterizations. With the exception of the case with 5 total clusters, the multi-level fit always produces a lower BIC value. We select  $M = 9$  clusters total (cf. 8 clusters with the single-level fit), and  $K = 7$  1st level clusters. That is, we gain one more cluster. One can view this as a re-allocation of model complexity. In model selection we aim to balance the fit and model complexity (number of parameters). By setting some cluster parameters to 0 (within-cluster subset selection), and letting some clusters share parameters at the 1st level (between-cluster parameter constraints), we save on complexity and can "afford" to form another cluster. In Table 2 we summarize the results, listing for each cluster the number of parameters set to 0 via subset selection, as well as the number of parameter constraints from the multi-level fit. For this data set, the larger gains are made when the number of clusters increase. For example, 11 out of 49 parameters

were set to 0 (or constrained) in the ( $M = 11, K = 8$ ) multi-level fit.

The cluster profile that is the most unique in the multi-level fit is cluster 9 (Figure 2 (b) compared with (a)), which as we shall see in the discussion below provides some interesting insight into neuron specific activity. In addition, we have identified two groups of gene clusters (3 and 4, 5 and 6) for which the glial like population exhibits identical expression patterns, and a sub-division of genes exhibit radically different expression patterns in neurons. Expression profiles of the 9 clusters generated by the  $\mathcal{MIX}_{\mathcal{L}}$  fit are depicted in supplementary Figure 1. Clusters 1 to 6 have an almost complete correspondence between the single-level fit (Figure 2(a)) and the multi-level fit (Figure 2(b)). However, the multi-level model allows us to objectively state that clusters 3 and 4, as well as 5 and 6, constitute clusters for which there is a neuron-specific difference of expression only. Clusters 7 to 9 do not have clear counterparts among the single-level clusters.

### 3.4 Interpreting the clustering outcome

The  $\mathcal{MIX}_{\mathcal{L}}$  model description allows for one extra cluster compared with the single-level fit. In addition,  $\mathcal{MIX}_{\mathcal{L}}$  provides a sparse representation for each cluster. Clusters 3 and 4, as well as clusters 5 and 6, form sub-clusters for which the expression pattern coincides in the glial like population, but differs in the neuron like population. We used GOstat (Beissbarth and Speed (2004)) to identify the significant Gene Ontology (GO) categories for each of the 9 clusters. Tables 1-4 in the supplementary materials report top 10 significant GO categories for the 9 clusters. Among all clusters (780 genes), developmental terms and neurogenesis are over-represented compared with all annotated probes (15.111) on the array.

Cluster 9 was detected as a result of the efficient  $\mathcal{MIX}_{\mathcal{L}}$  parametrization, and contains genes that are upregulated in neurons compared with glia at all times. Many of the top GO categories associated with this cluster are linked to phosphorus binding. Phosphor is an activator of BDNF (Brain-derived neurotrophic factor) binding, a primary regulator of dendrite branching during neuron development.

Clusters 3 and 4 form one set of sub-clusters. The expression of the glia cell-line is steadily increasing for both clusters. In the neuron cell-line, cluster 3 represents genes with a steady high level of expression, whereas cluster 4 represents a profile of expression which increases

over time. Genes in cluster 3 are associated with neuron development and axon formation. Genes in cluster 4 are linked with dendrite formation and growth. These sub-clusters thus appear to be related to different neuron specific developmental processes.

Clusters 5 and 6 is another sub-cluster formation. The expression profile of the glia cell-line is decreasing in both clusters. Genes in cluster 6 are persistently underexpressed in the neuron cell-line, whereas genes in cluster 5 are overexpressed, compared with glia. Genes in cluster 6 (underexpressed in neurons) are primarily associated with acid synthesis. Glial cells are believed to synthesize some acids that assist in neuron development and migration. Cluster 5 (overexpressed in neurons) is associated with acid metabolism, which is one process by which neurons generate neurotransmitters.

The annotations of the 9 clusters are summarized in the supplementary materials. There, we also discuss several transcription factors detected by mining the genomic sequences of genes in the neuron-specific sub-clusters for regulatory motifs.

## 4 Simulation studies

We use the estimated best single-level (SF) and multi-level (MF) fits (see Section 3) to generate mixtures of multivariate normal data from realistic scenarios. We use the simulated data sets to validate; (a) selection of the number of (sub)clusters; and (b) subset model selection for each cluster. The best single level model is referred to as  $Mod(1)$ , and the best multi-level model as  $Mod(2)$ .

$Mod(1)$  is a single-level model with  $K = 8$  clusters. The cluster means for this model are depicted in Figure 2 (a). The cluster means are parameterized with  $5 * 8$  coefficients, and 5 parameters were set to 0 by subset selection.  $Mod(2)$  is the multi-level model with  $M = 9$  clusters, and  $K = 7$  1st level clusters. 4 parameters were set to 0 by the subset selection, and 4 parameters constrained by the multi-level structure of the model (see Table 2). We generate 50 simulated data sets (of the same dimensions as the original data) each from  $Mod(1)$  and  $Mod(2)$ . We then fit single- and multi-level models, and perform cluster model subset selection. For each simulated data set, we record the selected number of (sub)clusters. We also compare the selected subset model (for the true number of clusters) to the true model, and record the

total number of selection errors (the number of coefficients erroneously set to 0, or non-zero).

In Table 3 we summarize the results (see also supplementary Figure 3). Cluster subset selection always produces a better model in terms of the BIC validation index (supplementary Figure 3 (a)). In addition, a multi-level fit always reduces the BIC compared with a single-level fit when a multi-level structure is truly present ( $Mod(2)$ ), and produces comparable results with a single-level fit when a single-level structure is true ( $Mod(1)$ ) (see supplementary Figures 3 (c) and (d)). The subset selection performance is highly satisfactory (supplementary Figure 3 (b)), i.e., the generative subset model is often correctly identified. The multi-level approach produces subset selection results closer to the correct subset model, compared with the single-level approach. This is because selection is undertaken separately for the 1st and 2nd level clusters (see supplementary materials).

	SF( $Mod\ 1$ )	MF( $Mod\ 1$ )		
	K=M	K=7	K=8	K=9
M=8	88	2	<b>86</b>	
M=9	12	0	4	<b>8</b>

	SF( $Mod\ 2$ )	MF( $Mod\ 2$ )				
	K=M	K=6	K=7	K=8	K=9	K=10
M=7	2	<b>2</b>	0			
M=8	14	<b>4</b>	<b>6</b>	2		
M=9	30	<b>0</b>	<b>14</b>	<b>16</b>	4	
M=10	54	<b>0</b>	<b>4</b>	<b>14</b>	<b>22</b>	12

Table 3: Top panel: The percent of times a number ( $M, K$ ) of clusters are selected with the single-level and multi-level fits for the  $Mod(1)$  data. The correct  $K = 8$ . Both fitting strategies perform well, and the multi-level fit correctly identifies a single-level fit (bold face in table) in almost all cases. Lower panel: The percent of times a number ( $M, K$ ) of clusters are selected with the single- and multi-level fits for the  $Mod(2)$  data. The correct  $M = 9$ , with 7 1st level clusters ( $K = 7$ ). In almost all cases, the multi-level fit correctly identifies a sub-cluster structure (bold face in table), rather than single-level model.

In Table 3 we present the selected number of clusters for the  $Mod(1)$  and  $Mod(2)$  data sets, using the single-level and multi-level fits. In the case of  $Mod(1)$  data, the multi-level fit in almost all cases identifies the single-level fit as the correct model structure. In the case of  $Mod(2)$  data, the multi-level fit in almost all cases identifies a multi-level fit (with sub-clusters) as the correct model structure. In the case of  $Mod(2)$ , both fitting strategies have

trouble identifying the correct total number of clusters. The reason for this is that cluster 7 in  $Mod(2)$  is sparsely populated. In some simulations, cluster 7 is split into 2 clusters, producing a total of  $M = 10$  clusters. Sometimes "genes" in cluster 7 are simply allocated to nearby clusters, producing a total of  $M = 8$  clusters.

In summary, the multi-level fit can correctly identify a single-level model as well as a multi-level model. In addition, the BIC is much reduced if the multi-level structure of the data is accounted for. Subset selection also reduces the BIC, in both single-level and multi-level models. Our simulations thus illustrate the benefits of sparse multi-level representations of cluster profiles in model-based clustering.

## 5 Discussion

We propose a mixture model with multiple levels to more efficiently model multiple-factor experimental data. In addition, we introduce a subset selection method to generate sparse representations of cluster profiles, under various parameterizations. We illustrate on real and simulated data that sparse, multi-level mixture models can substantially improve the fit, significantly reducing the BIC, compared with standard mixture models. We show that our multi-level mixture modeling approach with subset selection can correctly identify both single-level and multi-level data structures. Furthermore, in our simulation setting, we show that the subset selection procedure is highly accurate.

Our multi-level approach identifies interesting and biologically relevant groups of genes in the proliferating cell-line data. A more thorough study of our findings is now underway in collaboration with biologists at Rutgers university. We stress that the findings we presented in this paper are preliminary. A small perturbation study, where we randomly altered 5 percent of the selected gene list, did provide clustering results that substantially overlapped with the original outcome (IQR 84 to 98 percent concordance). However, as additional data become available we expect that other cluster structures may be detected.

Efficient cluster model representations (multiple levels and subset selection) will have a larger impact in high-dimensional settings, e.g., time-course data with more time points. It is in these cases that a multi-level approach with subset selection has the largest potential

to substantially reduce the number of parameters in the model. In addition, while we did not consider efficient representations of the cluster covariances, this is another area in which modeling efficiency may be explored. Fraley and Raftery (2002) compared mixture models with parameterized cluster covariances. Incorporating covariance parametrization and subset selection into our multi-level approach is an interesting future research topic.

While we demonstrated our multi-level approach on a proliferating cell-line data, with a two-level factor of interest, the method can in theory be extended to more factors, and factors with more levels. Let us consider the case where we have a three-level factor of interest (e.g. three cell-lines), and denote the specific data sets by  $\mathbf{x}_g, \mathbf{y}_g$  and  $\mathbf{v}_g$  respectively. The most simple extension is to use one of the cell-lines as reference, e.g.  $\mathbf{x}_g$ . At the 2nd level of the modeling hierarchy we thus model contrasts with respect to the reference:  $\mathbf{y}_g, \mathbf{v}_g | \mathbf{x}_g$ . This approach is quite reasonable in experiments involving multiple species or strains, where a "wild-type" or reference strain constitutes a natural basis for comparison. In other experiments, a modeling hierarchy is induced by ordering the factor levels. Thus, the 1st modeling hierarchy models  $\mathbf{x}_g$ , the 2nd level models  $\mathbf{y}_g | \mathbf{x}_g$ , and the third level involves  $\mathbf{v}_g | \mathbf{x}_g, \mathbf{y}_g$ . The modeling structure at the third level is of the same mathematical form as the 2nd level, so falls under the  $\mathcal{MIX}_{\mathcal{L}}$  framework we presented in the paper. The generalization of the  $\mathcal{MIX}_{\mathcal{L}}$  model requires the selection of the optimal assignment of the factor levels to the levels of the modeling hierarchy. This constitutes an interesting research problem we plan to explore in the future.

The R implementation of the two-level  $\mathcal{MIX}_{\mathcal{L}}$  model is available from the corresponding author upon request.

## Acknowledgement

We thank Professor Ron Hart, and members of the Hart lab for generously sharing their data with us, and for helping us with a preliminary interpretation of the analysis outcome. We also thank the two referees, the associate editor, and the co-editor for many helpful suggestions.

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. SK is supported by a WARF grant from the University of Wisconsin, Madison, and NIH grant 1-R01-H603747-01.

## List of Figures and Tables

Figure 1: (a) An illustration of a Mixture Model with 2 levels; Solid line (black and gray): Two 1st level clusters for cell-line 1. The two sets of dashed lines (black and gray) represent the corresponding sub-clusters (level 2) for cell-line 2. Thus, here  $K = 2$ , and  $L_1 = 2$ ,  $L_2 = 2$ . (b) The "Dynamic DE (differential expression) parametrization". The  $\alpha$  parameters model the time course expression profile for cell-line 1, whereas the  $\gamma$  parameters model the time course of *cell-line differential* expression.

Figure 2: (a) Cluster mean profiles of the best single-level fit ( $K = 8$ ). The glial like population is depicted in the left panel, the neuron like in the right panel (parametrization  $W_{III}$ ) (b) Cluster mean profiles of the best multi-level fit  $K = 7$ ,  $M = 9$ , with two sets of sub-clusters (parametrization  $W_{III}$ ). (c) The BIC curves obtained using the single-level fit. Solid line: no subset selection. Dashed and dotted curves annotated with the respective parametrization ( $W_I, W_{II}, W_{III}$ ). The  $W_{III}$ -BIC curve is the lowest, indicating that the  $W_{III}$  parametrization is the most efficient for this data set. (d) The BIC curves obtained from the single- and multi-level fits, using the  $W_{III}$  parametrization. The multi-level fit always gives a lower BIC for the same total number of clusters ( $M$ ). The numbers in the figures ( $M, K$ ) refers to the total number of clusters, and the number of 1st level clusters respectively. The best BIC values is obtained with  $M = 9$  clusters total, and  $K = 7$  1st level clusters.

Supplementary Figure 1: The 9 clusters generated by the best  $\mathcal{MIX}_{\mathcal{L}}$  fit. The solid line represents the cluster mean, the dashed lines are  $\pm 1$  standard deviation (pointwise). The gray lines represent individual genes that have been allocated to each cluster.

Supplementary Figure 2: Cluster mean profiles of the best multi-level fit  $K = 7$ ,  $M = 9$ , with two sets of sub-clusters (parametrization  $W_{III}$ ).

Supplementary Figure 3: (a) The BIC value of the full model minus the BIC value after subset selection for both simulation settings: ( $Mod(1), Mod(2)$ ), and both fitting strategies (single- (SF) and multi-level (MF) fits). The BIC is always smaller after subset selection. (b)

Histograms of the total number of subset selection errors for the *Mod(1)* data (40 parameters total) (top panel) and the *Mod(2)* data (41 parameters total) (lower panel). The multi-level fit produce fewer selection errors in both cases. (c) The BIC of the single-level fit minus the BIC of the multi-level fit for *Mod(1)* data, before and after subset selection. After subset selection, the multi-level fit improves on the single-level fit, even when the single-level model is correct. (d) The BIC of the single-level fit minus the BIC of the multi-level fit for *Mod(2)* data, before and after subset selection. The multi-level fit improves on the single-level fit in almost all cases.

Table 1: Number of coefficients set to 0 by subset selection for the single-level fits with the three parameterizations.

Table 2: Number of coefficients set to 0 by subset selection for the single- and multi-level fits using parametrization  $W_{III}$ .

Table 3: Top panel: The selected number of clusters with the single-level and multi-level fits for the *Mod(1)* data. The correct  $K = 8$ . Both fitting strategies perform well, and the multi-level fit correctly identifies a single-level fit (bold face in table) in almost all cases. Lower panel: The selected number of clusters with the single- and multi-level fits for the *Mod(2)* data. The correct  $M = 9$ , with 7 1st level clusters ( $K = 7$ ). In almost all cases, the multi-level fit correctly identifies a sub-cluster structure (bold face in table), rather than single-level model.

Supplementary Table 1: Top 10 GO categories of all clusters, and clusters 1 and 2.

Supplementary Table 2: Top 10 GO categories for clusters 3 and 4.

Supplementary Table 3: Top 10 GO categories for clusters 5 and 6.

Supplementary Table 4: Top 10 GO categories for clusters 7, 8 and 9.

Supplementary Table 5: *Position weights matrices selected by the GLMpath algorithm.* 5-

fold cross-validation is employed to select the optimal regularization parameter in each of the logistic regression models.

## References

- Banfield, J. D. and Raftery, A. E. (1993). Model-based gaussian and non-gaussian clustering, *Biometrics* **48**: 803–821.
- Beissbarth, T. and Speed, T. P. (2004). Gostat: Find statistically overrepresented gene ontologies within a group of genes, *Bioinformatics* **20(9)**: 1464–1465.
- Fraley, C. and Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation, *Journal of the American Statistical Association* **97**: 611–631.
- Fraley, C. and Raftery, A. E. (2004). Bayesian regularization for normal mixture estimation and model-based clustering, *Technical Report 486*, Department of Statistics, University of Washington.
- Friedman, J. and Meulman, J. (2004). Clustering objects on subsets of attributes, *Journal of the Royal Statistical Society - Series B* **66(4)**: 815–849.
- 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.
- Hastie, T. and Tibshirani, R. (1996). Discriminant analysis by gaussian mixtures, *Journal of the Royal Statistical Society - Series B* **58**(158-176).
- Hoff, P. (2006). Model-based subspace clustering, *Bayesian Analysis* **1(2)**: 321–344.
- Jornsten, R. (2004). Clustering and classification based on the l1 data depth, *Journal of Multivariate Analysis* **90**: 67–89.
- Jornsten, R., Vardi, Y. and Zhang, C.-H. (2002). *Statistical data analysis based on the L1norm and related methods*, Editor: Y. Dodge, Statistics for industry and technology, Birkhauser, chapter "A Robust Clustering Method and Visualization Tool Based on Data Depth", pp. 353–366.

- Kaufman, L. and Rousseeuw, P. J. (1990). *Finding Groups in Data: An introduction to cluster analysis*, Wiley, New York.
- Law, M. H., Figueiredo, M. A. and Jain, A. K. (2004). Simultaneous feature selection and clustering using mixture models, *IEEE Pattern Analysis and Machine Intelligence* **26**(9): 1154–1166.
- Li, J. (2005). Clustering based on a multi-layer mixture model, *Journal of Computational and Graphical Statistics* **14**(3): 547–568.
- Meng, X. and Rubin, D. (1993). Maximum likelihood estimation via the ecm algorithm: A general framework, *Biometrika* **80**(2): 267–278.
- Raftery, A. and Dean, N. (2006). Variable selection for model-based clustering, *Journal of the American Statistical Association* **101**(473): 168–178.
- Smyth, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments, *Statistical Applications in Genetics and Molecular Biology* **3**(1): Article 3.
- Tadesse, M. G., Sha, N. and Vannucci, M. (2005). Bayesian variable selection in clustering high-dimensional data., *Journal of the American Statistical Association* **100**: 602–617.
- Yuan, M. and C.Kendziorski (2006). A unified approach for simultaneous gene clustering and differential expression identification, *Biometrics* **62**(4): 1089–1098.

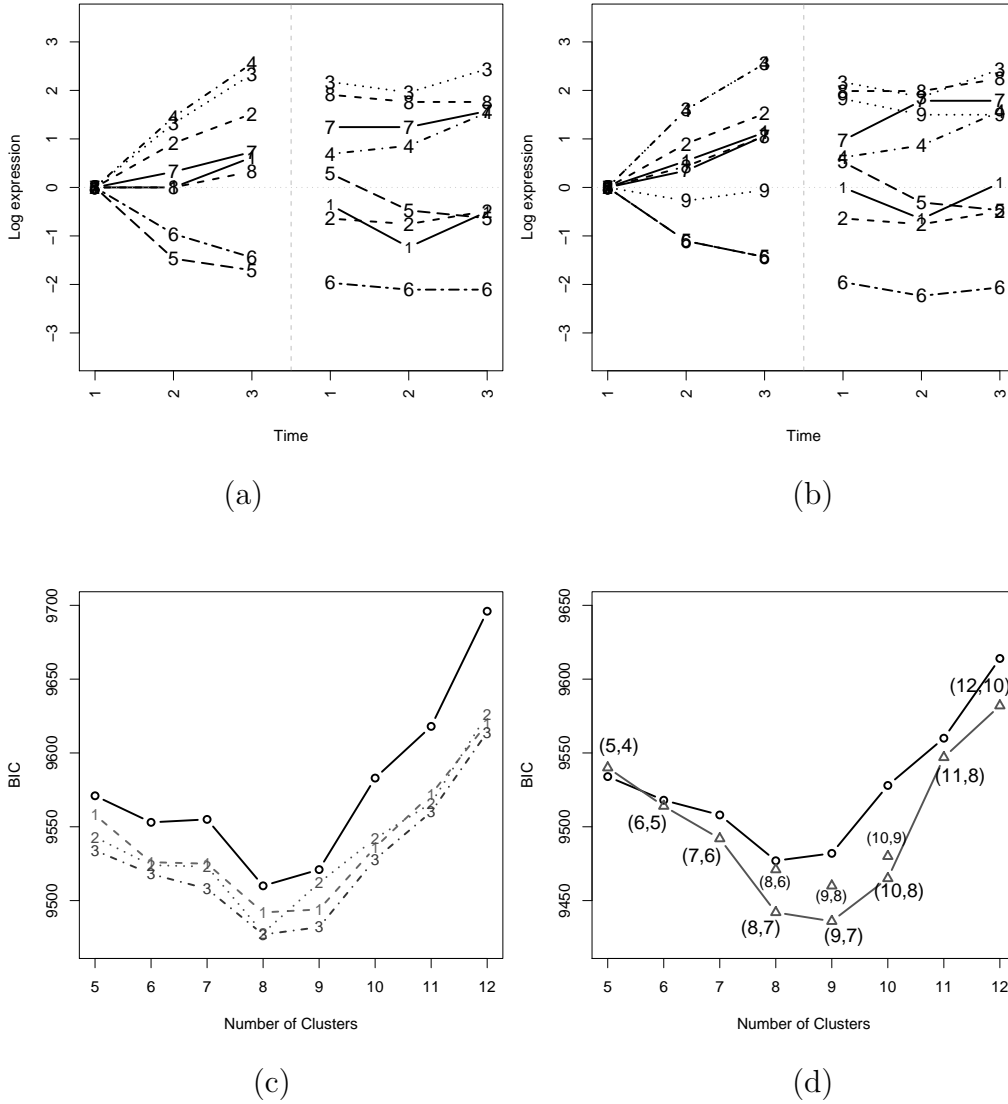


Figure 2: (a) Cluster mean profiles of the best single-level fit ( $K = 8$ ). The glial like population is depicted in the left panel, the neuron like in the right panel (parametrization  $W_{III}$ ) (b) Cluster mean profiles of the best multi-level fit  $K = 7, M = 9$ , with two sets of sub-clusters (parametrization  $W_{III}$ ). (c) The BIC curves obtained using the single-level fit. Solid line: no subset selection. Dashed and dotted curves annotated with the respective parametrization ( $W_I, W_{II}, W_{III}$ ). The  $W_{III}$ -BIC curve is the lowest, indicating that the  $W_{III}$  parametrization is the most efficient for this data set. (d) The BIC curves obtained from the single- and multi-level fits, using the  $W_{III}$  parametrization. The multi-level fit always gives a lower BIC for the same total number of clusters ( $M$ ). The numbers in the figures ( $M, K$ ) refers to the total number of clusters, and the number of 1st level clusters respectively. The best BIC values is obtained with  $M = 9$  clusters total, and  $K = 7$  1st level clusters.