Clustering with multiple distance metrics

- multi-level mixture models with profile transformations.

Rebecka Jörnsten

Department of Mathematical Statistics
Chalmers and Gothenburg University
jornsten@chalmers.se, http://www.stat.rutgers.edu/∼rebecka

September 16, 2009

Cluster analysis

General problem

- Popular approach for dimension reduction
- Wide range of applications: engineering, geological data, social networks, high-throughput biology

Thinking about the problems in terms of Model Selection

Clustering

1. How many clusters?
2. Can we find an efficient *within-cluster* parameterization? (E.g., are cluster profiles flat, or increasing?)
3. Can we find an efficient *between-cluster* parameterization? (Do cluster profiles share a similar shape, or coincide for a subset of data dimensions?)

Clustering gene expression data

Tasks

- To group genes by similarity of expression profiles across experimental conditions
- Assign functions to genes - 'guilt by association'

Goals

- We need an objective interpretation of cluster profiles - gene function inferences
- Avoid "Waste of parameters"

Within-cluster parameterizations

Mathematical Formulation

For each gene g we observe a feature vector $\mathbf{x}_g$:

$$\mathbf{x}_g \mid R_g = k \sim MVN(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k),$$

where R_g is the cluster membership indicator.

- We seek efficient/interpretable parameterizations of the cluster profiles; $\boldsymbol{\mu}_k = W\boldsymbol{\theta}_k$
- A sparse representation of $\boldsymbol{\mu}_k$ is obtained if we set some of $\boldsymbol{\theta}_k$ to 0.
- The design matrix W is chosen to represent a scientific question of interest.

Fitting the model: EM with a *modified M-step* incorporating the $\boldsymbol{\theta}_k$ constraints (see Jornsten and Keles, Biostatistics, 2008).

Model Selection

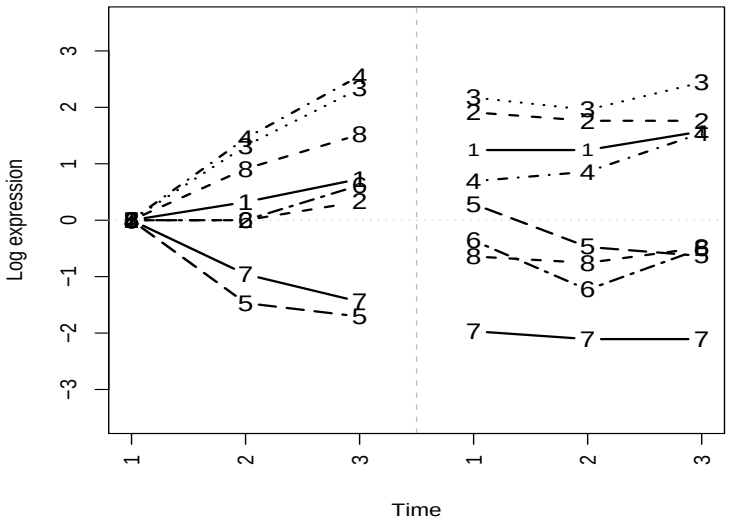
Searching for an optimal sparse within-cluster representation:

- Classification EM (CEM) approach, see Jornsten and Keles, 2008.
- Simultaneous selection via rate-distortion theory, see Jornsten, JCGS, 2009.

Multi-level Mixture models: Between-cluster parameterizations

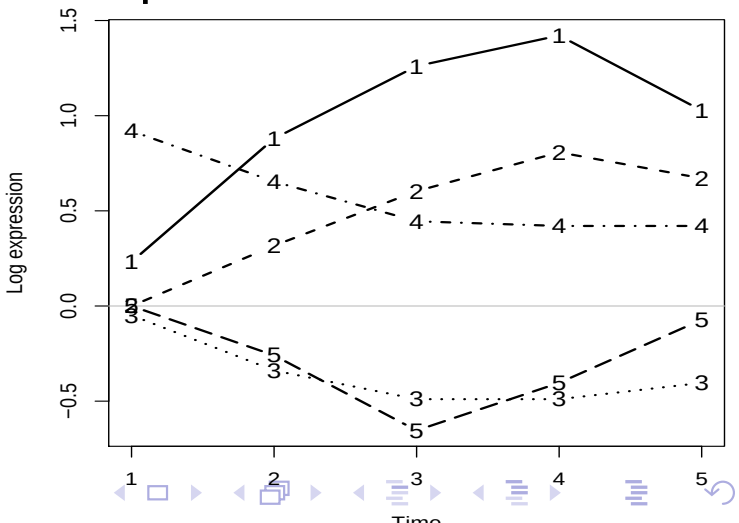
- Cell-line data
- There seem to be more neuron-specific clusters than glia specific
- $\mathcal{MIX}_{\mathcal{L}}$: Mixture models for multi-factor experiments

Example 1:



- mRNA expression time course after spinal cord injury
- Clusters seem to share a similar shape, and differ in terms of offset and scale.
- $\mathcal{MIX}_{\mathcal{T}}$: Mixture models with profile transformations.

Example 2:



$\mathcal{MIX}_{\mathcal{T}}$ - Model formulation

Multi-level structure

- For each gene g we observe a feature vector $\mathbf{x}_g$.
- Indicators $R_g = k, U_g = l$: cluster k and *sub-cluster* $l \in \{1, \cdots, L_k\}$.

$$\mathbf{x}_g \mid R_g = k, U_g = l \sim MVN(A_{kl}\boldsymbol{\mu}_k + b_{kl}, A_{kl}\boldsymbol{\Sigma}_k A'_{kl})$$

- *Within-cluster parameterization*: $\boldsymbol{\mu}_k = W\boldsymbol{\theta}_k$
- *Between-cluster parameterization*: affine transformation $(A_{kl}, b_{kl}) \rightarrow$ still Gaussian mixture.

Clustering with Multiple distance metrics

- Here, consider sign-flip and/or scale transformations:

$$\boldsymbol{\mu}_{kl} = T_{kl}(\boldsymbol{\mu}_k) = \beta_{kl}(\boldsymbol{\mu}_k + \alpha_{kl}), \quad \boldsymbol{\Sigma}_{kl} = \beta_{kl}^2 \boldsymbol{\Sigma}_k$$

- Equivalent to **clustering with three distance metrics simultaneously**: $1 - |correlation|$, $1 - correlation$, and *euclidean distance*!

Model fitting

- Fitting the model: EM with a *profile M-step* incorporating the constraints.
- Separate the tasks into
  1. The *top-level* estimation problem: $(\boldsymbol{\mu}_k = W_k \boldsymbol{\theta}_k, \boldsymbol{\Sigma}_k)$
  2. The *transformation parameter* estimation problem: $(\alpha_{kl}, \beta_{kl})$ (where $\beta_{k1} = 1, \alpha_{k1} = 0, \forall k$).
- How can we do this? Well, given $(\alpha_{kl}, \beta_{kl})$, maximizing the conditional likelihood translates to an application of the inverse profile transformation, and

$$\left(\frac{\mathbf{x}_g}{\beta_{kl}} - \alpha_{kl}\right) | R_g = k, U_g = l \sim MVN(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k).$$

Solve the aggregate problem first.

The aggregate problem

For each cluster k, and its sub-clusters $l = 1, \cdots, L_k$, compute

$$T_{kl}^{-1}(\mathbf{x}_g) = \frac{\mathbf{x}_g}{\beta_{kl}} - \alpha_{kl}$$

1. Update

$$\boldsymbol{\Sigma}_k = \frac{\sum_{g,l} \eta_{gkl} (T_{kl}^{-1}(\mathbf{x}_g) - \boldsymbol{\mu}_k)(T_{kl}^{-1}(\mathbf{x}_g) - \boldsymbol{\mu}_k)'}{\sum_{g,l} \eta_{gkl}}$$

2. Condition on $\boldsymbol{\Sigma}_k$, solve for the constrained $\boldsymbol{\theta}_k$:

$$\boldsymbol{\theta}_k = \frac{\sum_{g,l} \eta_{gkl} (W_k' \Sigma_k^{-1} W_k)^{-1} W_k' \boldsymbol{\Sigma}_k^{-1} \left(T_{kl}^{-1}(\mathbf{x}_g)\right)}{\sum_{g,l} \eta_{gkl}}$$

Note that $\mathbf{x}_g$ contributes to the top-level parameters in multiple ways, moderated by the sub-cluster membership probabilities.

Estimating the transform parameters

Condition on $(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, and consider

$$\max_{\alpha_{kl},\beta_{kl}} \sum_g \eta_{gkl} - \frac{1}{2}\frac{(\mathbf{x}_g - \beta_{kl}\boldsymbol{\mu}_k - \alpha'_{kl}1_J)'\Sigma_k^{-1}(\mathbf{x}_g - \beta_{kl}\boldsymbol{\mu}_k - \alpha'_{kl}1_J)}{\beta_{kl}^2} +$$

$$-\frac{J}{2}\log(\beta_{kl}^2) + c, \text{ where } \alpha'_{kl} = \beta_{kl}\alpha_{kl}, \ c \text{ is a constant.}$$

Solutions:

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

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

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

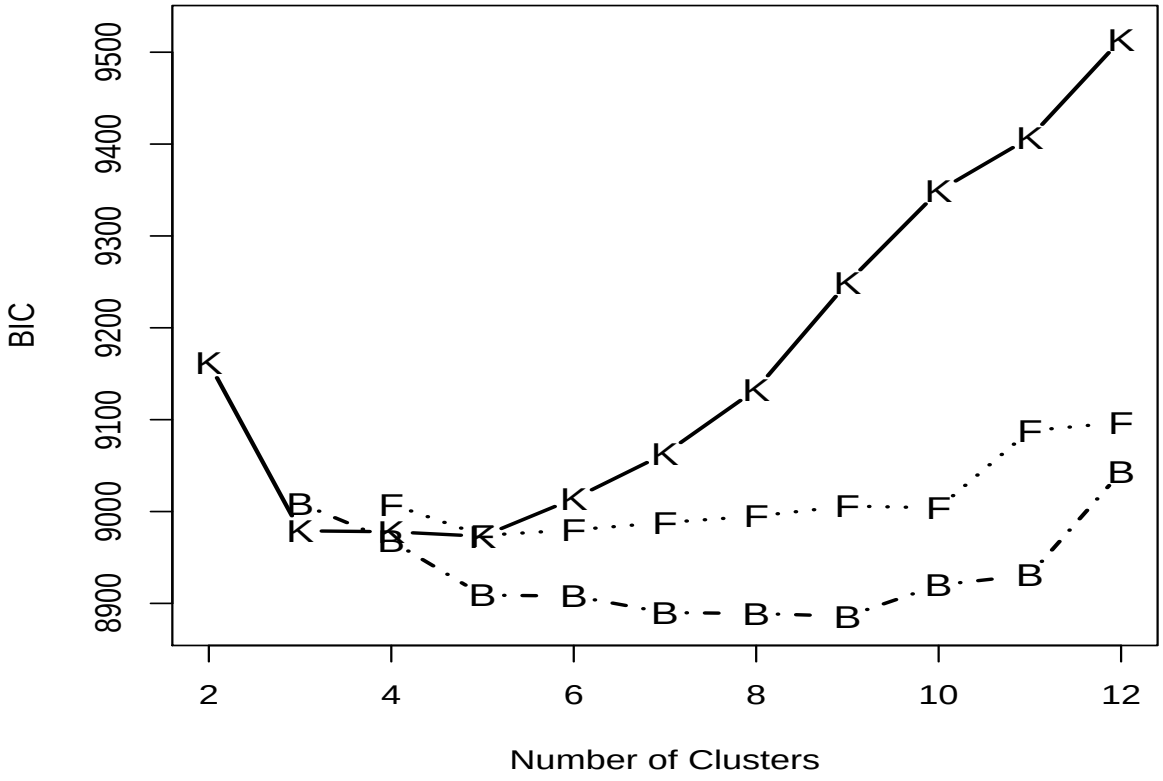
Making it work!

Setting up the model is the easy bit...

- Starting values: The search space $(K, \{L_k, k = 1, \cdots, K\})$ is huge ...
- We parameterize this as (M, K), where $M = \sum_k L_k$ is the total number of clusters.
  1. Initialize with M kmeans clusters
  2. Use PAM (robust kmeans) with $1 - correlation$ or $1 - |correlation|$ to group the M centroids into K top-level clusters.
  3. Add an additional random element by initializing $\mathbf{\Sigma}_k = |z| \times Cov(\tilde{\mathbf{x}}|R_g = k)$, where $z \sim N(r, s)$.
- Search forward for M, backwards for $K \in \{M, \cdots, 1\}$ to minimize BIC.
- More details in the papers (regularization, parameter value initialization,...).

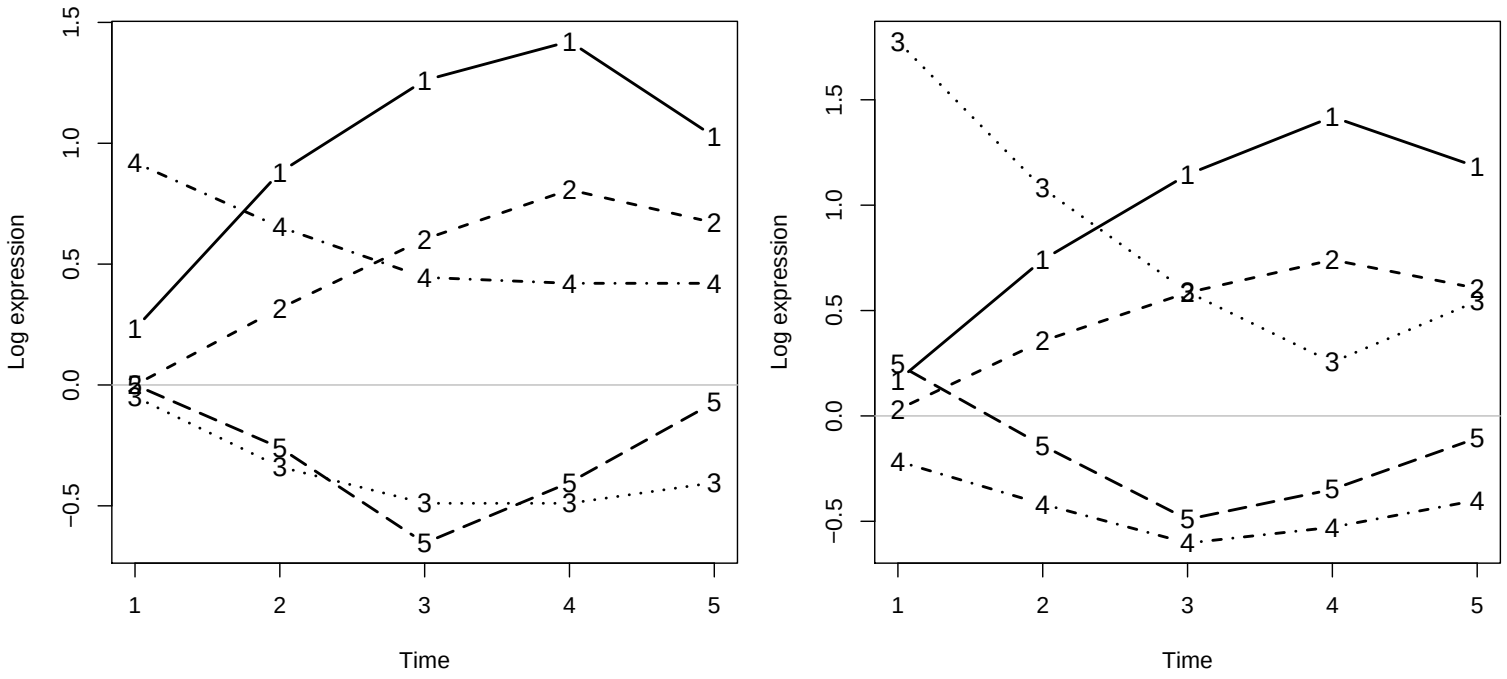
Clustering results

BIC as a function of the number of clusters, for the standard method (K), the sign-flip (F), and sign-flip+scale (B) models.



We select 9 clusters with an efficient *between-cluster* parameterization, compared to 5 using the standard mixture model.

Clustering results

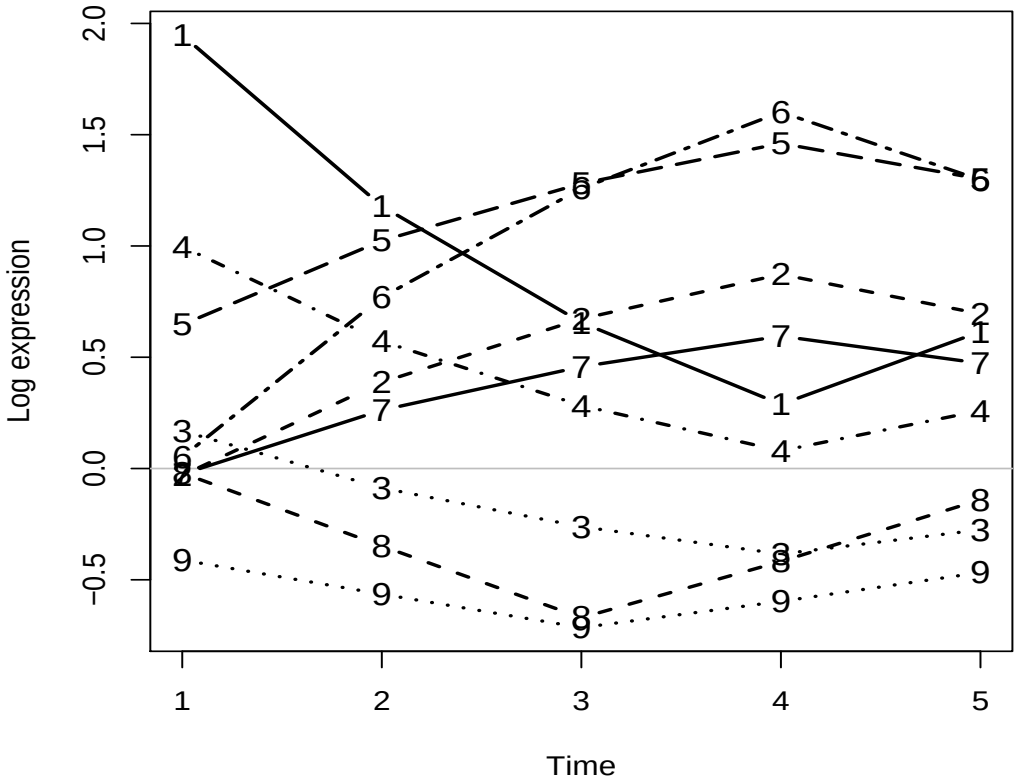


The $\mathcal{MIX}_{\mathcal{T}}$ with $M = 5$ has $K = 2$ profile clusters, with sub-clusters (1,2,3), where cluster 3 is a sign-flip cluster, and sub-clusters (4,5), both with positive β_{kl}.

Clustering results

The 9 cluster profiles

$k = 1$: (1,3,4) sign-flip of (2,5,6,7), $k = 2$: (8,9)



Clusters 7,2,5,6: increased level of response - distinct functional gene classes in each cluster.

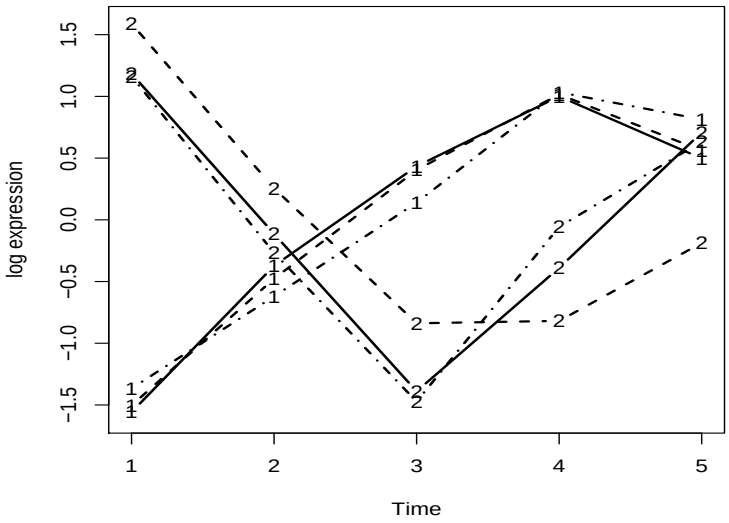
Clusters 1,3,4: wave of early immune response

Clusters 8,9: under-expression of neuron-specific genes, marker of neuron death after injury.

Clustering results

- $\mathcal{MIX}_{\mathcal{T}}$ allows for clustering via multiple distance metrics simultaneously
  - Sub-clusters: Mahalanobis distance
  - Top-level: $1 - cor$ or $1 - |cor|$.

- Compare cluster profiles:
  - PAM($1 - cor$) (dash)
  - PAM($1 - |cor|$) (dash-dot)
  - $\mathcal{MIX}_{\mathcal{T}}$ top-level profiles $\boldsymbol{\mu}_k, k = 1, 2$ (solid).



- With $\mathcal{MIX}_{\mathcal{T}}$ we can choose to investigate clusters at either level of the modeling hierarchy.

Multi-factor $\mathcal{MIX}_{\mathcal{T}}$

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

$$(\mathbf{x}_g, \mathbf{y}_g)|R_g = k, U_g = l \sim$$

$$\sim N\left((\beta_{kl}(\boldsymbol{\mu}_k^X + \alpha_{kl}1_J), \delta_{kl}(\boldsymbol{\mu}_k^Y + \gamma_{kl}1_J)), \boldsymbol{\Sigma}_{kl}\right),$$

where

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

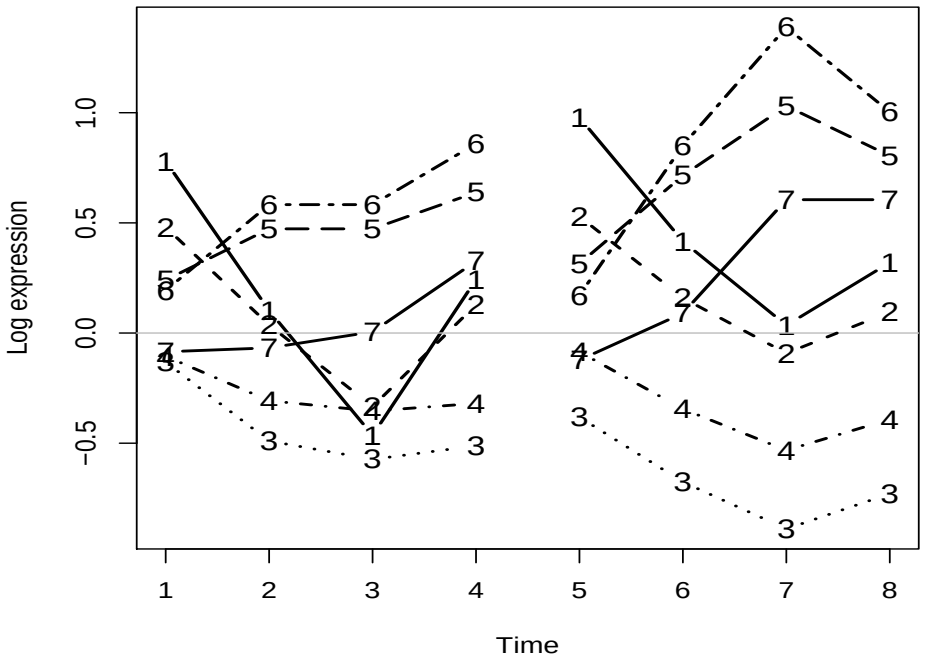
$\boldsymbol{\mu}_k^Y = \boldsymbol{\mu}_k^X + \Delta_k = W_k^X\boldsymbol{\theta}_k^X + \Delta_k$, where Δ_k is a vector of cluster contrast parameters between the two levels of injury. This parameterization provides a sparse representation of $\boldsymbol{\mu}_k$ if the cluster profiles $\boldsymbol{\mu}_k^X \simeq \boldsymbol{\mu}_k^Y$.

Multi-factor $\mathcal{MIX}_{\mathcal{T}}$

- 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.

Clustering results

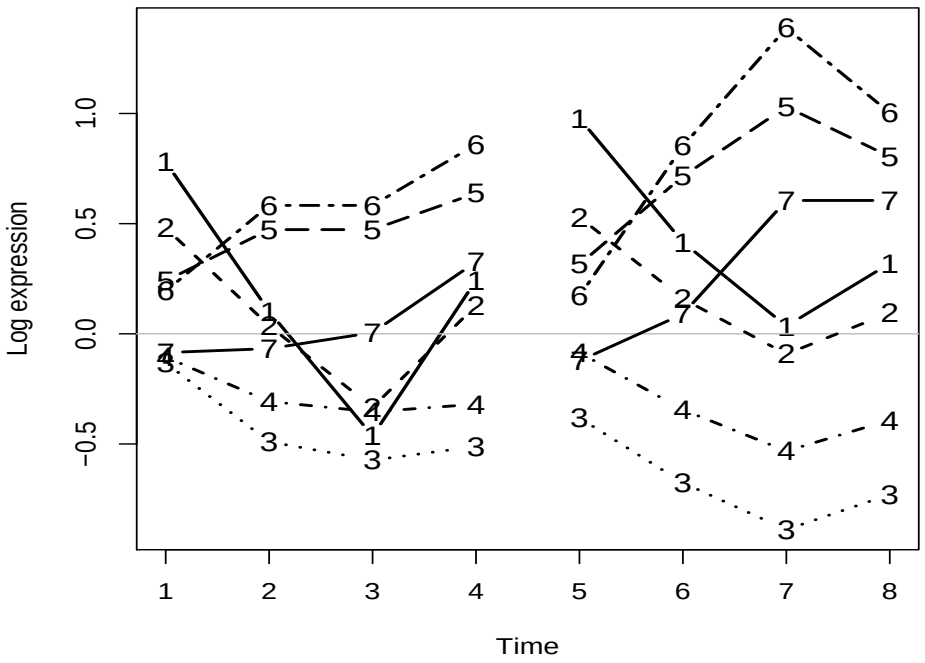
Cluster mean profiles. The left sub-panel: mild injury data. The right figure sub-panel: the moderate injury data.



The $\mathcal{MIX}_{\mathcal{T}}$ model with $M = 7$ clusters has $K = 4$ profile clusters. No sign-flip clusters are detected.

Clustering results

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 smaller than the sub-cluster separation for the moderate injury level data, $|\delta_{k1} - \delta_{k2}|$.

Conclusions

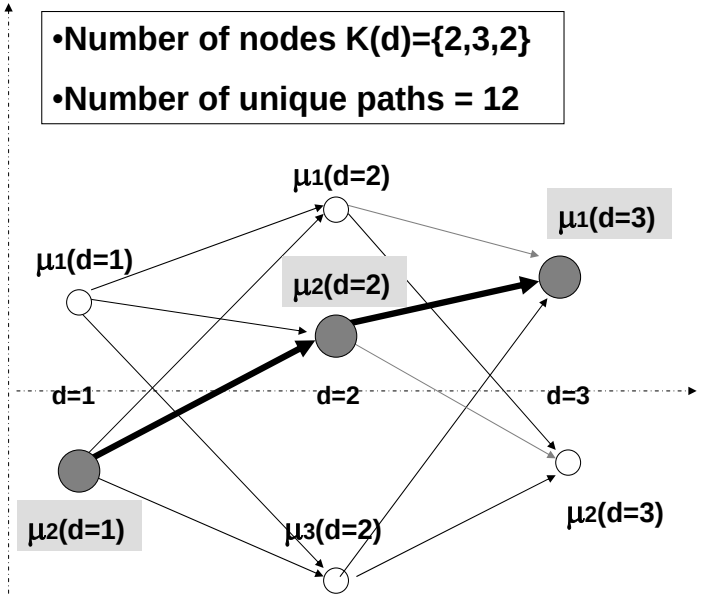
Multi-level mixture models

- Between-cluster parameterizations provide an objective inference tool for comparing clusters...
- ... and constitutes a substantial savings in terms of the number of model parameters.
- We detect clusters that a standard model-based clustering method may miss.
- Multi-level mixture models allow for clustering using multiple distance metrics simultaneously.

- Papers are available at `http://www.stat.rutgers.edu/~rebecka`

Future Work

- More complex experiments - need to automate subset selection as much as possible (e.g. consider different parameterizations simultaneously).
- Time course experiments: no multi-level structure is known a-priori. Define clusters by a unique path through a set of nodes. This allows clusters to coincide for a subset of data dimensions.



Path P_l (bold arrows) = {k(d),d=1,…,D}={2,2,1}
This path connects the shaded nodes.

Acknowledgements

- Prof. Ron Hart and Loyal Goff, W.M. Keck center for Collaborative Neuroscience, Rutgers
- Sunduz Keles, U. Wisconsin, Madison
- NSF, EPA