Clustering using Objective Functions and Stochastic Search

Jim Booth
University of Florida

April 2004

http://web.stat.ufl.edu/∼jbooth/

Joint work with George Casella and Jim Hobert

Background

- Basic problem is to partition n objects into K clusters
- Suppose there is an *objective function* that can be used to rank partitions
- K known - Sterling numbers (see e.g. Stanley, 1997)
- K unknown - Bell number, $B(n)$

$$B(40) = 1.6 \times 10^{35} \qquad B(100) = 4.8 \times 10^{115}$$

- Random partitions (see Pitman, 1997)

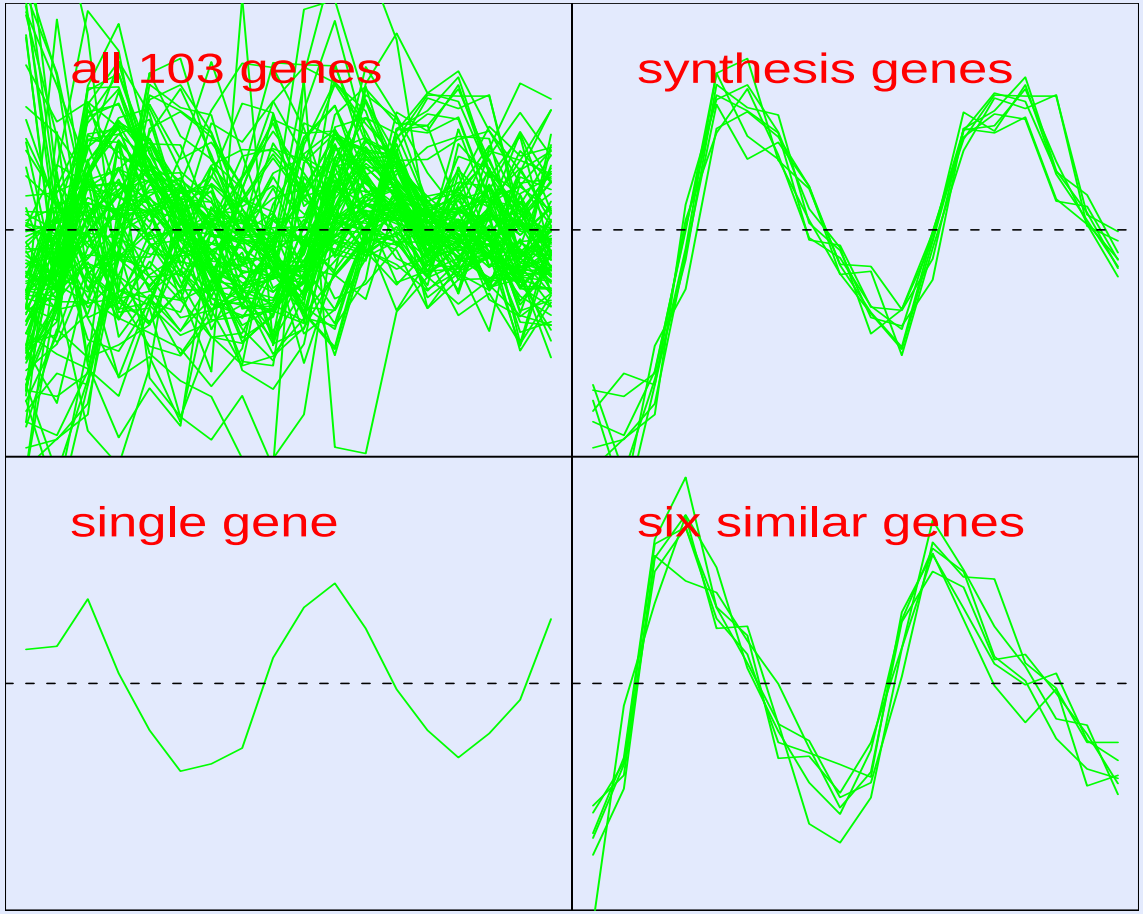
$$P(\text{at least one of top 1000 in } 10^9 \text{ draws} \,|\, n = 20) \approx .02$$

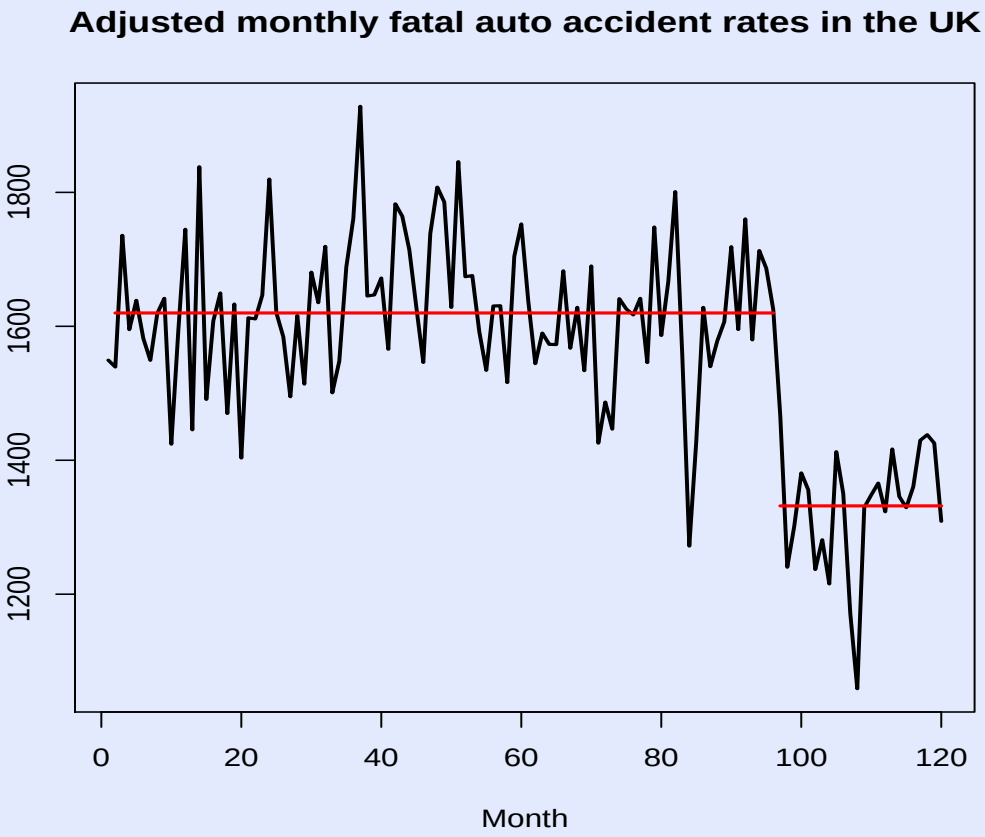
Summary

- Model-based clustering
  - Classification likelihood
  - Mixture model and EM algorithm
- Linear mixed model formulation
- Predictive likelihood
- Stochastic search
- Applications to gene expression profiles

Yeast Cell Cycle Data

- Alpha-factor microarray experiment (see Spellman, 1998)
- 18 cDNA arrays at 7 minute intervals
- Response is *log expression ratio* (synchronized/unsynchronized)
- Subset of 103 *goldstar* gene profiles (previously verified)


all 103 genes
synthesis genes
single gene
six similar genes


Adjusted monthly fatal auto accident rates in the UK
1200
1400
1600
1800
0
20
40
60
80
100
120
Month

A Statistical Model

- $K =$ number of sub-populations (clusters)
- $y_i =$ ith p-variate response (profile)
- $\gamma_i =$ cluster label for ith response
- Density of ith response,

$$f_{\gamma_i}(y_i) = f(y_i; \theta_{\gamma_i})$$

e.g. multivariate normal model

$$y_i = \mu_{\gamma_i} + \varepsilon_i\,, \qquad \varepsilon_i \sim N(0, \Sigma_{\gamma_i})$$

$$f(y; \mu_k, \Sigma_k) = |2\pi\Sigma_k|^{-1/2} \exp\left\{ -\frac{1}{2}(y-\mu_k)'\Sigma_k^{-1}(y-\mu_k) \right\}$$

Classification Likelihood

- Let $E_k = \{i|\gamma_i = k\}$ and $n_k = |E_k|$. Then

$$l_K(\theta,\gamma) = \log\prod_{i=1}^{n} f(y_i;\theta_{\gamma_i}) = \log\prod_{k=1}^{K}\prod_{i\in E_k} f(y_i;\theta_k)$$

- For the multivariate normal model, we have $\hat{\mu}_k = \bar{y}_k$, the sample mean profile in the ith cluster, and so

$$l_K(\hat{\mu},\Sigma,\gamma) = \text{const.} - \frac{1}{2}\sum_{k=1}^{K}\left\{\text{tr}(W_k\Sigma_k^{-1}) + n_k\log|\Sigma_k|\right\}$$

where

$$W_k = \sum_{i\in E_k}(y_i - \bar{y}_k)(y_i - \bar{y}_k)'$$

- Different assumptions about the structure of Σ_k lead to different ML clustering criteria (see Banfield and Raftery, 1993).
- If Σ_k is unconstrained, then $\hat{\Sigma}_k = W_k/n_k$, and

$$l_K(\hat{\mu}, \hat{\Sigma}, \gamma) = \text{const.} - \frac{1}{2}\sum_{k=1}^{K} n_k \log |W_k/n_k|$$

- If $\Sigma_k = \sigma^2 I$, then $\hat{\Sigma}_k = \text{tr}(W/n)I$, where $W = \sum_k W_k$, and

$$l_K(\hat{\mu}, \hat{\Sigma}, \gamma) = \text{const.} - \frac{1}{2} n \,\text{tr}(W/n)$$

- *Ward's method* - hierarchical agglomerative clustering - fails to find global maximum.
- Maximization of $l_K(\hat{\theta}, \hat{\gamma})$ with respect to K leads to n clusters - no penalty for complexity.

Mixture Model

- View the data as an i.i.d. random sample from a mixture of K sub-populations, with probabilities, $\tau_1, \ldots, \tau_K$
- We can attempt to maximize the likelihood

$$L(\theta, \tau) = \prod_{i=1}^{n} \sum_{k=1}^{K} \tau_k f(y_i; \theta_k)$$

- In this framework the cluster labels are random variables!
- Let $z_{ik} = I(\gamma_i = k)$, then

$$z_i = (z_{i1}, \ldots, z_{iK}) \sim \text{Multinomial}(1; \tau_1, \ldots, \tau_K)$$

- The mixture model formulation specifies the joint distribution of (y, z)

$$f(y, z; \theta, \tau) = \prod_{i=1}^{n} \prod_{k=1}^{K} \{\tau_k f(y_i; \theta_k)\}^{z_{ik}} = \prod_{k \in \mathcal{K}} \tau_k^{n_k} \prod_{i \in E_k} f(y_i; \theta_k)$$

- where $\mathcal{K} = \{k \,|\, z_{ik} = 1 \text{ for some } i\}$.
- The problem is now to *predict* z given the data y.
- The ML classification rule assigns the ith response to to the sub-population with the largest estimated posterior probability, $\hat{z}_{ik}$, $k = 1, \ldots, K$.

$$\hat{z}_{ik} \leftarrow \frac{\hat{\tau}_k f(y_i; \hat{\theta}_k)}{\sum_{j=1}^{K} \hat{\tau}_j f(y_i; \hat{\theta}_j)}$$

EM Algorithm

- Consider the linear model

$$y_i \,|\, z_{ik} = 1 \sim N(X\beta_k, \sigma_k^2 I)$$

where X is $p \times q$.

- E-step: compute posterior probabilities, $\hat{z}_{ik}$
- M-step: compute parameter estimates

$$\hat{\tau}_k \leftarrow \frac{n_k}{n} \qquad \hat{\beta}_k \leftarrow \frac{1}{n_k}(X'X)^{-1}X'\sum_{i=1}\hat{z}_{ik}y_i\,,$$

where $n_k = \sum_{i=1}^n \hat{z}_{ik}$, and

$$\hat{\sigma}_k^2 \leftarrow \frac{1}{n_k p}\sum_{i=1}^n \hat{z}_{ik}(y_i - X\hat{\beta}_k)'(y_i - X\hat{\beta}_k)$$

Issues

- Multimodality is a problem!
- Maximization is for fixed K!
- BIC for model selection (see Fraley and Raftery, 2002)

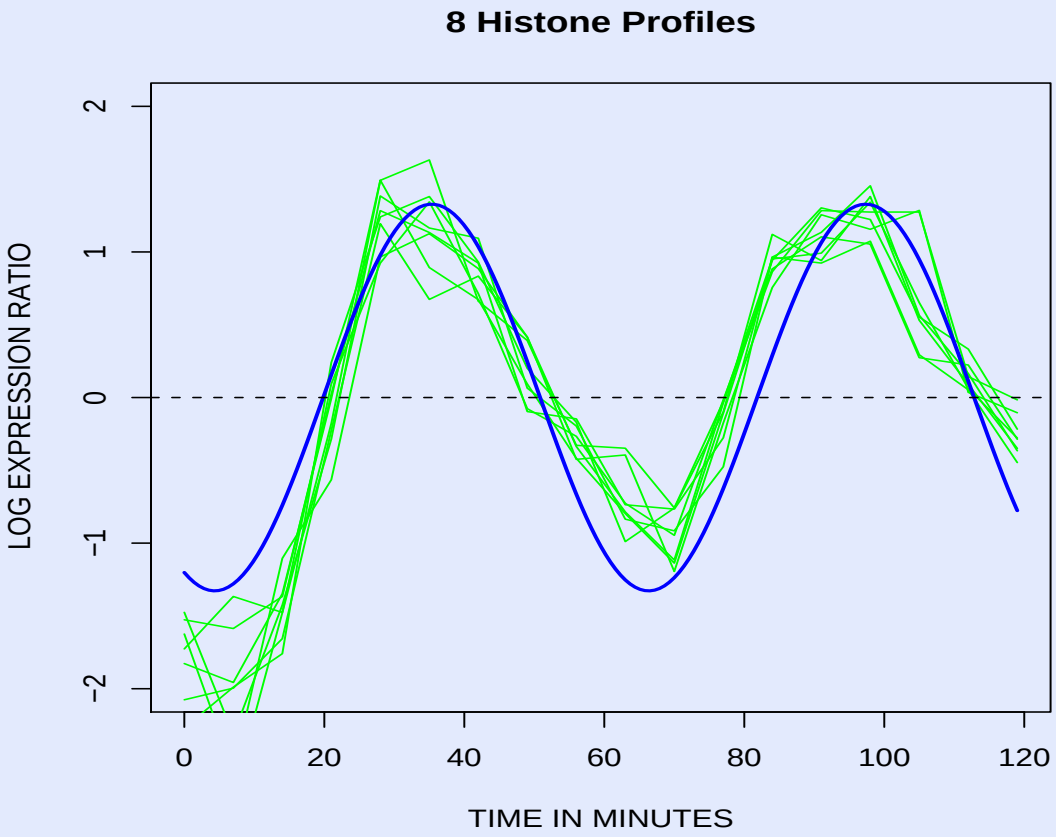
Cell Cycle Data

- Spellman et al, 1998 proposed a first-order Fourier series model:

$$y_{ij} = \beta_0 + \beta_1 \cos\left(\frac{2\pi t_j}{w}\right) + \beta_2 \sin\left(\frac{2\pi t_j}{w}\right) + \varepsilon_{ij}$$

  where w is the period of the cell cycle, and the t_j's are the observation times.

- Suppose that the coefficients and error variance differ among clusters.


8 Histone Profiles
LOG EXPRESSION RATIO
2
1
0
−1
−2
0
20
40
60
80
100
120
TIME IN MINUTES

Linear Mixed Model Formulation

- Stage 1: Select labels, γ_i as an iid sample from a probability distribution on N^+

$$P(\gamma_i = k) = \tau_k, \quad k = 1, 2, \ldots$$

  Let $n_k = \#\{i : \gamma_i = k\}$.

- Stage 2: Suppose that, conditional on the partition, the joint distribution of the responses from cluster is described by a *linear mixed model*:

$$Y_k \sim (1 \otimes X)\beta_k + (1 \otimes Z)b_k + \varepsilon_k$$

  where Y_k is the $n_k p \times 1$ stacked vector of responses, X is $p \times q$, Z is $p \times r$, and

$$\varepsilon_k \sim N(0, \sigma_k^2 I_{n_k p}) \qquad b_k \sim N(0, \nu\sigma_k^2 I_r)$$

Comments

- This model implies *observations in the same cluster are dependent*!
- Conditional on the partition, the mixed model formulation implies,

$$Y_k \sim N\left[(1 \otimes X)\beta_k, \sigma_k^2 A_k\right] ,$$

where

$$A_k = I_{n_k p} + \nu J_{n_k} \otimes ZZ'$$

Estimation for a given partition

- If ν is known, ML estimates of β_k and σ_k^2, for a given partition, are given by

$$\hat{\beta}_k = (X^T B_k^{-1} X)^{-1} X^T B_k^{-1} \bar{y}_k \,,$$

where $B_k = I_p + n_k \nu Z Z^T$, and

$$\hat{\sigma}_k^2 = \frac{1}{n_k p} \left[\sum_{i \in E_k} ||y_i - \bar{y}_k||^2 + n_k (\bar{y}_k - X\hat{\beta}_k)^T B_k^{-1} (\bar{y}_k - X\hat{\beta}_k) \right] .$$

Predictive likelihood

- The conditional distribution, $z|y$, is proportional to the joint distribution of (y,z)
- Model specifies the joint distribution of (y,z), but depends on parameters $\{\tau_k, \beta_k, \sigma_k^2 : k \in \mathcal{K}\}$
- Sufficient statistics are $\{n_k, \hat{\beta}_k, \hat{\sigma}_k^2 : k \in \mathcal{K}\}$
- A *predictive likelihood* for $z|y$ is proportional to the conditional distribution of (y,z) given the sufficient statistics. This results in

$$p(z|y) \propto \prod_{k \in \mathcal{K}} \frac{n_k! \Gamma\left[(n_k p - q)/2\right] |X^T B_k^{-1} X|^{-1/2} \pi^{q/2}}{n_k^{n_k p/2} p^{(n_k p - q)/2} (\hat{\sigma}_k^2)^{(n_k p - q)/2 - 1} \prod_{j=1}^{r} (1 + n_k \nu \lambda_i)^{1/2}}$$

- Butler (1986), Bjornstad (1990)

Stochastic Search

- Markov chain on the partition space with stationary distribution proportional to the predictive likelihood.
- Candidate moves chosen by *biased random walk*:
  - Pick an object at random
  - If it's a singleton, move it to another cluster at random
  - If it's not a singleton, move it to another cluster or to a singleton cluster at random
- *Symmetric transition kernel*
- Metropolis-Hastings step: candidate move from z to z' accepted with probability

$$\min\left[1, \frac{p(z'|y)}{p(z|y)}\right]$$

- Slow. Gets stuck in local modes!

Tempered Transitions (Neal, 1996)

- Consider a collection of probability distributions on the partition space
$$\pi_i(z) \propto p(z|y)^{1/\alpha_i}, \quad i = 0, 1, \ldots, r$$
e.g. $\alpha_i = 2^i$.

- For each i, let $T_i(z'|z)$ denote a Markov transition kernel that is reversible with respect to π_i; e.g. the MH algorithm described above.

- A candidate tempered transition is obtained by applying these base transitions in the sequence:
$$T_1 \cdots T_r T_r \cdots T_1$$

- The candidate is accepted with probability
$$\min\left[1, \frac{\pi_1(\hat{z}_0)}{\pi_0(\hat{z}_0)} \cdots \frac{\pi_r(\hat{z}_{r-1})}{\pi_{r-1}(\hat{z}_{r-1})} \frac{\pi_{r-1}(\check{z}_{r-1})}{\pi_r(\check{z}_{r-1})} \cdots \frac{\pi_0(\check{z}_0)}{\pi_1(\check{z}_0)}\right]$$

Shrinkage Estimation of Mean Profile

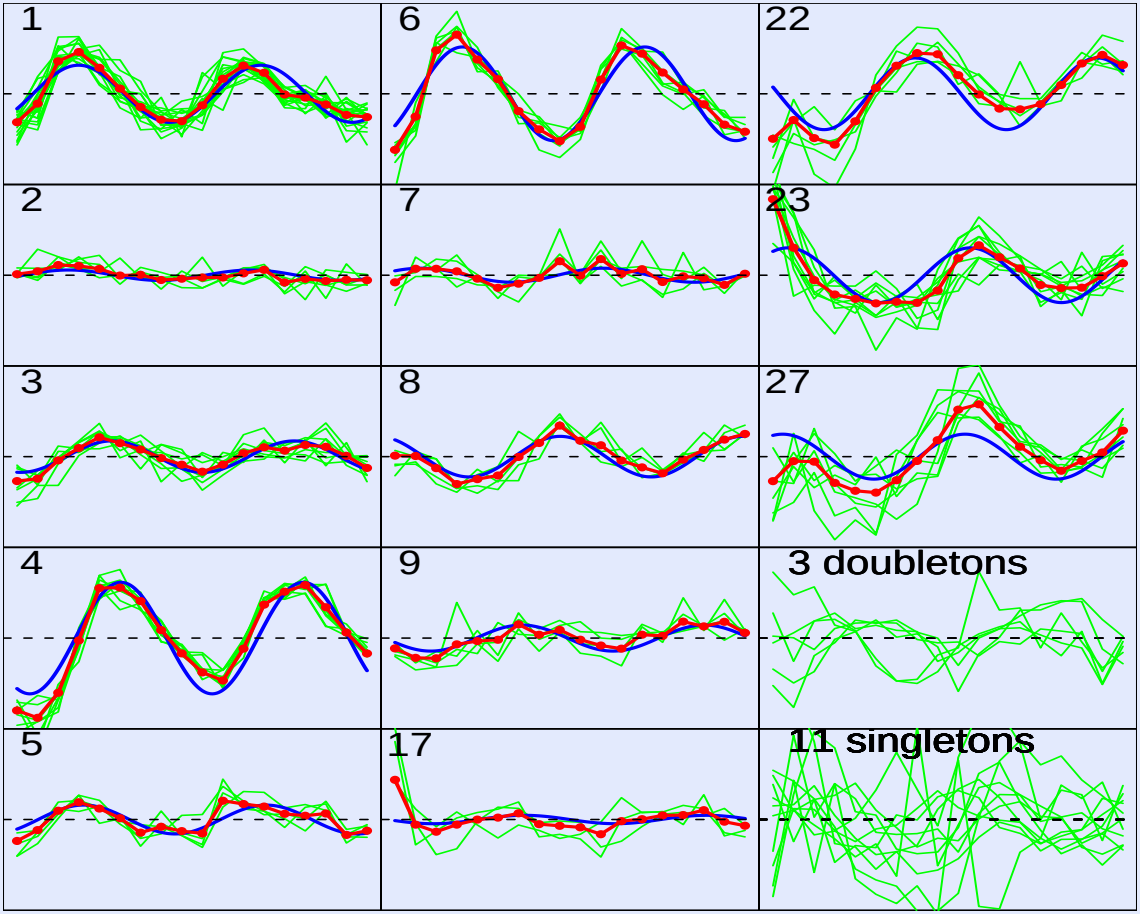
- The BLUP for b_k is

$$\hat{b}_k = n_k \nu (I_r + n_k \nu Z'Z)^{-1} Z'(\bar{y}_k - X\hat{\beta}_k)$$

- If $Z = I$, $B_k = (1 + n_k \nu) I_p$, and the predicted mean profile is a *convex combination of the sample mean* (unstructured model), *and the fitted values* from the linear model:

$$X\hat{\beta}_k + \hat{b}_k = \frac{n_k \nu}{1 + n_k \nu} \bar{y}_k + \frac{1}{1 + n_k \nu} X\hat{\beta}_k$$

- Best partition with clusters ordered by contribution to predictive likelihood ...


1
6
22
2
7
23
3
8
27
4
9
3 doubletons
5
17
11 singletons

Penalized Splines as BLUPs

- (Quadratic) penalized splines fits can be formulated as BLUPs arising from a linear mixed model of the form

$$Y_j = \beta_0 + \beta_1 x_j + \beta_2 x_j^2 + \sum_{l=1}^{L} b_l z_{jl} + \varepsilon_j$$

where

$$z_{jl} = (x_j - \xi_l)_+^2, \quad l = 1, \ldots, L$$

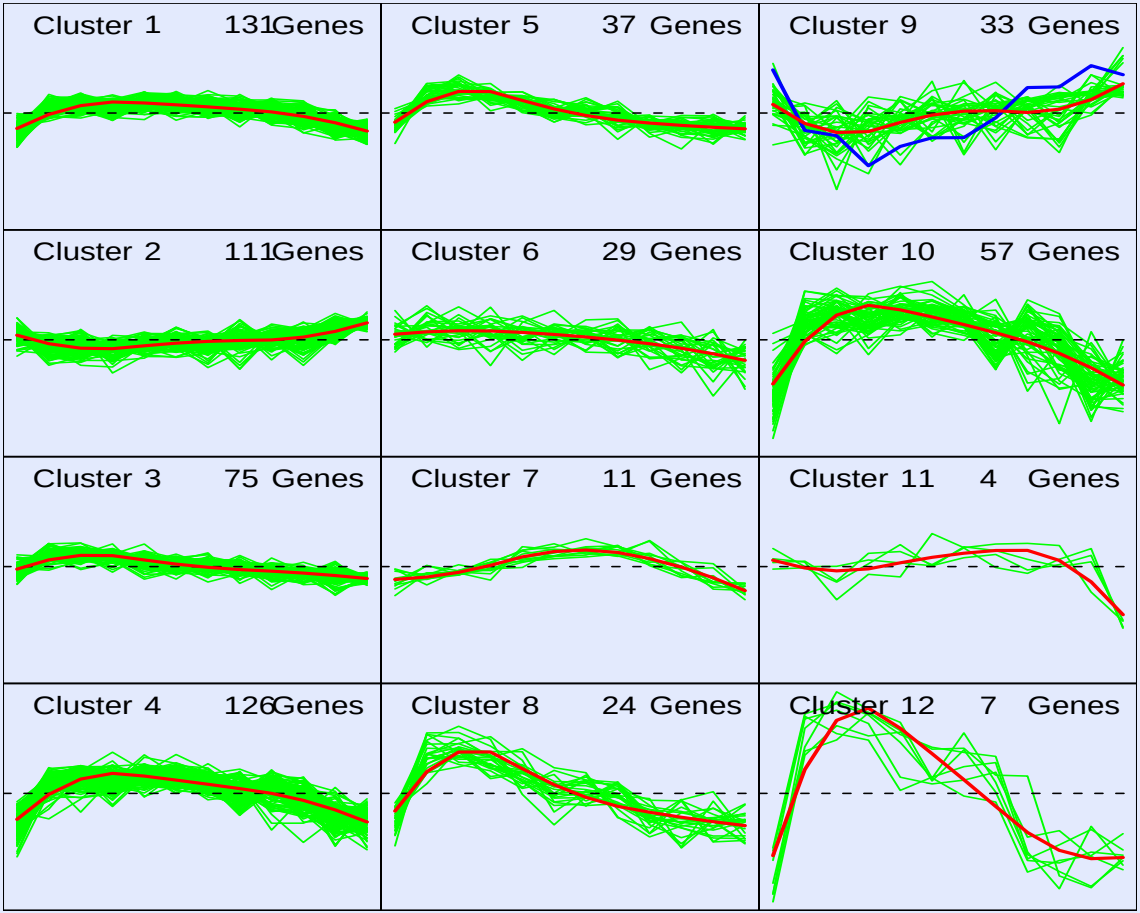
- Ruppert, Wand and Carroll, 2003

Corneal Wound Healing

- 646 gene expression profiles at days 0, 1, 2, 3, 4, 5, 6, 7, 14, 21, 42, and 98.
- The day 0 sample was taken prior to treatment
- Transformed time-scale

$$t_j = (j - 5.5), \quad j = 0, 1, \ldots, 11$$

- Two knots, $\xi_1 = -2.0$ and $\xi_2 = 2.0$


Cluster 1 131 Genes
Cluster 5 37 Genes
Cluster 9 33 Genes
Cluster 2 111 Genes
Cluster 6 29 Genes
Cluster 10 57 Genes
Cluster 3 75 Genes
Cluster 7 11 Genes
Cluster 11 4 Genes
Cluster 4 126 Genes
Cluster 8 24 Genes
Cluster 12 7 Genes

THANKS?