Biostatistics Seminar
May 2005, Harvard

Clustering Using Objective Functions and Stochastic Search

George Casella
University of Florida

Jim Hobert
University of Florida

Jim Booth
Cornell University

casella@stat.ufl.edu

Overview

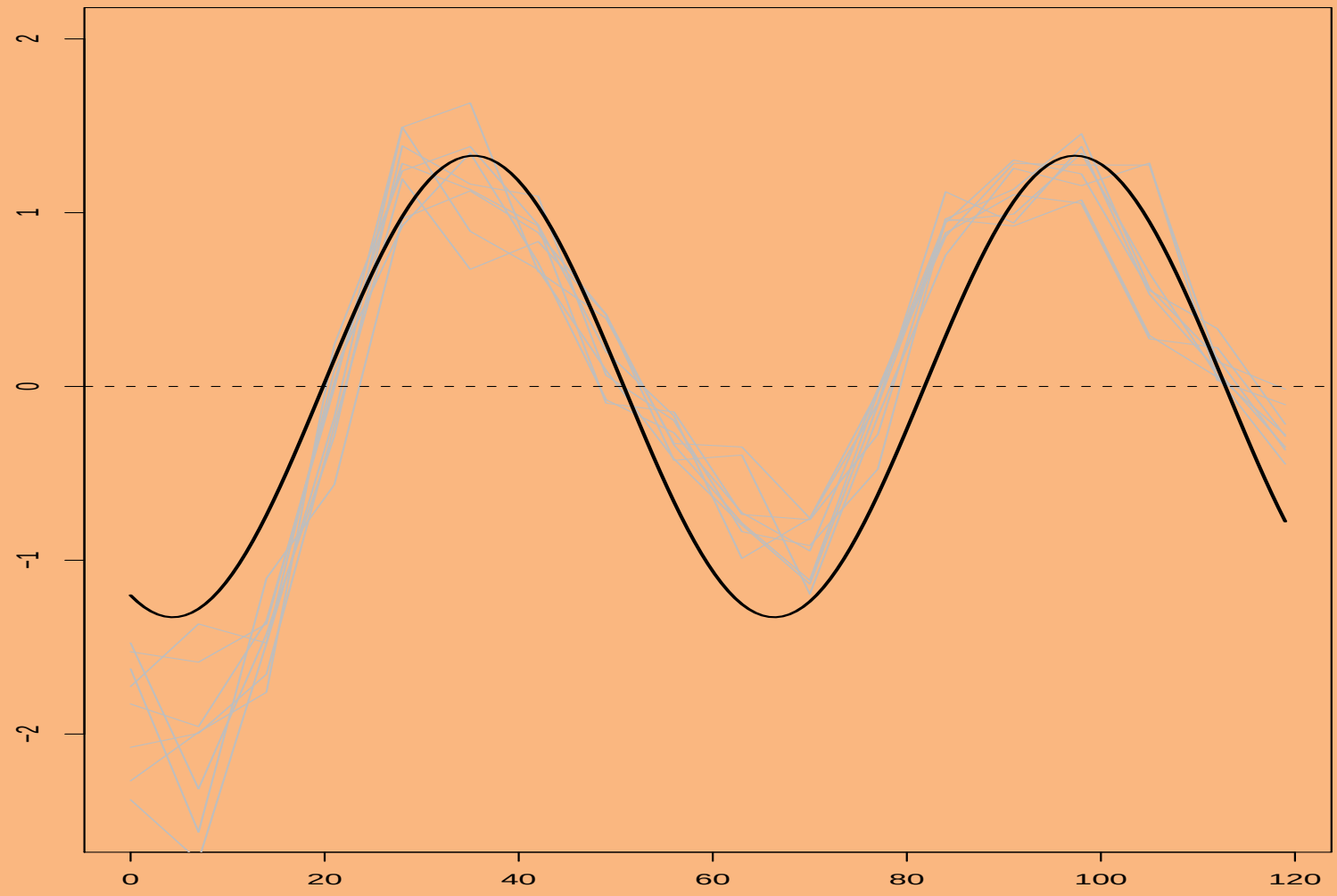
- Introduction - Models for Clustering
- Objective Functions
- Stochastic Search
- Example - Yeast Cell Cycle (Parametric Base Model)
- Example Corneal Scarring (Nonparametric Base Model)
- Conclusions and Other Stuff

1. Introduction

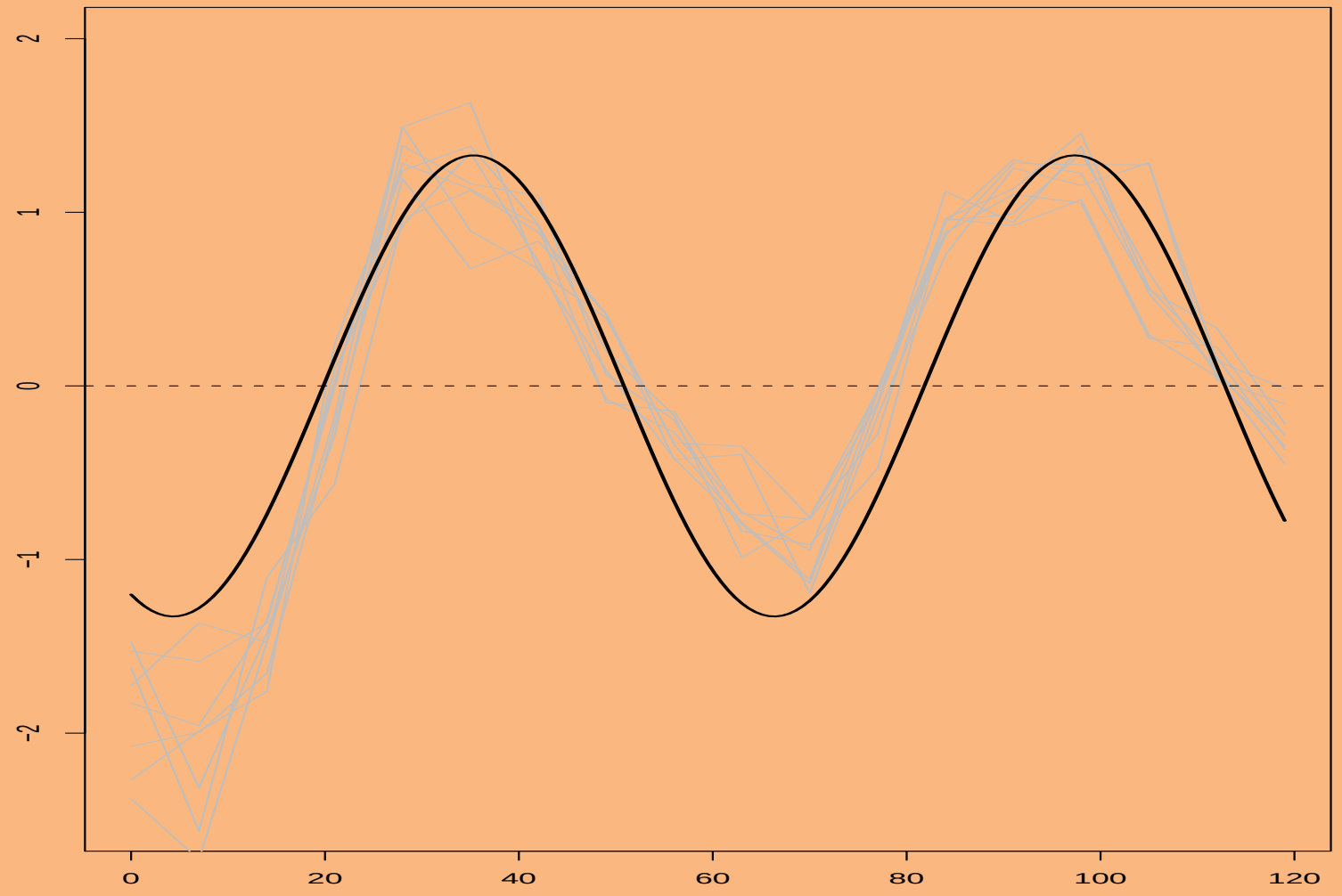
- Clustering and Classification are very popular
- Standard algorithms
  - Cluster vectors of observations
  - No structure assumed
- Example: Yeast Cell Cycle
  - Time course gene expressions
  - Goal: Cluster similar profiles
  - Fourier Analysis Suggested

▷ Gene expression profiles for the eight histones.

▷ Solid line = first-order Fourier series model



- ▷ Model OK in identifying phase of peak expression.
- ▷ Substantial lack-of-fit to profiles



- Given n distinguishable objects
  - Divide objects into groups
  - Within groups similar
  - Between groups different
- $\mathbb{N}_n := \{1, 2, \ldots, n\}$ are the objects
- Clustering is a partition $\mathbb{P}_n$ of $\mathbb{N}_n$
- Objective function $\pi : \mathbb{P}_n \to \mathbb{R}^+$

- Cluster n p-dimensional vectors $y_i = (y_{i1}, \ldots, y_{ip})^T$, $i = 1, 2, \ldots, n$.

- Many authors assume

$$y_i \sim \sum_{k=1}^{K} \tau_k \, f(y; \theta_k) \; , \quad \text{iid}$$

  - K is fixed
  - $\sum_{k=1}^{K} \tau_k = 1$
  - $\{ f(\cdot\,; \theta) : \theta \in \Theta \}$ is a parametric family of densities

- McLachlan and co-authors, Titterington *et al.* 1985

Previous Model-Based Clustering -2

- $\mathbb{P}_n$ is obtained as a byproduct of EM
  - EM finds MLEs of $\left\{(\tau_k)_{k=1}^K, (\theta_k)_{k=1}^K\right\}$
  - Missing data $= W_i =$ cluster to which y_i belongs
  - Final iteration of EM: Compute E(W_i|MLEs)

- Fast and convenient, but...
  - Statistical Procedure and Computational Algorithm Fused Unnaturally
  - The EM algorithm Itself is part of the inference
- Knowledge of MLEs insufficient to construct clusters
- Indeed, knowledge of Parameters insufficient
  - One run of EM STILL needed to assign clusters

An Alternate Model

- Mixture Model often unrealistic for clustering
- A more realistic assumption
  - There is a fixed, unknown partition of $\mathbb{N}_n$, ω
  - It has $c = c(\omega)$ clusters, $\mathcal{C}_1, \ldots, \mathcal{C}_c$
  - The data have density

$$f(y|\theta, \omega) = \prod_{k=1}^{c(\omega)} \prod_{j \in \mathcal{C}_k} f(y_j|\theta_k) \ .$$

An Alternate Model-2

- For the density

$$f(y|\theta,\omega) = \prod_{k=1}^{c(\omega)} \prod_{j\in\mathcal{C}_k} f(y_j|\theta_k) \ .$$

  - $\cup_{k=1}^{c(\omega)}\mathcal{C}_k = \mathbb{N}_n$ and $\mathcal{C}_i \cap \mathcal{C}_j = \emptyset$ whenever $i \neq j$.
  - The parameter ω is directly relevant to the clustering problem.

2. Objective Functions

- The data vector, Y_i, measured on the ith object, consists of r replicate profiles;

$$Y_i = (Y_{i1}^T, \ldots, Y_{ir}^T)^T \ ,$$

for $i = 1, \ldots, n$.

  - Y_{i1}^T is a vector of length p
  - Can model microarray experiments with replicate measurements on each gene.

A Particular Model for Clustering

- For objects in the kth cluster

$$Y_{ij} = X\beta_k + Z_1 U_i + Z_2 V_k + \varepsilon_{ij}\,,$$

  - $i = 1, \ldots, n_k$, $j = 1, \ldots, r$,
  - $\varepsilon_{ij} \sim \mathrm{N}_p(0, \sigma_k^2 I_p)$,
  - $U_i \sim \mathrm{N}_{s_1}(0, \lambda_1 \sigma_k^2 I_{s_1})$ Subject Effect
  - $V_k \sim \mathrm{N}_{s_2}(0, \lambda_2 \sigma_k^2 I_{s_2})$ Cluster Effect
  - all mutually independent.

A Particular Model for Clustering -

- $$Y_{ij} = X\beta_k + Z_1 U_i + Z_2 V_k + \varepsilon_{ij}\,,$$
  - ▷ Standard Priors
  - ▷ Default Prior on β and σ
  $$\pi(\beta, \sigma^2|\omega) \propto \prod_{k=1}^{c(\omega)} (1/\sigma_k^2)^{\alpha+1}$$
  - ▷ All Priors conditional on ω
- $\pi(\omega) \propto \prod_{k=1}^{c(\omega)} n_k!$ (Crowley 1997 + heuristics)

An Objective Function

- For Clustering, all parameters are nuisance parameters
- Except ω, the cluster!
- Distribution of interest is $\pi(\omega|\mathbf{y})$ - Objective Function
  - Marginal Posterior
  - Integrate out all other parameters

An Objective Function - 2

- After all the integrations and *fun with matrix algebra*

$$\pi(\omega|y) \propto \frac{1}{(\hat{\sigma}^2)^{(nrp-q)/2+\alpha}} \prod_{k=1}^{c(\omega)} \left(\frac{n_k!}{|\mathrm{Var}(Y_k)|^{1/2} |\mathrm{Var}(\hat{\beta}_k)|^{1/2}} \right)$$

where

$$\hat{\sigma}^2 = \frac{1}{nrp} \left\{ \sum_{i=1}^{n} (Y_i - 1_r \otimes \bar{Y}_k)^T A^{-1} (Y_i - 1_r \otimes \bar{Y}_k) + \sum_{k=1}^{c(\omega)} n_k r (\bar{Y}_k - X\hat{\beta}_k)^T W_k (\bar{Y}_k - X\hat{\beta}_k) \right\} .$$

An Objective Function - 3

- Note the two pieces of $\hat{\sigma}^2$

$$\hat{\sigma}^2 = \frac{1}{nrp}\left\{\sum_{i=1}^{n}(Y_i - 1_r \otimes \bar{Y}_k)^T A^{-1}(Y_i - 1_r \otimes \bar{Y}_k) + \sum_{k=1}^{c(\omega)} n_k r(\bar{Y}_k - X\hat{\beta}_k)^T W_k(\bar{Y}_k - X\hat{\beta}_k)\right\} .$$

- The first piece measures deviation from the cluster mean
- The second piece measures deviation from the base model

3. Stochastic Search

- Objective function,

$$\pi : \mathbb{P}_n \to \mathbb{R}^+$$

  measures the goodness of partitions

- Maximizing Objective function $\Rightarrow$ Potentially difficult

- Impossible by brute force enumeration

- Amenable to MCMC

- Success of the search requires that the Markov chain
  - Has stationary distribution putting high mass on good partitions
  - Moves freely between isolated modes
- Can do this with Metropolis-Hastings
  - Stationary distribution $\propto \pi$
  - Obtain this by "correcting" a candidate

- Neighborhood structure on $\mathbb{P}_n$
  - ω_i and ω_j are neighbors if and only if they share an edge
  - edge shared $\Leftrightarrow$ $\omega_i \rightarrow \omega_j$ by moving one object
- Example, $n = 3$ has partitions

$$\omega_1 : \{1,2,3\} \;\; \omega_2 : \{1,2\}\{3\} \;\; \omega_3 : \{1,3\}\{2\}$$

$$\omega_4 : \{2,3\}\{1\} \;\; \omega_5 : \{1\}\{2\}\{3\}.$$

Only ω_1 and ω_5 do not share an edge.

Candidate Distribution - Random Walk - 2

- $d(\omega)$ = number of neighbors
- Nearest neighbor random walk
- $\omega' \to \omega$
  - with probability $1/d(\omega')$ if ω and ω' are neighbors
  - zero otherwise.
- Problem:

  $Prob(\omega_1 \to \omega_2) = 1/3,$ but $Prob(\omega_2 \to \omega_1) = 1/4$

  - Random walk not symmetric
  - Need to calculate $d(\omega)$ to sample neighbors uniformly
    - Really hard

We fixed it - Biased Random Walk

- Let c = number of clusters in current state
  - $c = 1 \Rightarrow$ choose one object uniformly and move it to its own cluster
  - $c > 1 \Rightarrow$ choose one object uniformly
    - if a singleton, randomly assign it to one of the $c-1$ clusters
    - if not, move to a cluster or make a singleton, each with probability $1/c$

- Symmetric Transition Matrix

$$Prob(\omega_1 \rightarrow \omega_2) = Prob(\omega_2 \rightarrow \omega_1) = 1/3$$

- In the MH algorithm, the acceptance probability is simply
$$\min\{1, \pi(\omega)/\pi(\omega')\}.$$
- Running this algorithm does not require finding or counting the neighbors of ω and ω'.
- This alternative gives an MH algorithm that is
  - ▷ Much easier to program
  - ▷ Faster in the sense of more iterations per unit time.

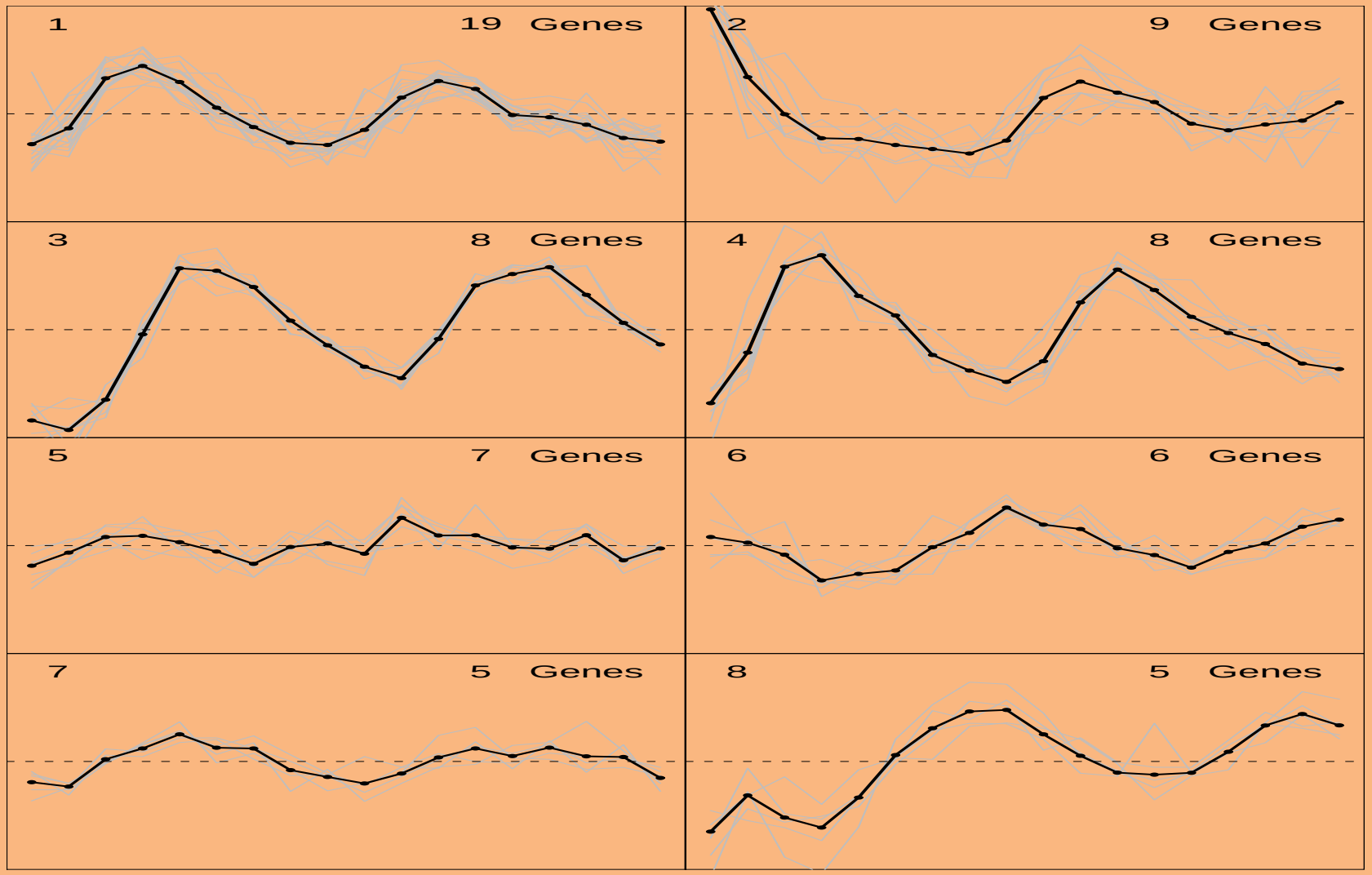
4. Examples - Yeast Cell Cycle

- Profiles of 104 identified cell cycle-regulated genes
  - Log expression ratios
  - 18 cDNA microarrays
  - 7 minute intervals
- Model: For objects in the kth cluster

$$Y_{ij} = X\beta_k + V_k + \varepsilon_{ij}\,,$$

  - Base model = first-order Fourier series

- Clustering of the Yeast Cell Cycle Data
  - The solid black = best linear unbiased predictor, Cluster 3 are the histones



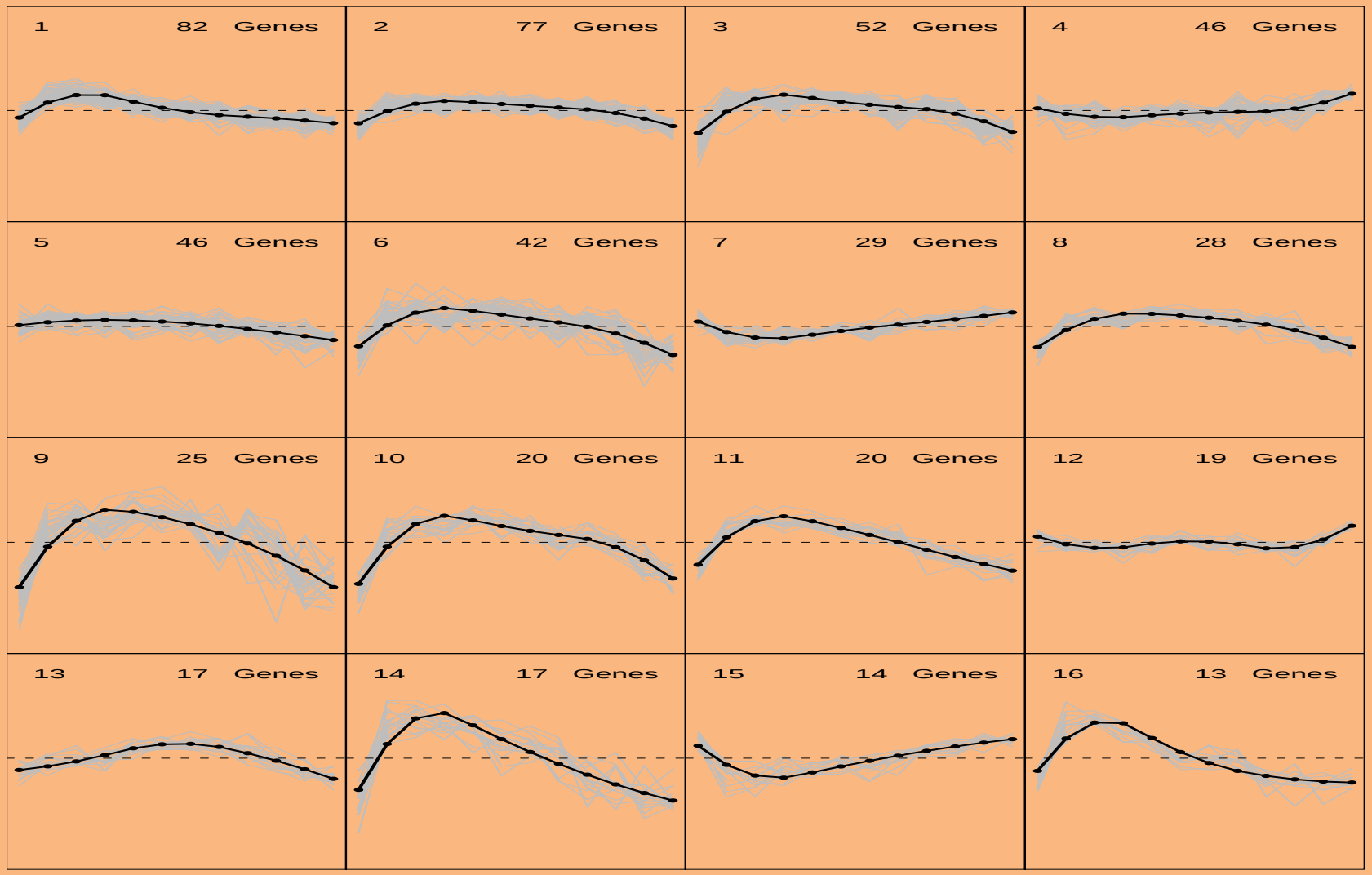
- Study of Corneal Wound Healing in Rats
  - 646 gene expression profiles, Affymetrix chip
  - Measured at Days 0,1,2,3,4,5,6,7,14,21,42,96
  - Day 0 sample prior to eye surgery

- Model: For objects in the kth cluster

$$Y_{ij} = X\beta_k + Z_2 V_k + \varepsilon_{ij}\,,$$

  - Base model = quadratic penalized spline with 2 knots

- Clustering of the Wound Healing Data: 16 largest clusters
  - The solid black = best linear unbiased predictor

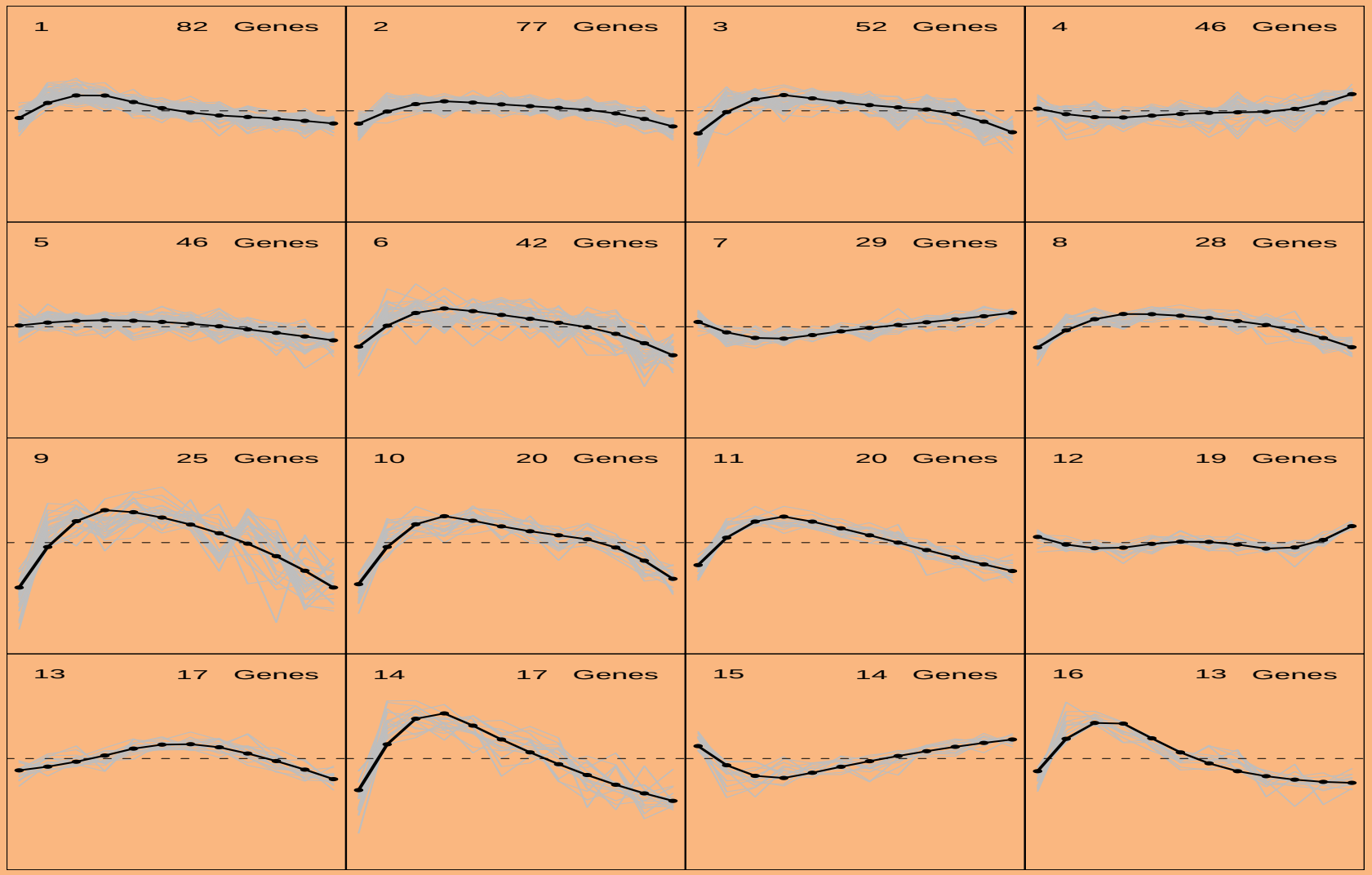


- Model: For objects in the kth cluster

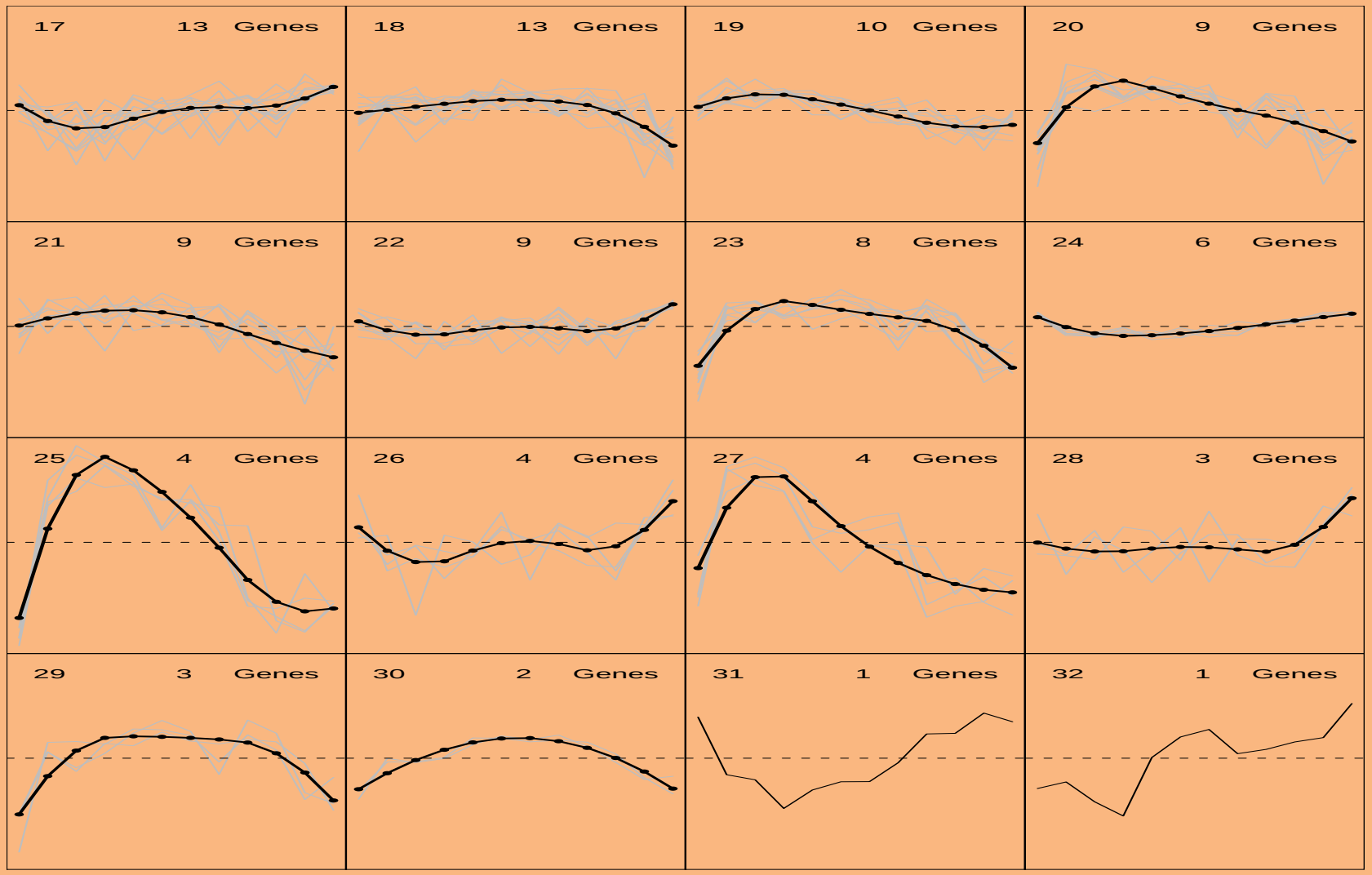
$$Y_{ij} = X\beta_k + Z_2 V_k + \varepsilon_{ij}\,,$$

- Following Ruppert *et al.* (2003) *Semparametric Regression*
  - Base Model: $X_{12\times 3}$ has jth row $(1, t_j, t_j^2)$.
  - Cluster effect: $Z_{2\ 12\times 2}$ has ith column $(t_j - \xi_i)_+^2$, $j = 0, \ldots, 11$, where $\xi_1 = -2$ and $\xi_2 = +2$.
  - Gene Effect: Variance component estimated as zero

- Clustering of the Wound Healing Data: 16 largest clusters
  - Clusters increase or decrease, then return to baseline
  - Dotted line = zero



- Clustering of the Wound Healing Data: 16 smallest clusters
  - Greater Within Cluster Variability



5. Conclusions and Other Stuff

- Multi-Level Mixed Model for Clustering Multivariate Data
- Allows for Cluster-Specific Effects (correlation)
- Clusters according to Base Model
  + Parsimonious Deviation (Booth wrote that)
- Also allows for object-specific correlation
  (repeated measures)

- This model quite different from previous ones
  - Not based on a mixture model
  - Contains a cluster-specific parameter
    - the object of inference
- Similar to *Partition Models* of Hartigan (1990) and Crowley (1997)

Other Stuff to Think About

- Other Priors on ω (the cluster parameter)
  - Can favor *small* or *large* clusters
- Other Search Algorithms (we tried most of them)
  - Gibbs sampler
  - Split-merge moves
  - Simulated annealing and simulated tempering

Other Stuff to Think About

- Variable Cluster Size
  - Extremely Important - Algorithm must allow for this
  - Misspecified cluster size $\Rightarrow$ Multimodal Likelihood (we think)
  - Misspecified cluster size $\Rightarrow$ Difficult Search
- Reporting many partitions
  - Can Report "Top ten most visited partitions"
  - Information about cluster and object variability