# Clustering through Mixture Models

**General references:**

- Lindsay B.G. (1995), *Mixture models: theory, geometry and applications*, NFS-CBMS Regional Conference Series in Probability and Statistics.
- McLachlan G.J., Basford K.E. (1988), *Mixture Models: Inference and Applications to Clustering*, Marcel Dekker, New York.
- Fraley C., Raftery A.E. (1998), How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis, *The Computer Journal*, **41**, 570-588.
- Dempster A.P., Laird N. M., Rubin, D. B. (1977) Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society*, Series B, **39**, 1-22.

**Applications to Microarray data:** reading list, several papers.

Examples are joint work with F. Bartolucci, Dept. of Stat. Univ. of Perugia, IT

## **Issues**:

- Reliability; arbitrariness (natural "lumpiness" of the data): bringing partitions and characteristic patterns within the domain of ***statistical inference***; substitute membership with ***membership probabilities***.
- Multiple and compounding sources of experimental error: ***robustification*** towards anomalies, while keeping an adequate degree of sensitivity.
- Much is unknown, but some aspects are well known or object of well defined hypotheses: integrating ***exploration*** and ***substantive modeling***.

An approach based on multivariate normal mixtures and maximum likelihood may provide some answers…

**The Mixture Approach**: data is a size $N$ sample from

$$X \in R^T \quad , \quad X \sim \sum_{c=1}^{C-1} \pi_c N(\mu_c, \Sigma_c) + \pi_C \Gamma \quad , \quad \pi_c \geq 0 \, , \sum_{c=1}^{C} \pi_c = 1$$

…each profile comes from one of $C$ alternative components

$C$; contamination term

Uniform on data range or sparse and spherical ("absorbs" anomalous profiles)

$$\Gamma = Un(\text{data range}) \quad \text{or}$$

$$\Gamma = N(\mu_C, \sigma_C^2 I) \quad \sigma_C^2 \geq \underline{\sigma}_C^2$$

"coverage radius"

$$\pi_C \leq \overline{\pi}_C$$

"degree of contamination"

$c$=1…$C$-1; regular components

Model means and within component covariance to various degrees of specificity

$$N(\mu_c, \Sigma_c)$$

$$\mu_c = Z_c \beta_c \, , \beta_c \in R^{p_c}$$

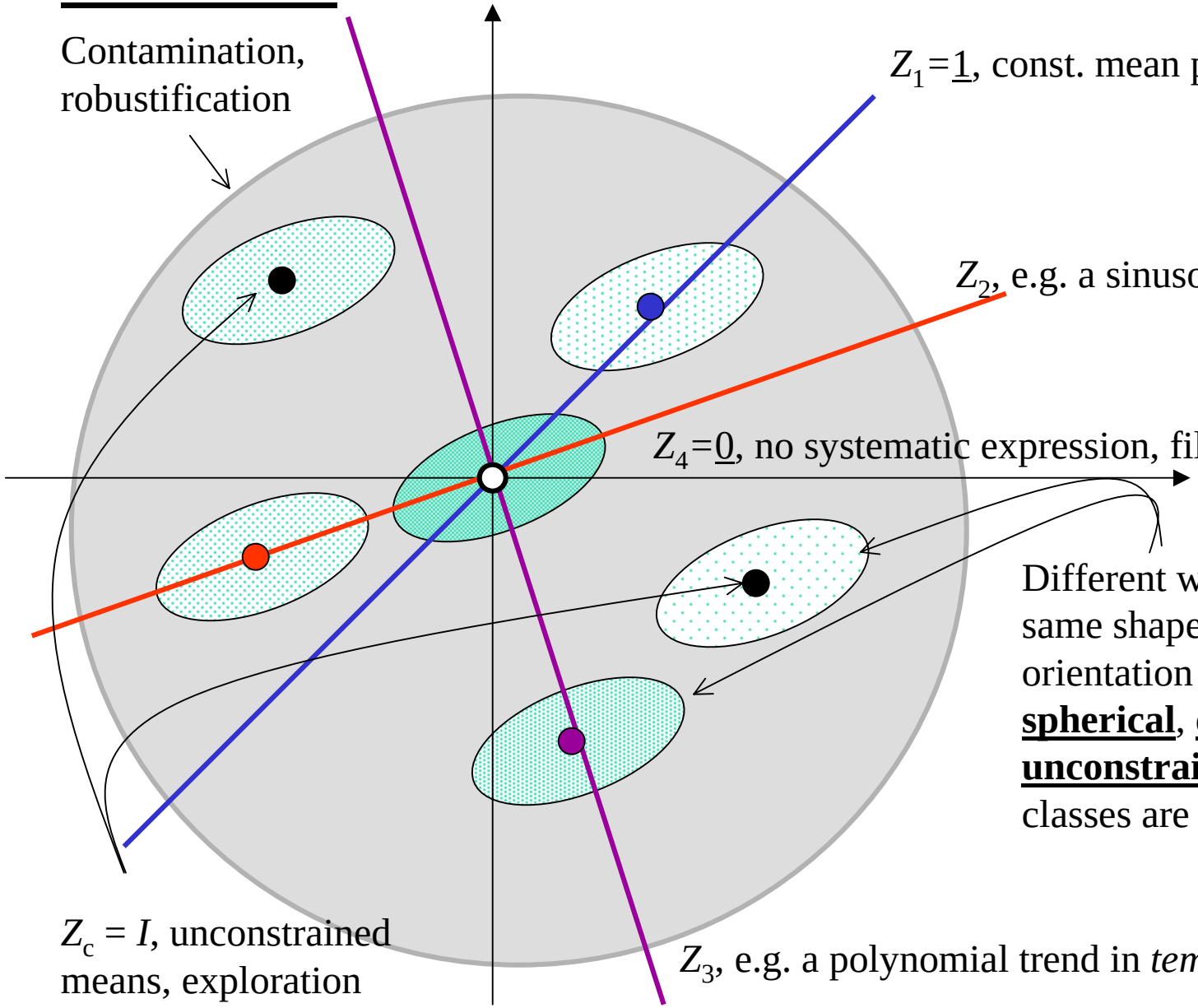
$$\Sigma_c \in S \quad \text{maybe} \quad \Sigma_c = \Sigma \in S$$

Linear re-param of means

Assume equal (better discrimination of within-between variation), and model

# A cartoon…

**Linear constraints** on mean patterns



Contamination, robustification

$Z_1$=1, const. mean pattern, filtering

$Z_2$, e.g. a sinusoidal trend in *time*$_t$

$Z_4$=0, no systematic expression, filtering

Different weights ($\pi_c$), but same shape, size and orientation ($\Sigma$) from *S*. e.g. **spherical**, **diagonal** and **unconstrained**, but other classes are possible.

$Z_c = I$, unconstrained means, exploration

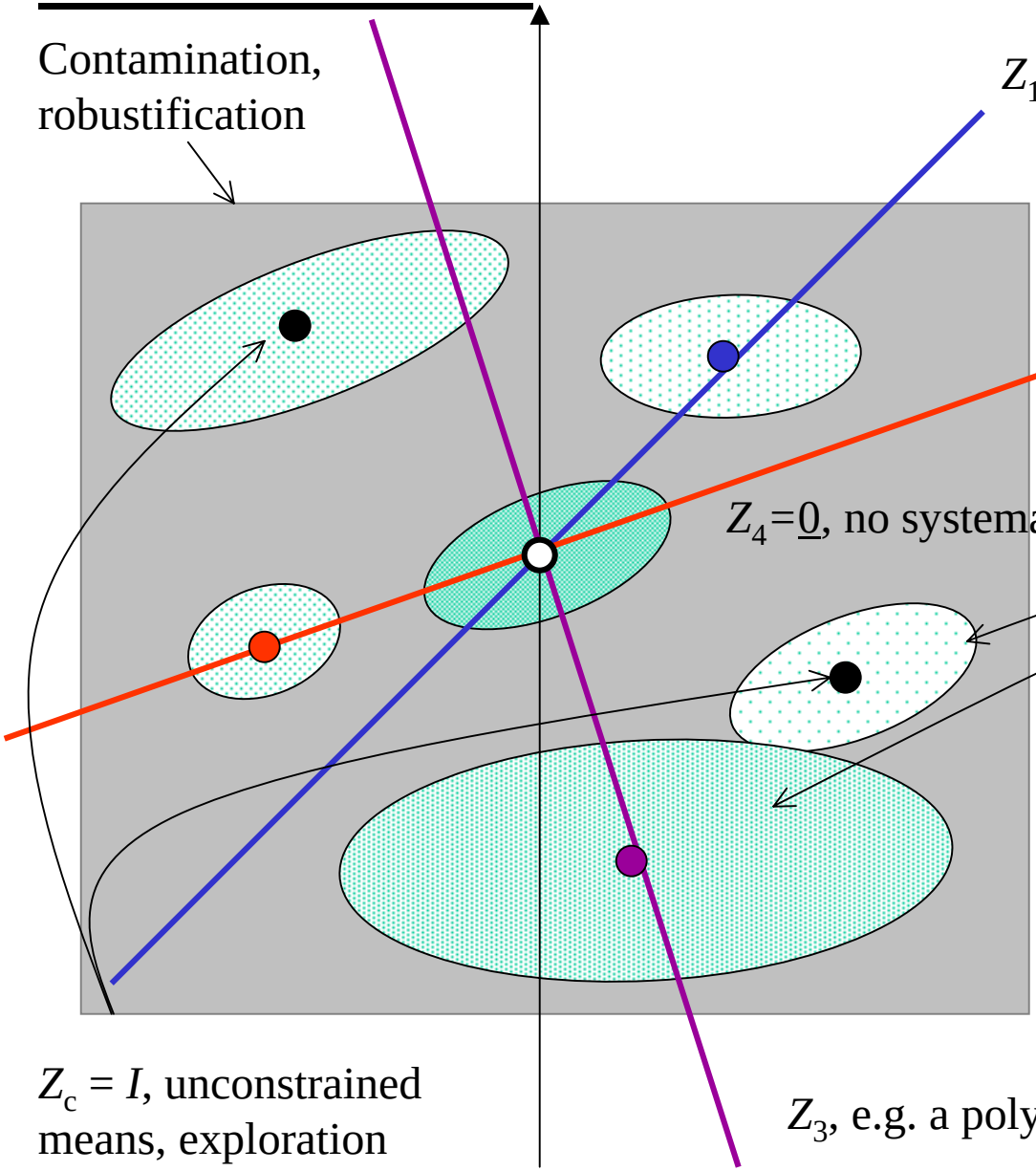
$Z_3$, e.g. a polynomial trend in *temp*$_t$

# Another cartoon…

**Linear constraints** on mean patterns

Contamination, robustification

$Z_1$=1, const. mean pattern, filtering



$Z_4$=0, no systematic expression, filtering

Different weights ($\pi_c$), Shape, size and orientation ($\Sigma_c$) from *S*. e.g. **modeled to share certain features** (but not others), or **unconstrained**.

$Z_c = I$, unconstrained means, exploration

$Z_3$, e.g. a polynomial trend in $temp_t$

**Log likelihood(s)**:

Unobserved component membership vectors

$$X_i \in R^T, m_i \in \{0,1\}^C, i = 1...N$$

$$\pi = (\pi_1 \ldots \pi_{C1})'$$

*T*-variate normal density

… or uniform on data range

$$f_i(\tau) = (\varphi(X_i; Z_1\beta_1, \Sigma) \ldots \varphi(X_i; Z_{C-1}\beta_{C-1}, \Sigma), \varphi(X_i; \mu_C, \sigma_C^2 I))'$$

$$l_X(\vartheta) = \sum_{i=1}^{N} \log(\pi' f_i(\tau))$$ ← ***"incomplete"***

$$l_{X,M}(\vartheta) = \sum_{i=1}^{N} m_i' \log(f_i(\tau)) + \sum_{i=1}^{N} m_i' \log(\pi)$$ ← ***"complete"***

## **Numerical maximization via EM algorithm**:

E) Using the current parameter values compute

$$\overline{m}_i = E(m_i \mid X_i) = (\hat{\pi}' f_i(\hat{\tau}))^{-1} diag(\hat{\pi}) f_i(\hat{\tau}), \quad i = 1...N$$

M) Substitute the current parameter values with the maximum of

$$\bar{l}_{X,M}(\vartheta) = E(l_{X,M}(\vartheta) \mid X) = \sum_{i=1}^{N} \overline{m}_i' \log(f_i(\tau)) + \sum_{i=1}^{N} \overline{m}_i' \log(\pi)$$

Iterate until convergence.

**Initialization**: $\overline{m}_i^{(0)}, i = 1..N$

memberships from a k-means clustering with $k=C$-1. Or other strategies (dependence on initialization is an issue also here)

## Outcomes, from the last iteration:

**(Very rich; lots of information!)**

$\hat{\pi}_c$ , $c = 1...C-1$ ← Estimated ***weights***

$\hat{\mu}_c = Z_c \hat{\beta}_c$ , $c = 1...C-1$ ← Estimated ***mean patterns***

$\hat{\Sigma}_c \in S$ , $c = 1...C-1$ or $\hat{\Sigma} \in S$ ← Estimated ***within-component variability structure(s)***

$\hat{\pi}_C$ and possibly $\hat{\mu}_C$ , $\hat{\sigma}^2_C$ ← Estimated ***contamination parameters***

$\hat{p}_i = \overline{m}_i$ , $i = 1...N$

Estimated vectors of conditional prob's; ***membership probabilities***

## Cluster formation:

$i \in Cluster(c) \quad \leftrightarrow \quad \max\{\hat{p}_{i1} \ldots \hat{p}_{iC}\} = p_i^* = \hat{p}_{ic}$

or, threshold $\gamma \in (0,1)$

$i \in Cluster(c) \quad \leftrightarrow \quad \max\{p_i^*; \gamma\} = p_i^* = \hat{p}_{ic}$

residual $(C+1)$th class for $i : p_i^* < \gamma$

Their distribution's high end concentration gives interesting info on ***"lumpiness"*** of the data, in the context established by choice of $C$ and constraints specification

**First application**:

Spellman *et al*., 1998, expression of yeast genes on a time course covering 2+ cell cycles. Log ratios; baseline = unsynchronized culture. Select 800 genes with periodic expression profiles. Halter et al., 2000 restrict attention to T=12 equispaced time points recovering 2 cell cycles, and profiles without missing values (most of the variability of the data cloud is captured by the first two principal components;  data do not appear "lumpy").

Here we use an N=696 by 12 data matrix (neglect genes with missing values in the first 12 time points), but do not center and standardize by row/gene profile.

- No missing value imputation;
- contamination = spherical normal;
- common within component covariance structure.

## Fits in first application:

- K-means, k=8 (initialization for all mixture fits below)
- Mix. Fit A: closest to k-means. *C*-1=8 regular components, plus contamination. Unconstrained mean patterns. Spherical within-comp. cov. structure (var. about mean pattern equal and uncorr. over t's).
- Mix. Fit B: relaxation of A; diagonal within-comp. cov. structure (var. about mean pattern different but uncorrelated over t's).
- [Mix. Fit C: relaxation of B; unconstrained within-comp. cov. structure (var. about mean pattern different and freely correlated over t's)].
- Mix. Fit D: a restriction of B; mean patterns modeled as

$$\mu_{ct} = (\beta_{c1} + \beta_{c2} t) + (\beta_{c3} + \beta_{c4} t) \sin\left(\frac{(t - shift_c)2\pi}{period}\right), \quad t = 1...12, c = 1...8$$

$\beta$ 's (continuously) optimized by EM

optimized at the outset over a grid

**Imputing missing values**:

the *X*'s may contain missing values that will end up in the category of unobserved data (not in the incomplete likelihood), and will be imputed by the EM algorithm

Unobserved, component membership vectors

$$X_i = (X_i^{obs}; X_i^{miss}) \in R^T, m_i \in \{0,1\}^C, i = 1...N$$

Unobserved, missing values

E) Using the current parameter values compute

$$(\overline{X}_i^{miss}; \overline{m}_i) = E(X_i^{miss}; m_i \mid X_i^{obs})$$

M) Substitute the current parameter values with the maximum of

$$\bar{l}_{X^{obs}, X^{miss}, M}(\vartheta) = E(l_{X^{obs}, X^{miss}, M}(\vartheta) \mid X^{obs})$$

From last iteration, among other outputs:

$$\hat{X}_i^{miss} = \overline{X}_i^{miss} \qquad \hat{p}_i = \overline{m}_i, i = 1...N$$

## **Second application**:

Gasch A.P., Spellman P.T., Kao C.M., Carmel-Harel O., Eisen M.B., Storz G., Botstein D., Brown P.O. (2001), Genomic Expression Programs in the Response of Yeast Cells to Environmental Changes, *Molecular Biology of the Cell* **11** 4241-4257.

N=6152 known and putative genes on over 140 conditions. We concentrate on a T=8 time course for heat shock (25 to 37C, minute 5, 10, 15, 20, 30,40, 60, 80). Log ratios; baseline=pooling equal amounts of all experimental samples. The profiles of 2509 genes (40.78% of the total) have missing values.

We use this 6152x8 matrix, without centering and standardize by row/gene profile.

- Missing value imputation;
- contamination = uniform on data range;
- allow for different within component covariance specifications (also different)

**Fits in second application**:

- free means, EEE covariances: C-1=7, common within component covariance structure, unconstrained.
- free means and UUE covariances: C-1=7, each component has a common (but not fixed) correlation structure, but differences in overall variability volume and distribution over the time course are allowed.

(many more, also modeling means, not presented)