Combining Mixture Components for Clustering

Gilles Celeux
INRIA, Saclay Île-de-France

Joint work with Jean-Patrick Baudry, Adrian Raftery, Kenneth Lo and Raphaël Gottardo
Supported by NICHD and NSF

Journées Franco-Roumaines 2010, Poitiers
27 août 2010

Outline

Outline

- Model-based clustering

Outline

- Model-based clustering
- Choice of the number of components: BIC and ICL

Outline

- Model-based clustering
- Choice of the number of components: BIC and ICL
- Combining mixture components for clustering

Outline

- Model-based clustering
- Choice of the number of components: BIC and ICL
- Combining mixture components for clustering
- Simulation example

Outline

- Model-based clustering
- Choice of the number of components: BIC and ICL
- Combining mixture components for clustering
- Simulation example
- Flow cytometry example

Basic Ideas of Model-Based Clustering

Basic Ideas of Model-Based Clustering

- Based on a finite mixture of multivariate normal distributions:

$$y_i \sim \sum_{g=1}^{G} \tau_g \mathrm{MVN}_d(\mu_g, \Sigma_g),$$

Basic Ideas of Model-Based Clustering

- Based on a finite mixture of multivariate normal distributions:

$$y_i \sim \sum_{g=1}^{G} \tau_g \mathrm{MVN}_d(\mu_g, \Sigma_g),$$

- where $\Sigma_g = \lambda_g D_g A_g D_g^T$

Basic Ideas of Model-Based Clustering

- Based on a finite mixture of multivariate normal distributions:

$$y_i \sim \sum_{g=1}^{G} \tau_g \mathrm{MVN}_d(\mu_g, \Sigma_g),$$

- where $\Sigma_g = \lambda_g D_g A_g D_g^T$
- λ_g = determinant of Σ_g: controls the *volume* of the gth cluster

Basic Ideas of Model-Based Clustering

- Based on a finite mixture of multivariate normal distributions:

$$y_i \sim \sum_{g=1}^{G} \tau_g \mathrm{MVN}_d(\mu_g, \Sigma_g),$$

- where $\Sigma_g = \lambda_g D_g A_g D_g^T$
- λ_g = determinant of Σ_g: controls the *volume* of the gth cluster
- $A_g = \mathrm{diag}\{1, \alpha_{2g}, \ldots, \alpha_{dg}\}$

Basic Ideas of Model-Based Clustering

- Based on a finite mixture of multivariate normal distributions:

$$y_i \sim \sum_{g=1}^{G} \tau_g \mathrm{MVN}_d(\mu_g, \Sigma_g),$$

- where $\Sigma_g = \lambda_g D_g A_g D_g^T$
- λ_g = determinant of Σ_g: controls the *volume* of the gth cluster
- $A_g = \text{diag}\{1, \alpha_{2g}, \ldots, \alpha_{dg}\}$
  - controls the *shape* of the gth cluster

Basic Ideas of Model-Based Clustering

- Based on a finite mixture of multivariate normal distributions:

$$y_i \sim \sum_{g=1}^{G} \tau_g \mathrm{MVN}_d(\mu_g, \Sigma_g),$$

- where $\Sigma_g = \lambda_g D_g A_g D_g^T$
- λ_g = determinant of Σ_g: controls the *volume* of the gth cluster
- $A_g = \mathrm{diag}\{1, \alpha_{2g}, \ldots, \alpha_{dg}\}$
  - controls the *shape* of the gth cluster
  - $(1 \geq \alpha_2 \geq \ldots \geq \alpha_d > 0)$

Basic Ideas of Model-Based Clustering

- Based on a finite mixture of multivariate normal distributions:

$$y_i \sim \sum_{g=1}^{G} \tau_g \mathrm{MVN}_d(\mu_g, \Sigma_g),$$

- where $\Sigma_g = \lambda_g D_g A_g D_g^T$
- λ_g = determinant of Σ_g: controls the *volume* of the gth cluster
- $A_g = \mathrm{diag}\{1, \alpha_{2g}, \ldots, \alpha_{dg}\}$
    - controls the *shape* of the gth cluster
    - $(1 \geq \alpha_2 \geq \ldots \geq \alpha_d > 0)$
    - E.g. α_2 close to zero: Cluster g concentrated about a line.

Basic Ideas of Model-Based Clustering

- Based on a finite mixture of multivariate normal distributions:

$$y_i \sim \sum_{g=1}^{G} \tau_g \mathrm{MVN}_d(\mu_g, \Sigma_g),$$

- where $\Sigma_g = \lambda_g D_g A_g D_g^T$
- λ_g = determinant of Σ_g: controls the *volume* of the gth cluster
- $A_g = \mathrm{diag}\{1, \alpha_{2g}, \ldots, \alpha_{dg}\}$
    - controls the *shape* of the gth cluster
    - $(1 \geq \alpha_2 \geq \ldots \geq \alpha_d > 0)$
    - E.g. α_2 close to zero: Cluster g concentrated about a line.
    - E.g. $\alpha_{2g}, \ldots, \alpha_{dg}$ all close to 1: Cluster g nearly spherical.

Basic Ideas of Model-Based Clustering

- Based on a finite mixture of multivariate normal distributions:

$$y_i \sim \sum_{g=1}^{G} \tau_g \mathrm{MVN}_d(\mu_g, \Sigma_g),$$

- where $\Sigma_g = \lambda_g D_g A_g D_g^T$
- λ_g = determinant of Σ_g: controls the *volume* of the gth cluster
- $A_g = \mathrm{diag}\{1, \alpha_{2g}, \ldots, \alpha_{dg}\}$
    - controls the *shape* of the gth cluster
    - $(1 \geq \alpha_2 \geq \ldots \geq \alpha_d > 0)$
    - E.g. α_2 close to zero: Cluster g concentrated about a line.
    - E.g. $\alpha_{2g}, \ldots, \alpha_{dg}$ all close to 1: Cluster g nearly spherical.
- D_g = Eigenvectors: Control the *orientation* of the gth cluster

Basic Ideas of Model-Based Clustering

- Based on a finite mixture of multivariate normal distributions:

$$y_i \sim \sum_{g=1}^{G} \tau_g \mathrm{MVN}_d(\mu_g, \Sigma_g),$$

- where $\Sigma_g = \lambda_g D_g A_g D_g^T$
- λ_g = determinant of Σ_g: controls the *volume* of the gth cluster
- $A_g = \text{diag}\{1, \alpha_{2g}, \ldots, \alpha_{dg}\}$
  - controls the *shape* of the gth cluster
  - $(1 \geq \alpha_2 \geq \ldots \geq \alpha_d > 0)$
  - E.g. α_2 close to zero: Cluster g concentrated about a line.
  - E.g. $\alpha_{2g}, \ldots, \alpha_{dg}$ all close to 1: Cluster g nearly spherical.
- D_g = Eigenvectors: Control the *orientation* of the gth cluster
- Different clustering models can be obtained by constraining each of *volume*, *shape* and *orientation* to be constant across clusters, or by allowing them to vary (Banfield & Raftery, 93, Celeux & Govaert 95)

Model-Based Clustering Strategy

Model-Based Clustering Strategy

- Maximum likelihood estimation for the mixture model parameters $\theta = (\tau, \mu, \Sigma)$, via the EM algorithm

Model-Based Clustering Strategy

- Maximum likelihood estimation for the mixture model parameters $\theta = (\tau, \mu, \Sigma)$, via the EM algorithm
- Initialization of EM via repeated small runs of EM from many radom positions.

Model-Based Clustering Strategy

- Maximum likelihood estimation for the mixture model parameters $\theta = (\tau, \mu, \Sigma)$, via the EM algorithm
- Initialization of EM via repeated small runs of EM from many radom positions.
- Choosing the Number of Clusters and the Clustering Method/Model:

Model-Based Clustering Strategy

- Maximum likelihood estimation for the mixture model parameters $\theta = (\tau, \mu, \Sigma)$, via the EM algorithm
- Initialization of EM via repeated small runs of EM from many radom positions.
- Choosing the Number of Clusters and the Clustering Method/Model:
  - Both are reduced to statistical model selection problems, and solved simultaneously.

Model-Based Clustering Strategy

- Maximum likelihood estimation for the mixture model parameters $\theta = (\tau, \mu, \Sigma)$, via the EM algorithm
- Initialization of EM via repeated small runs of EM from many radom positions.
- Choosing the Number of Clusters and the Clustering Method/Model:
  - Both are reduced to statistical model selection problems, and solved simultaneously.
  - Each combination of (Number of Clusters, Clustering Model) is viewed as a separate statistical model

Model-Based Clustering Strategy

- Maximum likelihood estimation for the mixture model parameters $\theta = (\tau, \mu, \Sigma)$, via the EM algorithm
- Initialization of EM via repeated small runs of EM from many radom positions.
- Choosing the Number of Clusters and the Clustering Method/Model:
  - Both are reduced to statistical model selection problems, and solved simultaneously.
  - Each combination of (Number of Clusters, Clustering Model) is viewed as a separate statistical model
  - We use the Bayes factor, i.e. the ratio of posterior to prior odds for one model against another.

Model-Based Clustering Strategy

- Maximum likelihood estimation for the mixture model parameters $\theta = (\tau, \mu, \Sigma)$, via the EM algorithm
- Initialization of EM via repeated small runs of EM from many radom positions.
- Choosing the Number of Clusters and the Clustering Method/Model:
  - Both are reduced to statistical model selection problems, and solved simultaneously.
  - Each combination of (Number of Clusters, Clustering Model) is viewed as a separate statistical model
  - We use the Bayes factor, i.e. the ratio of posterior to prior odds for one model against another.
  - This allows comparison of the multiple, nonnested models considered.

Model-Based Clustering Strategy

- Maximum likelihood estimation for the mixture model parameters $\theta = (\tau, \mu, \Sigma)$, via the EM algorithm
- Initialization of EM via repeated small runs of EM from many radom positions.
- Choosing the Number of Clusters and the Clustering Method/Model:
  - Both are reduced to statistical model selection problems, and solved simultaneously.
  - Each combination of (Number of Clusters, Clustering Model) is viewed as a separate statistical model
  - We use the Bayes factor, i.e. the ratio of posterior to prior odds for one model against another.
  - This allows comparison of the multiple, nonnested models considered.
  - We approximate the Bayes factors via

$$\text{BIC} = 2 \log \text{maximized likelihood} - (\#\ \text{parameters}) \log(n)$$

Model-Based Clustering Strategy

- Maximum likelihood estimation for the mixture model parameters $\theta = (\tau, \mu, \Sigma)$, via the EM algorithm
- Initialization of EM via repeated small runs of EM from many radom positions.
- Choosing the Number of Clusters and the Clustering Method/Model:
  - Both are reduced to statistical model selection problems, and solved simultaneously.
  - Each combination of (Number of Clusters, Clustering Model) is viewed as a separate statistical model
  - We use the Bayes factor, i.e. the ratio of posterior to prior odds for one model against another.
  - This allows comparison of the multiple, nonnested models considered.
  - We approximate the Bayes factors via

$$\mathrm{BIC} = 2 \log \text{maximized likelihood} - (\#\ \text{parameters}) \log(n)$$

  - This is consistent for the number of components (Keribin 2000), and also provides consistent density estimates (Roeder and Wasserman 1997).

Choice of Number of Components: Simulation Study

10 experiments based on distribution of estimates in literature (Steele & Raftery 2010)

Choice of Number of Components: Simulation Study

10 experiments based on distribution of estimates in literature (Steele & Raftery 2010)

Times right model chosen/50 (bigger is better)

Expt.	BIC	Stephens	AIC	ICL	UIP	DIC
1	50	49	45	50	44	20
2	50	48	38	50	39	17
3	50	50	42	50	40	22
4	49	48	34	50	30	14
5	49	46	33	49	19	16
6	23	29	35	0	40	20
7	50	42	46	19	34	23
8	47	45	45	16	33	14
9	50	41	37	39	22	10
10	50	43	39	50	7	20
Total	468	441	394	373	308	176
% Correct	94	88	79	75	62	35

Choice of Number of Components: Simulation Study

10 experiments based on distribution of estimates in literature (Steele & Raftery 2010)

MISE of density estimate (smaller is better)

Expt.	BIC	Stephens	AIC	ICL	UIP	DIC
1	0.19	0.21	0.22	0.19	0.23	0.67
2	0.21	0.24	0.33	0.21	0.31	0.65
3	0.35	0.35	0.41	0.35	0.50	1.32
4	0.48	0.51	1.30	0.48	1.35	2.24
5	0.60	1.00	1.58	0.60	2.75	3.20
6	1.53	1.13	0.86	2.31	0.77	0.76
7	0.23	0.24	0.23	2.18	0.25	0.28
8	0.55	0.39	0.37	2.45	0.42	0.61
9	0.37	0.75	0.47	0.61	0.58	0.77
10	0.34	0.44	0.39	0.34	0.75	0.58
Mean	0.48	0.53	0.62	0.97	0.79	1.11

Choosing the Number of Clusters: ICL, a first solution

Choosing the Number of Clusters: ICL, a first solution

- Problem: Cluster $\neq$ One mixture component, if its distribution is not Gaussian

Choosing the Number of Clusters: ICL, a first solution

- Problem: Cluster $\neq$ One mixture component, if its distribution is not Gaussian
  - It might be better represented by two or more mixture components

Choosing the Number of Clusters: ICL, a first solution

- Problem: Cluster $\neq$ One mixture component, if its distribution is not Gaussian
  - It might be better represented by two or more mixture components
  - Thus # Clusters $\leq$ # Mixture components

Choosing the Number of Clusters: ICL, a first solution

- Problem: Cluster $\neq$ One mixture component, if its distribution is not Gaussian
  - It might be better represented by two or more mixture components
  - Thus # Clusters $\leq$ # Mixture components
- First solution: Instead of BIC, which approximates the log integrated likelihood of the data,

$$\log p(\mathbf{x}|K) = \int p(\mathbf{x}|K, \theta_K)\pi(\theta_K)d\theta_K,$$

Choosing the Number of Clusters: ICL, a first solution

- Problem: Cluster $\neq$ One mixture component, if its distribution is not Gaussian
  - It might be better represented by two or more mixture components
  - Thus # Clusters $\leq$ # Mixture components
- First solution: Instead of BIC, which approximates the log integrated likelihood of the data,

$$\log p(\mathbf{x}|K) = \int p(\mathbf{x}|K, \theta_K)\pi(\theta_K)d\theta_K,$$

Choosing the Number of Clusters: ICL, a first solution

- Problem: Cluster $\neq$ One mixture component, if its distribution is not Gaussian
  - It might be better represented by two or more mixture components
  - Thus # Clusters $\leq$ # Mixture components
- First solution: Instead of BIC, which approximates the log integrated likelihood of the data,

$$\log p(\mathbf{x}|K) = \int p(\mathbf{x}|K, \theta_K)\pi(\theta_K)d\theta_K,$$

use ICL, which approximates the log integrated likelihood of the *completed data*,

$$\text{ICL}(K) = \log p(\mathbf{x}, \mathbf{z} \mid K) \quad = \quad \int_{\Theta_K} p(\mathbf{x}, \mathbf{z} \mid K, \theta)\pi(\theta \mid K)d\theta$$

Choosing the Number of Clusters: ICL, a first solution

- Problem: Cluster $\neq$ One mixture component, if its distribution is not Gaussian
  - It might be better represented by two or more mixture components
  - Thus # Clusters $\leq$ # Mixture components
- First solution: Instead of BIC, which approximates the log integrated likelihood of the data,

$$\log p(\mathbf{x}|K) = \int p(\mathbf{x}|K, \theta_K)\pi(\theta_K)d\theta_K,$$

use ICL, which approximates the log integrated likelihood of the *completed data*,

$$\begin{aligned} \mathrm{ICL}(K) = \log p(\mathbf{x}, \mathbf{z} \mid K) &= \int_{\Theta_K} p(\mathbf{x}, \mathbf{z} \mid K, \theta)\pi(\theta \mid K)d\theta \\ &\approx log\, \mathbf{p}(\mathbf{x}, \hat{\mathbf{z}} \mid K, \hat{\theta}_K) - \frac{\nu_K}{2}\log n \end{aligned}$$

(Biernacki, Celeux & Govaert 2000)

ICL and Entropy

ICL and Entropy

- $\text{ICL}(K) \approx \text{BIC}(K) -$ the mean entropy, $\text{Ent}(K)$,

ICL and Entropy

- ICL(K) ≈ BIC(K) − the mean entropy, Ent(K),
  - $\mathrm{Ent}(K) = -\sum_{k=1}^{K} \sum_{i=1}^{n} t_{ik}(\hat{\theta}_K) \log t_{ik}(\hat{\theta}_K) \geq 0$

ICL and Entropy

- ICL(K) $\approx$ BIC(K) $-$ the mean entropy, Ent(K),
  - $\mathrm{Ent}(K) = -\sum_{k=1}^{K} \sum_{i=1}^{n} t_{ik}(\hat{\theta}_K) \log t_{ik}(\hat{\theta}_K) \geq 0$
  - where t_{ik} = conditional probability that $\mathbf{x}_i$ is from kth mixture component

ICL and Entropy

- ICL(K) ≈ BIC(K) − the mean entropy, Ent(K),
  - $\mathrm{Ent}(K) = -\sum_{k=1}^{K}\sum_{i=1}^{n} t_{ik}(\hat{\theta}_K)\log t_{ik}(\hat{\theta}_K) \geq 0$
  - where t_{ik} = conditional probability that $\mathbf{x}_i$ is from kth mixture component
  - Thus ICL tends to find smaller K than BIC

ICL and Entropy

- ICL(K) ≈ BIC(K) − the mean entropy, Ent(K),
  - $\mathrm{Ent}(K) = -\sum_{k=1}^{K} \sum_{i=1}^{n} t_{ik}(\hat{\theta}_K) \log t_{ik}(\hat{\theta}_K) \geq 0$
  - where t_{ik} = conditional probability that $\mathbf{x}_i$ is from kth mixture component
  - Thus ICL tends to find smaller K than BIC
- Problem: If ICL is used to estimate the number of mixture components, it tends to underestimate it when there are poorly separated components, and so can fit the data poorly

ICL and Entropy

- ICL(K) $\approx$ BIC(K) $-$ the mean entropy, Ent(K),
  - $\mathrm{Ent}(K) = -\sum_{k=1}^{K} \sum_{i=1}^{n} t_{ik}(\hat{\theta}_K) \log t_{ik}(\hat{\theta}_K) \geq 0$
  - where t_{ik} = conditional probability that $\mathbf{x}_i$ is from kth mixture component
  - Thus ICL tends to find smaller K than BIC
- Problem: If ICL is used to estimate the number of mixture components, it tends to underestimate it when there are poorly separated components, and so can fit the data poorly
- Goal: Find a method that gives the best of both worlds:

ICL and Entropy

- ICL(K) $\approx$ BIC(K) $-$ the mean entropy, Ent(K),
  - $\mathrm{Ent}(K) = -\sum_{k=1}^{K} \sum_{i=1}^{n} t_{ik}(\hat{\theta}_K) \log t_{ik}(\hat{\theta}_K) \geq 0$
  - where t_{ik} = conditional probability that $\mathbf{x}_i$ is from kth mixture component
  - Thus ICL tends to find smaller K than BIC
- Problem: If ICL is used to estimate the number of mixture components, it tends to underestimate it when there are poorly separated components, and so can fit the data poorly
- Goal: Find a method that gives the best of both worlds:
  - fits the data well (like BIC), and

ICL and Entropy

- ICL(K) $\approx$ BIC(K) $-$ the mean entropy, Ent(K),
  - $\mathrm{Ent}(K) = -\sum_{k=1}^{K} \sum_{i=1}^{n} t_{ik}(\hat{\theta}_K) \log t_{ik}(\hat{\theta}_K) \geq 0$
  - where t_{ik} = conditional probability that $\mathbf{x}_i$ is from kth mixture component
  - Thus ICL tends to find smaller K than BIC
- Problem: If ICL is used to estimate the number of mixture components, it tends to underestimate it when there are poorly separated components, and so can fit the data poorly
- Goal: Find a method that gives the best of both worlds:
  - fits the data well (like BIC), and
  - identifies clusters rather than mixture components (like ICL)

Combining Mixture Components for Clustering

Combining Mixture Components for Clustering

- Start with a mixture model that fits the data well, with K chosen by BIC

Combining Mixture Components for Clustering

- Start with a mixture model that fits the data well, with K chosen by BIC
  - Design a *sequence* of soft clusterings with $K, K-1, \ldots, 1$ clusters by successively merging the components

Combining Mixture Components for Clustering

- Start with a mixture model that fits the data well, with K chosen by BIC
  - Design a *sequence* of soft clusterings with $K, K-1, \ldots, 1$ clusters by successively merging the components
  - At each stage we choose the two mixture components to be merged so as to minimize the entropy of the resulting clustering

Combining Mixture Components for Clustering

- Start with a mixture model that fits the data well, with K chosen by BIC
  - Design a *sequence* of soft clusterings with $K, K-1, \ldots, 1$ clusters by successively merging the components
  - At each stage we choose the two mixture components to be merged so as to minimize the entropy of the resulting clustering
- These clusterings all fit the data equally well:

Combining Mixture Components for Clustering

- Start with a mixture model that fits the data well, with K chosen by BIC
  - Design a *sequence* of soft clusterings with $K, K-1, \ldots, 1$ clusters by successively merging the components
  - At each stage we choose the two mixture components to be merged so as to minimize the entropy of the resulting clustering
- These clusterings all fit the data equally well:
  - the likelihood doesn't change.

Combining Mixture Components for Clustering

- Start with a mixture model that fits the data well, with K chosen by BIC
  - Design a *sequence* of soft clusterings with $K, K-1, \ldots, 1$ clusters by successively merging the components
  - At each stage we choose the two mixture components to be merged so as to minimize the entropy of the resulting clustering
- These clusterings all fit the data equally well:
  - the likelihood doesn't change.
  - Only the number and definition of clusters are different

Combining Mixture Components for Clustering

- Start with a mixture model that fits the data well, with K chosen by BIC
  - Design a *sequence* of soft clusterings with $K, K-1, \ldots, 1$ clusters by successively merging the components
  - At each stage we choose the two mixture components to be merged so as to minimize the entropy of the resulting clustering
- These clusterings all fit the data equally well:
  - the likelihood doesn't change.
  - Only the number and definition of clusters are different
  - one clustering for each number of clusters

Combining Mixture Components for Clustering

- Start with a mixture model that fits the data well, with K chosen by BIC
  - Design a *sequence* of soft clusterings with $K, K-1, \ldots, 1$ clusters by successively merging the components
  - At each stage we choose the two mixture components to be merged so as to minimize the entropy of the resulting clustering
- These clusterings all fit the data equally well:
  - the likelihood doesn't change.
  - Only the number and definition of clusters are different
  - one clustering for each number of clusters
- Choosing the number of clusters:

Combining Mixture Components for Clustering

- Start with a mixture model that fits the data well, with K chosen by BIC
  - Design a *sequence* of soft clusterings with $K, K-1, \ldots, 1$ clusters by successively merging the components
  - At each stage we choose the two mixture components to be merged so as to minimize the entropy of the resulting clustering
- These clusterings all fit the data equally well:
  - the likelihood doesn't change.
  - Only the number and definition of clusters are different
  - one clustering for each number of clusters
- Choosing the number of clusters:
  - substantive grounds, or

Combining Mixture Components for Clustering

- Start with a mixture model that fits the data well, with K chosen by BIC
  - Design a *sequence* of soft clusterings with $K, K-1, \ldots, 1$ clusters by successively merging the components
  - At each stage we choose the two mixture components to be merged so as to minimize the entropy of the resulting clustering
- These clusterings all fit the data equally well:
  - the likelihood doesn't change.
  - Only the number and definition of clusters are different
  - one clustering for each number of clusters
- Choosing the number of clusters:
  - substantive grounds, or
  - choose the number selected by ICL, or

Combining Mixture Components for Clustering

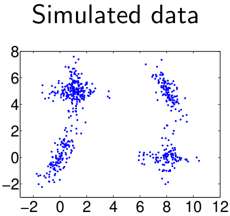
- Start with a mixture model that fits the data well, with K chosen by BIC
  - Design a *sequence* of soft clusterings with $K, K-1, \ldots, 1$ clusters by successively merging the components
  - At each stage we choose the two mixture components to be merged so as to minimize the entropy of the resulting clustering
- These clusterings all fit the data equally well:
  - the likelihood doesn't change.
  - Only the number and definition of clusters are different
  - one clustering for each number of clusters
- Choosing the number of clusters:
  - substantive grounds, or
  - choose the number selected by ICL, or
  - seek an elbow in the plot of the entropy versus # clusters, or

Combining Mixture Components for Clustering

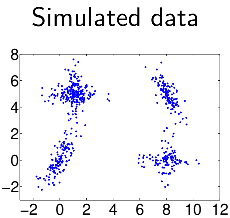
- Start with a mixture model that fits the data well, with K chosen by BIC
  - Design a *sequence* of soft clusterings with $K, K-1, \ldots, 1$ clusters by successively merging the components
  - At each stage we choose the two mixture components to be merged so as to minimize the entropy of the resulting clustering
- These clusterings all fit the data equally well:
  - the likelihood doesn't change.
  - Only the number and definition of clusters are different
  - one clustering for each number of clusters
- Choosing the number of clusters:
  - substantive grounds, or
  - choose the number selected by ICL, or
  - seek an elbow in the plot of the entropy versus # clusters, or
  - use piecewise regression to find the elbow (Byers & Raftery 1998)

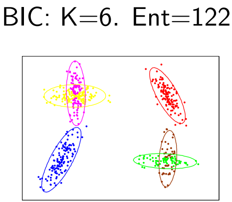
Simulated Example

Simulated Example



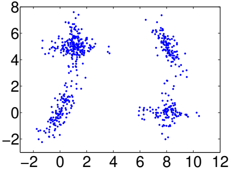
Simulated Example



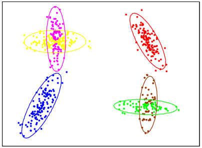


Simulated Example

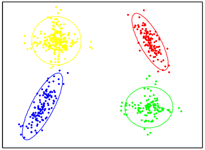
Simulated data



BIC: K=6. Ent=122

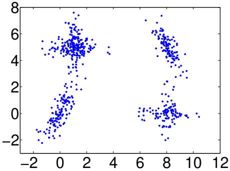


ICL: K=4. Ent=3

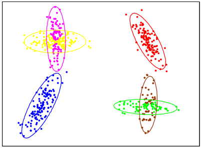


Simulated Example

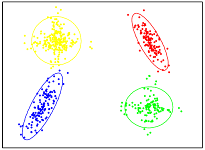
Simulated data



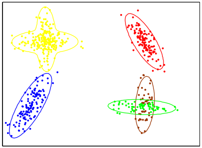
BIC: K=6. Ent=122



ICL: K=4. Ent=3

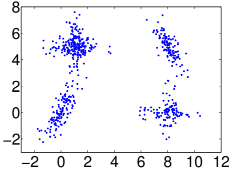


Combined: K=5. Ent=41

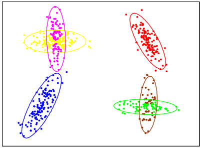


Simulated Example

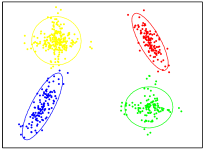
Simulated data



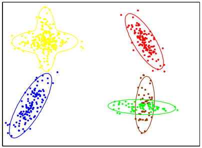
BIC: K=6. Ent=122



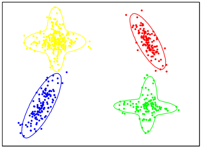
ICL: K=4. Ent=3



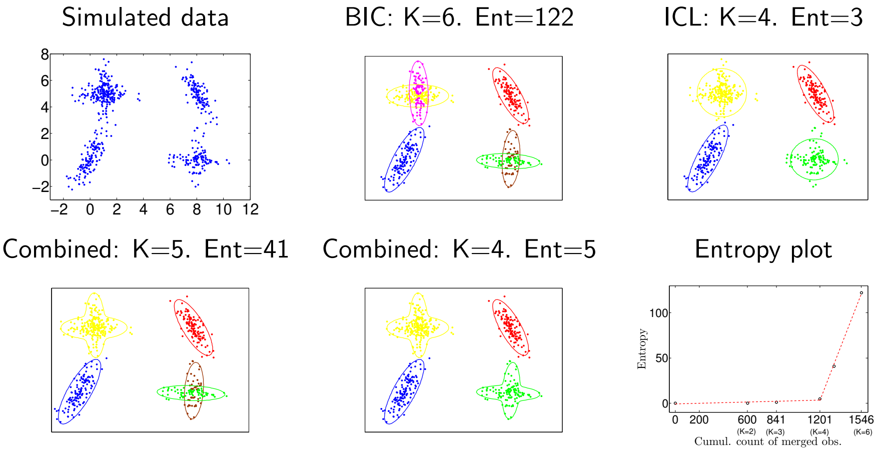
Combined: K=5. Ent=41



Combined: K=4. Ent=5



Simulated Example



Flow Cytometry Data

(Brinkman et al 2007; Lo et al 2008)

Flow Cytometry Data

(Brinkman et al 2007; Lo et al 2008)

- 9,083 cells from a graft-versus-host-disease (GvHD) patient

Flow Cytometry Data

(Brinkman et al 2007; Lo et al 2008)

- 9,083 cells from a graft-versus-host-disease (GvHD) patient
  - 4 biomarkers: CD4, CD8β, CD3, CD8

Flow Cytometry Data

(Brinkman et al 2007; Lo et al 2008)

- 9,083 cells from a graft-versus-host-disease (GvHD) patient
  - 4 biomarkers: CD4, CD8β, CD3, CD8
  - Goal: Find CD3+ CD4+ CD8β+ cell sub-populations

Flow Cytometry Data

(Brinkman et al 2007; Lo et al 2008)

- 9,083 cells from a graft-versus-host-disease (GvHD) patient
  - 4 biomarkers: CD4, CD8β, CD3, CD8
  - Goal: Find CD3+ CD4+ CD8β+ cell sub-populations
  - Clusters labeled CD3+ if mean of CD3 is $>$280.

Flow Cytometry Data

(Brinkman et al 2007; Lo et al 2008)

- 9,083 cells from a graft-versus-host-disease (GvHD) patient
  - 4 biomarkers: CD4, CD8β, CD3, CD8
  - Goal: Find CD3+ CD4+ CD8β+ cell sub-populations
  - Clusters labeled CD3+ if mean of CD3 is $>$280.
- ICL chose 9 clusters, of which 5 were CD3+.

Flow Cytometry Data

(Brinkman et al 2007; Lo et al 2008)

- 9,083 cells from a graft-versus-host-disease (GvHD) patient
  - 4 biomarkers: CD4, CD8β, CD3, CD8
  - Goal: Find CD3+ CD4+ CD8β+ cell sub-populations
  - Clusters labeled CD3+ if mean of CD3 is $>$280.
- ICL chose 9 clusters, of which 5 were CD3+.
  - Major CD3+ CD4+ CD8β- region lumped in with CD3- $\Longrightarrow$ not good

Flow Cytometry Data

(Brinkman et al 2007; Lo et al 2008)

- 9,083 cells from a graft-versus-host-disease (GvHD) patient
  - 4 biomarkers: CD4, CD8β, CD3, CD8
  - Goal: Find CD3+ CD4+ CD8β+ cell sub-populations
  - Clusters labeled CD3+ if mean of CD3 is $>$280.
- ICL chose 9 clusters, of which 5 were CD3+.
  - Major CD3+ CD4+ CD8β- region lumped in with CD3- $\Longrightarrow$ not good
- BIC chose 12 components, of which 6 were CD3+.

Flow Cytometry Data

(Brinkman et al 2007; Lo et al 2008)

- 9,083 cells from a graft-versus-host-disease (GvHD) patient
  - 4 biomarkers: CD4, CD8β, CD3, CD8
  - Goal: Find CD3+ CD4+ CD8β+ cell sub-populations
  - Clusters labeled CD3+ if mean of CD3 is $>$280.
- ICL chose 9 clusters, of which 5 were CD3+.
  - Major CD3+ CD4+ CD8β- region lumped in with CD3- $\Longrightarrow$ not good
- BIC chose 12 components, of which 6 were CD3+.
  - Known CD4+ CD8β+ region corresponds to cyan, green, red components.

Flow Cytometry Data

(Brinkman et al 2007; Lo et al 2008)

- 9,083 cells from a graft-versus-host-disease (GvHD) patient
  - 4 biomarkers: CD4, CD8β, CD3, CD8
  - Goal: Find CD3+ CD4+ CD8β+ cell sub-populations
  - Clusters labeled CD3+ if mean of CD3 is $>$280.
- ICL chose 9 clusters, of which 5 were CD3+.
  - Major CD3+ CD4+ CD8β- region lumped in with CD3- $\Longrightarrow$ not good
- BIC chose 12 components, of which 6 were CD3+.
  - Known CD4+ CD8β+ region corresponds to cyan, green, red components.
  - First 3 mergings (down to 9 clusters) make biological sense

Flow Cytometry Data

(Brinkman et al 2007; Lo et al 2008)

- 9,083 cells from a graft-versus-host-disease (GvHD) patient
  - 4 biomarkers: CD4, CD8β, CD3, CD8
  - Goal: Find CD3+ CD4+ CD8β+ cell sub-populations
  - Clusters labeled CD3+ if mean of CD3 is $>$280.
- ICL chose 9 clusters, of which 5 were CD3+.
  - Major CD3+ CD4+ CD8β- region lumped in with CD3- $\Longrightarrow$ not good
- BIC chose 12 components, of which 6 were CD3+.
  - Known CD4+ CD8β+ region corresponds to cyan, green, red components.
  - First 3 mergings (down to 9 clusters) make biological sense
  - 4th merging (to 8 clusters) doesn't

Flow Cytometry Data

(Brinkman et al 2007; Lo et al 2008)

- 9,083 cells from a graft-versus-host-disease (GvHD) patient
  - 4 biomarkers: CD4, CD8β, CD3, CD8
  - Goal: Find CD3+ CD4+ CD8β+ cell sub-populations
  - Clusters labeled CD3+ if mean of CD3 is $>$280.
- ICL chose 9 clusters, of which 5 were CD3+.
  - Major CD3+ CD4+ CD8β- region lumped in with CD3- $\Longrightarrow$ not good
- BIC chose 12 components, of which 6 were CD3+.
  - Known CD4+ CD8β+ region corresponds to cyan, green, red components.
  - First 3 mergings (down to 9 clusters) make biological sense
  - 4th merging (to 8 clusters) doesn't
  - $\Longrightarrow$ substantively choose 9 clusters

Flow Cytometry Data

(Brinkman et al 2007; Lo et al 2008)

- 9,083 cells from a graft-versus-host-disease (GvHD) patient
  - 4 biomarkers: CD4, CD8β, CD3, CD8
  - Goal: Find CD3+ CD4+ CD8β+ cell sub-populations
  - Clusters labeled CD3+ if mean of CD3 is $>$280.
- ICL chose 9 clusters, of which 5 were CD3+.
  - Major CD3+ CD4+ CD8β- region lumped in with CD3- $\Longrightarrow$ not good
- BIC chose 12 components, of which 6 were CD3+.
  - Known CD4+ CD8β+ region corresponds to cyan, green, red components.
  - First 3 mergings (down to 9 clusters) make biological sense
  - 4th merging (to 8 clusters) doesn't
  - $\Longrightarrow$ substantively choose 9 clusters
  - retains the 6 important CD3+ cell sub-populations

Flow Cytometry Data

(Brinkman et al 2007; Lo et al 2008)

- 9,083 cells from a graft-versus-host-disease (GvHD) patient
  - 4 biomarkers: CD4, CD8β, CD3, CD8
  - Goal: Find CD3+ CD4+ CD8β+ cell sub-populations
  - Clusters labeled CD3+ if mean of CD3 is $>$280.
- ICL chose 9 clusters, of which 5 were CD3+.
  - Major CD3+ CD4+ CD8β- region lumped in with CD3- $\Longrightarrow$ not good
- BIC chose 12 components, of which 6 were CD3+.
  - Known CD4+ CD8β+ region corresponds to cyan, green, red components.
  - First 3 mergings (down to 9 clusters) make biological sense
  - 4th merging (to 8 clusters) doesn't
  - $\Longrightarrow$ substantively choose 9 clusters
  - retains the 6 important CD3+ cell sub-populations
- Entropy plot also has elbow at 9 clusters

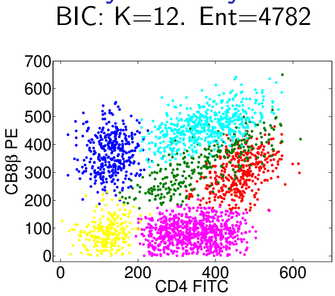
Flow Cytometry Data

(Brinkman et al 2007; Lo et al 2008)

- 9,083 cells from a graft-versus-host-disease (GvHD) patient
  - 4 biomarkers: CD4, CD8β, CD3, CD8
  - Goal: Find CD3+ CD4+ CD8β+ cell sub-populations
  - Clusters labeled CD3+ if mean of CD3 is $>$280.
- ICL chose 9 clusters, of which 5 were CD3+.
  - Major CD3+ CD4+ CD8β- region lumped in with CD3- $\Longrightarrow$ not good
- BIC chose 12 components, of which 6 were CD3+.
  - Known CD4+ CD8β+ region corresponds to cyan, green, red components.
  - First 3 mergings (down to 9 clusters) make biological sense
  - 4th merging (to 8 clusters) doesn't
  - $\Longrightarrow$ substantively choose 9 clusters
  - retains the 6 important CD3+ cell sub-populations
- Entropy plot also has elbow at 9 clusters
  - $\Longrightarrow$ statistical method recovers substantive result

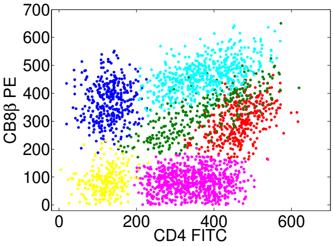
Flow Cytometry Data: Results for CD3+ Clusters

Flow Cytometry Data: Results for CD3+ Clusters

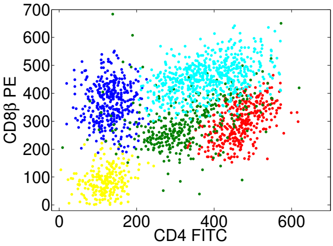


Flow Cytometry Data: Results for CD3+ Clusters

BIC: K=12. Ent=4782

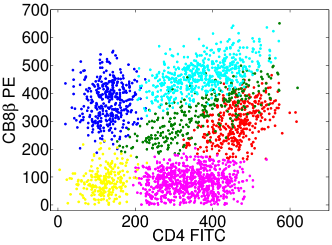


ICL: K=9. Ent=3235

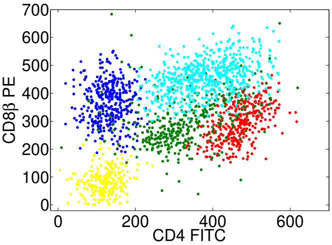


Flow Cytometry Data: Results for CD3+ Clusters

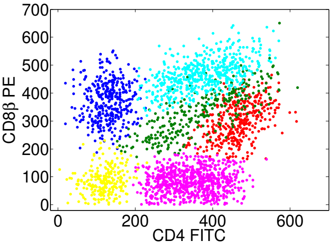
BIC: K=12. Ent=4782



ICL: K=9. Ent=3235

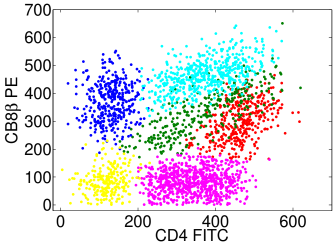


Combined: K=9. Ent=1478

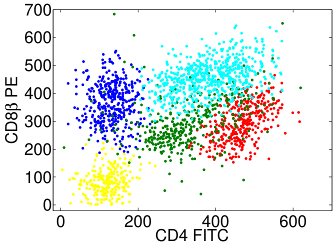


Flow Cytometry Data: Results for CD3+ Clusters

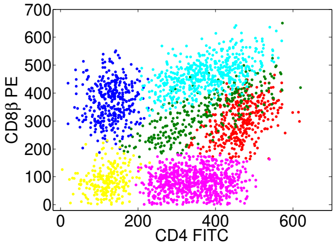
BIC: K=12. Ent=4782



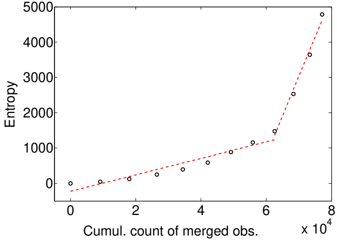
ICL: K=9. Ent=3235



Combined: K=9. Ent=1478



Entropy plot



Summary

Summary

- Model-based clustering with the number of mixture components, K, chosen by BIC, gives a good fit to data

Summary

- Model-based clustering with the number of mixture components, K, chosen by BIC, gives a good fit to data
  - But it can overstate the number of clusters because a non-Gaussian cluster can be represented by more than one mixture component

Summary

- Model-based clustering with the number of mixture components, K, chosen by BIC, gives a good fit to data
  - But it can overstate the number of clusters because a non-Gaussian cluster can be represented by more than one mixture component
- We propose a method for merging mixture components into clusters, by maximizing the change in entropy at each stage

Summary

- Model-based clustering with the number of mixture components, K, chosen by BIC, gives a good fit to data
  - But it can overstate the number of clusters because a non-Gaussian cluster can be represented by more than one mixture component
- We propose a method for merging mixture components into clusters, by maximizing the change in entropy at each stage
  - Yields a sequence of K soft clusterings

Summary

- Model-based clustering with the number of mixture components, K, chosen by BIC, gives a good fit to data
  - But it can overstate the number of clusters because a non-Gaussian cluster can be represented by more than one mixture component
- We propose a method for merging mixture components into clusters, by maximizing the change in entropy at each stage
  - Yields a sequence of K soft clusterings
  - User can choose between them substantively or using the entropy plot, or ICL

Summary

- Model-based clustering with the number of mixture components, K, chosen by BIC, gives a good fit to data
  - But it can overstate the number of clusters because a non-Gaussian cluster can be represented by more than one mixture component
- We propose a method for merging mixture components into clusters, by maximizing the change in entropy at each stage
  - Yields a sequence of K soft clusterings
  - User can choose between them substantively or using the entropy plot, or ICL
- Worked well in simulation experiments

Summary

- Model-based clustering with the number of mixture components, K, chosen by BIC, gives a good fit to data
  - But it can overstate the number of clusters because a non-Gaussian cluster can be represented by more than one mixture component
- We propose a method for merging mixture components into clusters, by maximizing the change in entropy at each stage
  - Yields a sequence of K soft clusterings
  - User can choose between them substantively or using the entropy plot, or ICL
- Worked well in simulation experiments
- Found a biologically satisfactory solution in the flow cytometry dataset

Summary

- Model-based clustering with the number of mixture components, K, chosen by BIC, gives a good fit to data
  - But it can overstate the number of clusters because a non-Gaussian cluster can be represented by more than one mixture component
- We propose a method for merging mixture components into clusters, by maximizing the change in entropy at each stage
  - Yields a sequence of K soft clusterings
  - User can choose between them substantively or using the entropy plot, or ICL
- Worked well in simulation experiments
- Found a biologically satisfactory solution in the flow cytometry dataset
- Paper is to appeared in the next issue of *Journal of Computational and Graphical Statistics*

Summary

- Model-based clustering with the number of mixture components, K, chosen by BIC, gives a good fit to data
  - But it can overstate the number of clusters because a non-Gaussian cluster can be represented by more than one mixture component
- We propose a method for merging mixture components into clusters, by maximizing the change in entropy at each stage
  - Yields a sequence of K soft clusterings
  - User can choose between them substantively or using the entropy plot, or ICL
- Worked well in simulation experiments
- Found a biologically satisfactory solution in the flow cytometry dataset
- Paper is to appeared in the next issue of *Journal of Computational and Graphical Statistics*
- All the described material is available in the MIXMOD software http://www.mixmod.org