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:

\[ y_i \sim \sum_{g=1}^{G} \tau_g \text{MVN}_d(\mu_g, \Sigma_g), \]
Basic Ideas of Model-Based Clustering

Based on a finite mixture of multivariate normal distributions:

\[ y_i \sim \sum_{g=1}^{G} \tau_g \text{MVN}_d(\mu_g, \Sigma_g), \]

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:

\[ y_i \sim \sum_{g=1}^{G} \tau_g \text{MVN}_d(\mu_g, \Sigma_g), \]

- where \( \Sigma_g = \lambda_g D_g A_g D_g^T \)

- \( \lambda_g = \) determinant of \( \Sigma_g \): controls the \textit{volume} of the \( g \)th cluster
Basic Ideas of Model-Based Clustering

- Based on a finite mixture of multivariate normal distributions:

\[ y_i \sim \sum_{g=1}^{G} \tau_g \text{MVN}_d(\mu_g, \Sigma_g), \]

- where \( \Sigma_g = \lambda_g D_g A_g D_g^T \)
- \( \lambda_g = \) determinant of \( \Sigma_g \): controls the volume of the \( g \)th cluster
- \( A_g = \text{diag}\{1, \alpha_{2g}, \ldots, \alpha_{dg}\} \)
Basic Ideas of Model-Based Clustering

Based on a finite mixture of multivariate normal distributions:

\[ y_i \sim \sum_{g=1}^{G} \tau_g \text{MVN}_d(\mu_g, \Sigma_g), \]

where \( \Sigma_g = \lambda_g D_g A_g D_g^T \)

\( \lambda_g \) = determinant of \( \Sigma_g \): controls the volume of the \( g \)th cluster

\( A_g = \text{diag}\{1, \alpha_{2g}, \ldots, \alpha_{dg}\} \)

- controls the shape of the \( g \)th cluster
Basic Ideas of Model-Based Clustering

- Based on a finite mixture of multivariate normal distributions:

\[ y_i \sim \sum_{g=1}^{G} \tau_g \text{MVN}_d(\mu_g, \Sigma_g), \]

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

E.g. \( \alpha_2 \) close to zero: Cluster \( g \) concentrated about a line.
E.g. \( \alpha_2, \ldots, \alpha_d \) all close to 1: Cluster \( g \) nearly spherical.
Basic Ideas of Model-Based Clustering

- Based on a finite mixture of multivariate normal distributions:

\[ y_i \sim \sum_{g=1}^{G} \tau_g \text{MVN}_d(\mu_g, \Sigma_g), \]

- where \( \Sigma_g = \lambda_g D_g A_g D_g^T \)
- \( \lambda_g = \) determinant of \( \Sigma_g \): controls the volume of the \( g \)th cluster
- \( A_g = \text{diag}\{1, \alpha_{2g}, \ldots, \alpha_{dg}\} \)
  - controls the shape of the \( g \)th cluster
  - \( (1 \geq \alpha_2 \geq \ldots \geq \alpha_d > 0) \)
  - E.g. \( \alpha_2 \) close to zero: Cluster \( g \) concentrated about a line.
Basic Ideas of Model-Based Clustering

- Based on a finite mixture of multivariate normal distributions:

\[ y_i \sim \sum_{g=1}^{G} \tau_g \text{MVN}_d(\mu_g, \Sigma_g), \]

where \( \Sigma_g = \lambda_g D_g A_g D_g^T \)

- \( \lambda_g \) = determinant of \( \Sigma_g \): controls the volume of the \( g \)th cluster
- \( A_g = \text{diag}\{1, \alpha_{2g}, \ldots, \alpha_{dg}\} \)
  - controls the shape of the \( g \)th cluster
  - \( 1 \geq \alpha_2 \geq \ldots \geq \alpha_d > 0 \)
  - E.g. \( \alpha_2 \) 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:

\[ y_i \sim \sum_{g=1}^{G} \tau_g \text{MVN}_d(\mu_g, \Sigma_g), \]

- \( \Sigma_g = \lambda_g D_g A_g D_g^T \)
- \( \lambda_g = \) determinant of \( \Sigma_g \): controls the \textit{volume} of the \( g \)th cluster
- \( A_g = \text{diag}\{1, \alpha_{2g}, \ldots, \alpha_{dg}\} \)
  - controls the \textit{shape} of the \( g \)th cluster
  - \((1 \geq \alpha_2 \geq \ldots \geq \alpha_d > 0)\)
  - E.g. \( \alpha_2 \) 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 \textit{orientation} of the \( g \)th cluster
Basic Ideas of Model-Based Clustering

- Based on a finite mixture of multivariate normal distributions:

\[ y_i \sim \sum_{g=1}^{G} \tau_g \text{MVN}_d(\mu_g, \Sigma_g), \]

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

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 random 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

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

This is consistent for the number of components (Keribin 2000), and also provides consistent density estimates (Roeder and Wasserman 1997).
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 random 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 $BIC = 2 \log \text{maximized likelihood} - (\# \text{parameters}) \log(n)$.
This is consistent for the number of components (Keribin 2000), and also provides consistent density estimates (Roeder and Wasserman 1997).
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 random 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 random 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

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

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

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

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

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

- 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)

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

<table>
<thead>
<tr>
<th>Expt.</th>
<th>BIC</th>
<th>Stephens</th>
<th>AIC</th>
<th>ICL</th>
<th>UIP</th>
<th>DIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.19</td>
<td>0.21</td>
<td>0.22</td>
<td>0.19</td>
<td>0.23</td>
<td>0.67</td>
</tr>
<tr>
<td>2</td>
<td>0.21</td>
<td>0.24</td>
<td>0.33</td>
<td>0.21</td>
<td>0.31</td>
<td>0.65</td>
</tr>
<tr>
<td>3</td>
<td>0.35</td>
<td>0.35</td>
<td>0.41</td>
<td>0.35</td>
<td>0.50</td>
<td>1.32</td>
</tr>
<tr>
<td>4</td>
<td>0.48</td>
<td>0.51</td>
<td>1.30</td>
<td>0.48</td>
<td>1.35</td>
<td>2.24</td>
</tr>
<tr>
<td>5</td>
<td>0.60</td>
<td>1.00</td>
<td>1.58</td>
<td>0.60</td>
<td>2.75</td>
<td>3.20</td>
</tr>
<tr>
<td>6</td>
<td>1.53</td>
<td>1.13</td>
<td>0.86</td>
<td>2.31</td>
<td>0.77</td>
<td>0.76</td>
</tr>
<tr>
<td>7</td>
<td>0.23</td>
<td>0.24</td>
<td>0.23</td>
<td>2.18</td>
<td>0.25</td>
<td>0.28</td>
</tr>
<tr>
<td>8</td>
<td>0.55</td>
<td>0.39</td>
<td>0.37</td>
<td>2.45</td>
<td>0.42</td>
<td>0.61</td>
</tr>
<tr>
<td>9</td>
<td>0.37</td>
<td>0.75</td>
<td>0.47</td>
<td>0.61</td>
<td>0.58</td>
<td>0.77</td>
</tr>
<tr>
<td>10</td>
<td>0.34</td>
<td>0.44</td>
<td>0.39</td>
<td>0.34</td>
<td>0.75</td>
<td>0.58</td>
</tr>
<tr>
<td>Mean</td>
<td>0.48</td>
<td>0.53</td>
<td>0.62</td>
<td>0.97</td>
<td>0.79</td>
<td>1.11</td>
</tr>
</tbody>
</table>
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,
  \[
  \log p(x|K) = \int p(x|K, \theta_K) \pi(\theta_K) d\theta_K,
  \]
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,

\[
\log p(x|K) = \int p(x|K, \theta_K)\pi(\theta_K)d\theta_K,
\]
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,

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

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

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

\[
\log p(x|K) = \int p(x|K, \theta_K) \pi(\theta_K) d\theta_K,
\]

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

\[
ICL(K) = \log p(x, z | K) = \int_{\Theta_K} p(x, z | K, \theta) \pi(\theta | K) d\theta
\approx \log p(x, \hat{z} | K, \hat{\theta}_K) - \frac{\nu_K}{2} \log n
\]

(Biernacki, Celeux & Govaert 2000)
ICL and Entropy

ICL(K) ≈ BIC(K) − the mean entropy, Ent(K),

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} is the conditional probability that x_i is from the 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).
ICL and Entropy

- $ICL(K) \approx BIC(K)$ — the mean entropy, $Ent(K)$,
ICL and Entropy

- ICL(K) ≈ BIC(K) — the mean entropy, Ent(K),
  
  \[ \text{Ent}(K) = -\sum_{k=1}^{K} \sum_{i=1}^{n} t_{ik}(\hat{\theta}_K) \log t_{ik}(\hat{\theta}_K) \geq 0 \]
ICL and Entropy

- $\text{ICL}(K) \approx \text{BIC}(K) - \text{the mean entropy, Ent}(K)$,
- $\text{Ent}(K) = - \sum_{k=1}^{K} \sum_{i=1}^{n} t_{ik}(\hat{\theta}_K) \log t_{ik}(\hat{\theta}_K) \geq 0$
- where $t_{ik} = \text{conditional probability that } x_i \text{ is from } k\text{th mixture component}
ICL and Entropy

- ICL(K) \approx BIC(K) – the mean entropy, Ent(K),
  - 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 \( x_i \) is from \( k \)th mixture component
  - Thus ICL tends to find smaller \( K \) than BIC
ICL and Entropy

- ICL(K) \approx BIC(K) – the mean entropy, Ent(K),
  \[ \text{Ent}(K) = - \sum_{k=1}^{K} \sum_{i=1}^{n} t_{ik}(\hat{\theta}_K) \log t_{ik}(\hat{\theta}_K) \geq 0 \]
  where \( t_{ik} \) = conditional probability that \( x_i \) is from \( k \)th 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) ≈ BIC(K) — the mean entropy, Ent(K),
  - \( \text{Ent}(K) = -\sum_{k=1}^{K} \sum_{i=1}^{n} t_{ik}(\hat{\theta}_K) \log t_{ik}(\hat{\theta}_K) \geq 0 \)
  - where \( t_{ik} \) = conditional probability that \( x_i \) is from \( k \)th 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) \text{ – the mean entropy, } \text{Ent}(K),
  - \text{Ent}(K) = - \sum_{k=1}^{K} \sum_{i=1}^{n} t_{ik}(\hat{\theta}_K) \log t_{ik}(\hat{\theta}_K) \geq 0
  - \text{where } t_{ik} = \text{conditional probability that } x_i \text{ is from } k\text{th mixture component}
  - \text{Thus ICL tends to find smaller } K \text{ 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) ≈ BIC(K) − the mean entropy, Ent(K),

- \( \text{Ent}(K) = - \sum_{k=1}^K \sum_{i=1}^n t_{ik}(\hat{\theta}_K) \log t_{ik}(\hat{\theta}_K) \geq 0 \)
- where \( t_{ik} = \) conditional probability that \( x_i \) is from \( k \)th 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

Start with a mixture model that fits the data well, with $K$ chosen by BIC. Design a sequence of soft clusterings with $K$, $K-1$, ... , 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).
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

---

Model-based clustering
BIC and ICL
Combining Components
Results
Summary
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 $K$, $K-1$, ..., 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 $K$, $K - 1$, . . . , 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 $K$, $K - 1$, ..., 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 data
Simulated Example

Simulated data

BIC: $K=6$. Ent=122
Simulated Example

Simulated data

BIC: $K=6$. $\text{Ent}=122$

ICL: $K=4$. $\text{Ent}=3$

Combined: $K=5$. $\text{Ent}=41$

Cumulative count of merged observations

Entropy plot

(K=3) (K=4) (K=6) (K=2)
Simulated Example

- Simulated data
- BIC: K=6. Ent=122
- ICL: K=4. Ent=3
- Combined: K=5. Ent=41


diagram with scatter plots and clusters

Entropy plot

Cumul. count of merged obs.

(K=3) (K=4) (K=6)(K=2)
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

**Simulated data**

- BIC: $K=6$. Ent=122
- ICL: $K=4$. Ent=3

**Combined**

- Combined: $K=5$. Ent=41
- Combined: $K=4$. Ent=5

**Entropy plot**

- Cumul. count of merged obs.
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

Results

ICL chose 9 clusters, of which 5 were CD3+. Major CD3+ CD4+ CD8β region lumped in with CD3- = 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 = substantively choose 9 clusters retains the 6 important CD3+ cell sub-populations.

Entropy plot also has elbow at 9 clusters = statistical method recovers substantive result.
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, CD8\(\beta\), CD3, CD8
  - Goal: Find CD3+ CD4+ CD8\(\beta\)+ 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- \(\Rightarrow\) 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- \Rightarrow 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- ⇒ 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- ⇒ 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- ⇒ 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- ⇒ 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
  - ⇒ 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− \(\Rightarrow\) 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
  - \(\Rightarrow\) 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- ⇒ 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
  - ⇒ 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- ⇒ 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
  - ⇒ substantively choose 9 clusters
  - retains the 6 important CD3+ cell sub-populations

- Entropy plot also has elbow at 9 clusters
  - ⇒ statistical method recovers substantive result
Flow Cytometry Data: Results for CD3+ Clusters
Flow Cytometry Data: Results for CD3+ Clusters

BIC: K=12. Ent=4782
Flow Cytometry Data: Results for CD3+ Clusters

BIC: $K=12$. Ent=$4782$

ICL: $K=9$. Ent=$3235$

Combined: $K=9$. Ent=$1478$

Entropy plot
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
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. The 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 appear in the next issue of the Journal of Computational and Graphical Statistics. All the described material is available in the mixmod software. http://www.mixmod.org
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
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.

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](http://www.mixmod.org).
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

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
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