Variable Selection for the Mixture Model Clustering Framework

Nema Dean
Department of Statistics, University of Glasgow
Joint work with Adrian Raftery

September 17, 2007

UNIVERSITY *of* GLASGOW

Outline

Outline

General Introduction to Clustering

- Classification involves categorizing subjects/items into predefined groups or looking at the different characteristics of the groups
  - Patients into healthy/unhealthy
  - Workers into blue/white collar
- Alternatively, sometimes we don't know what groups are in the data and want to find them
  - Data about patients with liver cancer: want to know if there are subtypes of the cancer
- If the data are only 2 or 3 dimensional we can plot the data and pick out clusters visually
- If data are higher dimensional we can't do this
- Clustering is an automatic, algorithmic method to do this

General Introduction to Clustering

- Classification involves categorizing subjects/items into predefined groups or looking at the different characteristics of the groups
  - Patients into healthy/unhealthy
  - Workers into blue/white collar
- Alternatively, sometimes we don't know what groups are in the data and want to find them
  - Data about patients with liver cancer: want to know if there are subtypes of the cancer
- If the data are only 2 or 3 dimensional we can plot the data and pick out clusters visually
- If data are higher dimensional we can't do this
- Clustering is an automatic, algorithmic method to do this

General Introduction to Clustering

- Classification involves categorizing subjects/items into predefined groups or looking at the different characteristics of the groups
  - Patients into healthy/unhealthy
  - Workers into blue/white collar
- Alternatively, sometimes we don't know what groups are in the data and want to find them
  - Data about patients with liver cancer: want to know if there are subtypes of the cancer
- If the data are only 2 or 3 dimensional we can plot the data and pick out clusters visually
- If data are higher dimensional we can't do this
- Clustering is an automatic, algorithmic method to do this

General Introduction to Clustering

- Classification involves categorizing subjects/items into predefined groups or looking at the different characteristics of the groups
  - Patients into healthy/unhealthy
  - Workers into blue/white collar
- Alternatively, sometimes we don't know what groups are in the data and want to find them
  - Data about patients with liver cancer: want to know if there are subtypes of the cancer
- If the data are only 2 or 3 dimensional we can plot the data and pick out clusters visually
- If data are higher dimensional we can't do this
- Clustering is an automatic, algorithmic method to do this

General Introduction to Clustering

- Classification involves categorizing subjects/items into predefined groups or looking at the different characteristics of the groups
  - Patients into healthy/unhealthy
  - Workers into blue/white collar
- Alternatively, sometimes we don't know what groups are in the data and want to find them
  - Data about patients with liver cancer: want to know if there are subtypes of the cancer
- If the data are only 2 or 3 dimensional we can plot the data and pick out clusters visually
- If data are higher dimensional we can't do this
- Clustering is an automatic, algorithmic method to do this

General Introduction to Clustering

- Classification involves categorizing subjects/items into predefined groups or looking at the different characteristics of the groups
  - Patients into healthy/unhealthy
  - Workers into blue/white collar
- Alternatively, sometimes we don't know what groups are in the data and want to find them
  - Data about patients with liver cancer: want to know if there are subtypes of the cancer
- If the data are only 2 or 3 dimensional we can plot the data and pick out clusters visually
- If data are higher dimensional we can't do this
- Clustering is an automatic, algorithmic method to do this

General Introduction to Clustering

- Classification involves categorizing subjects/items into predefined groups or looking at the different characteristics of the groups
  - Patients into healthy/unhealthy
  - Workers into blue/white collar
- Alternatively, sometimes we don't know what groups are in the data and want to find them
  - Data about patients with liver cancer: want to know if there are subtypes of the cancer
- If the data are only 2 or 3 dimensional we can plot the data and pick out clusters visually
- If data are higher dimensional we can't do this
- Clustering is an automatic, algorithmic method to do this

Different Clustering Methods

- How do we perform clustering?
- This depends on how we define our groups
- Could define a cost function and optimize over it (k means, hierarchical clustering)
- Could define a model for each cluster and fit it to the data (mixture model clustering)

Clustering/Classification Terminology

- General definition of clustering is: collection/classes of items more similar to others in their class than to items in other classes
- Group: “true” underlying partition or predefined classification
- Cluster: estimated partition

Outline

Mixture Models

- Mixture models are a simple method of extending single densities to a more flexible method of modeling data
- Instead of assuming data is modeled by a single density f we instead model it as a weighted sum of single densities

$$x \sim \sum_{k=1}^{K} \pi_k f_k \text{ where } 0 \leq \pi_k \leq 1, \sum_{k=1}^{K} \pi_k = 1$$

- Sometimes the single densities can be members of the same parametric family

$$x \sim \sum_{k=1}^{K} \pi_k f(\theta_k)$$

Mixture Models

- Mixture models are a simple method of extending single densities to a more flexible method of modeling data
- Instead of assuming data is modeled by a single density f we instead model it as a weighted sum of single densities

$$x \sim \sum_{k=1}^{K} \pi_k f_k \text{ where } 0 \leq \pi_k \leq 1, \sum_{k=1}^{K} \pi_k = 1$$

- Sometimes the single densities can be members of the same parametric family

$$x \sim \sum_{k=1}^{K} \pi_k f(\theta_k)$$

Mixture Model Clustering

- Simplest form of clustering involving mixture models: assume each group is modeled with its own density and the overall data is modeled as a weighted sum of these densities.
- The usual assumption for continuous data is that each group is distributed normally (model-based clustering).
- For discrete data we assume a multinomial or binomial distribution for each variable in each group with conditional independence between variables given the group membership (latent class analysis).
- If the true group shape is more complex more than one density will be needed to adequately model it.

Mixture Model Clustering

- Simplest form of clustering involving mixture models: assume each group is modeled with its own density and the overall data is modeled as a weighted sum of these densities.
- The usual assumption for continuous data is that each group is distributed normally (model-based clustering).
- For discrete data we assume a multinomial or binomial distribution for each variable in each group with conditional independence between variables given the group membership (latent class analysis).
- If the true group shape is more complex more than one density will be needed to adequately model it.

Mixture Model Clustering

- Simplest form of clustering involving mixture models: assume each group is modeled with its own density and the overall data is modeled as a weighted sum of these densities.
- The usual assumption for continuous data is that each group is distributed normally (model-based clustering).
- For discrete data we assume a multinomial or binomial distribution for each variable in each group with conditional independence between variables given the group membership (latent class analysis).
- If the true group shape is more complex more than one density will be needed to adequately model it.

Mixture Model Clustering

- Simplest form of clustering involving mixture models: assume each group is modeled with its own density and the overall data is modeled as a weighted sum of these densities.
- The usual assumption for continuous data is that each group is distributed normally (model-based clustering).
- For discrete data we assume a multinomial or binomial distribution for each variable in each group with conditional independence between variables given the group membership (latent class analysis).
- If the true group shape is more complex more than one density will be needed to adequately model it.

Outline

Selecting the Number of Clusters

- In addition to modeling group structure we also want to know how many clusters best model the data
- Because we are assuming a model for the clustering (that is defined by the number of clusters) we can use model selection techniques to decide the best model/number of clusters to fit to the data
- What we want: Bayes factor for model 1 versus model 2

$$B_{12} = \frac{p(Y \mid M_1)}{p(Y \mid M_2)}$$

  where $p(Y \mid M_1) = \int_\Theta f(Y \mid \theta, M_1) p(\theta \mid M_1) d\theta$ is the integrated likelihood for M_1.
- However, $p(Y \mid M_i)$ where M_i is a mixture model is not available in closed form

Selecting the Number of Clusters

- In addition to modeling group structure we also want to know how many clusters best model the data
- Because we are assuming a model for the clustering (that is defined by the number of clusters) we can use model selection techniques to decide the best model/number of clusters to fit to the data
- What we want: Bayes factor for model 1 versus model 2

$$B_{12} = \frac{p(Y \mid M_1)}{p(Y \mid M_2)}$$

where $p(Y \mid M_1) = \int_\Theta f(Y \mid \theta, M_1) p(\theta \mid M_1) d\theta$ is the integrated likelihood for M_1.

- However, $p(Y \mid M_i)$ where M_i is a mixture model is not available in closed form

Selecting the Number of Clusters

- In addition to modeling group structure we also want to know how many clusters best model the data
- Because we are assuming a model for the clustering (that is defined by the number of clusters) we can use model selection techniques to decide the best model/number of clusters to fit to the data
- What we want: Bayes factor for model 1 versus model 2

$$B_{12} = \frac{p(Y \mid M_1)}{p(Y \mid M_2)}$$

where $p(Y \mid M_1) = \int_\Theta f(Y \mid \theta, M_1) p(\theta \mid M_1) d\theta$ is the integrated likelihood for M_1.

- However, $p(Y \mid M_i)$ where M_i is a mixture model is not available in closed form

Selecting the Number of Clusters

- In addition to modeling group structure we also want to know how many clusters best model the data
- Because we are assuming a model for the clustering (that is defined by the number of clusters) we can use model selection techniques to decide the best model/number of clusters to fit to the data
- What we want: Bayes factor for model 1 versus model 2

$$B_{12} = \frac{p(Y \mid M_1)}{p(Y \mid M_2)}$$

where $p(Y \mid M_1) = \int_\Theta f(Y \mid \theta, M_1) p(\theta \mid M_1) d\theta$ is the integrated likelihood for M_1.

- However, $p(Y \mid M_i)$ where M_i is a mixture model is not available in closed form

Selecting the Number of Clusters

- We can approximate 2 times the log of the integrated likelihood $p(Y \mid M_i)$ by the fitted model's Bayesian Information Criterion (BIC) score where

  $$BIC(M) = 2 \times \log(\text{maximised likelihood of } M) - \nu \times \log(n)$$

  with ν being the number of independent parameters estimated in M and n being the number of observations in the data
- We can approximate the Bayes factor for model 1 versus model 2 by:

  $$2\log(B_{12}) \approx BIC(M_1) - BIC(M_2)$$

Selecting the Number of Clusters

- We can approximate 2 times the log of the integrated likelihood $p(Y \mid M_i)$ by the fitted model's Bayesian Information Criterion (BIC) score where

  $$BIC(M) = 2 \times \log(\text{maximised likelihood of } M) - \nu \times \log(n)$$

  with ν being the number of independent parameters estimated in M and n being the number of observations in the data
- We can approximate the Bayes factor for model 1 versus model 2 by:

  $$2\log(B_{12}) \approx BIC(M_1) - BIC(M_2)$$

How Good is BIC for Selecting the Number of Clusters?

- Keribin (2000) showed that under certain restrictions, BIC is consistent for estimating the number of mixture components for normal and poisson mixtures
- *However*, it was assumed that all variables in the data are mixture variables. There was no statement about consistency of BIC in the presence of noise variables.
- Rusakov and Geiger (2005) showed that for Latent Class Analysis, BIC is not consistent for model selection when there are noise variables present.

Outline

Model-Based Clustering

- Model-Based Clustering $\Rightarrow$ Mixture model with normally distributed components
- $f_g = f(\theta_g) = N(\mu_g, \Sigma_g)$
- Problem: Even with only 5 groups in 5 dimensions we have potentially 50 covariance parameters
- Need some way to restrict the model's covariances for more parsimonious clustering models
- Perform a spectral decomposition of the covariance matrices of the clusters and restrict elements of the decomposition to be the same across clusters

Model-Based Clustering

- Model-Based Clustering $\Rightarrow$ Mixture model with normally distributed components
- $f_g = f(\theta_g) = N(\mu_g, \Sigma_g)$
- Problem: Even with only 5 groups in 5 dimensions we have potentially 50 covariance parameters
- Need some way to restrict the model's covariances for more parsimonious clustering models
- Perform a spectral decomposition of the covariance matrices of the clusters and restrict elements of the decomposition to be the same across clusters

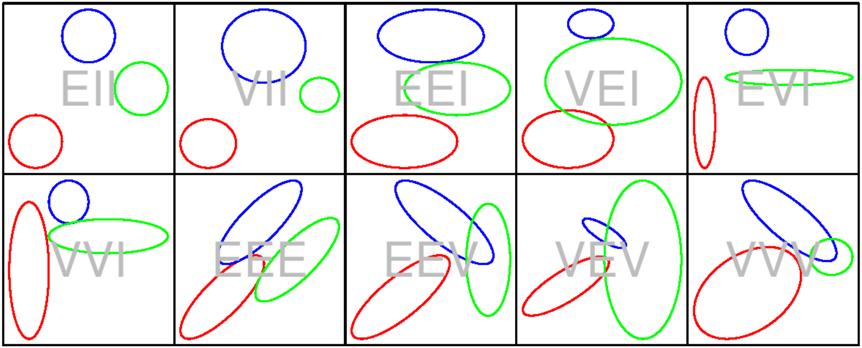
Model-Based Clustering

- Model-Based Clustering $\Rightarrow$ Mixture model with normally distributed components
- $f_g = f(\theta_g) = N(\mu_g, \Sigma_g)$
- Problem: Even with only 5 groups in 5 dimensions we have potentially 50 covariance parameters
- Need some way to restrict the model's covariances for more parsimonious clustering models
- Perform a spectral decomposition of the covariance matrices of the clusters and restrict elements of the decomposition to be the same across clusters

Model-Based Clustering

- Decompose the covariance matrices of the clusters $\Sigma_g = \lambda_g D_g A_g D_g^T$ such that:
  - λ_g is the largest eigenvalue of Σ_g controlling the volume of the g^{th} cluster
  - D_g is the matrix of eigenvectors of Σ_g controlling the orientation of the g^{th} cluster
  - A_g is the scaled diagonal matrix of eigenvalues of Σ_g controlling the shape of the g^{th} cluster
- We can restrict any of these elements across clusters to allow varying degrees of parsimony

Types of Cluster Constraints Available in mclust



Outline

Latent Class Analysis

- Latent Class Analysis $\Rightarrow$ Mixture model for discrete data
  - Variables are independent conditional on their class/cluster membership

$$\text{e.g. } X_i = (X_{i1}, \ldots, X_{id}), X_{ij} \perp X_{jk} \mid z_i = g, j \neq k$$

  - Each variable in each class is modeled with a multinomial distribution

$$\text{e.g. } f_g(x_{ij}) = \text{Mult}(p^j_{1g}, \ldots, p^j_{\ell_j g})$$

- Conditional Independence is necessary to give a parsimonious enough model to fit to data
- Idea: Any dependence in the data is modeled by the clustering

Latent Class Analysis

- Latent Class Analysis $\Rightarrow$ Mixture model for discrete data
  - Variables are independent conditional on their class/cluster membership

$$\text{e.g. } X_i = (X_{i1}, \ldots, X_{id}), X_{ij} \perp X_{jk} \mid z_i = g, j \neq k$$

  - Each variable in each class is modeled with a multinomial distribution

$$\text{e.g. } f_g(x_{ij}) = \text{Mult}(p^j_{1g}, \ldots, p^j_{\ell_j g})$$

- Conditional Independence is necessary to give a parsimonious enough model to fit to data
- Idea: Any dependence in the data is modeled by the clustering

Identification of Latent Class Models

- Problem: Have a limited amount of information and need to be able to check that there is enough information to fit models for certain numbers of clusters/classes
- Goodman (1978) provided a necessary condition for the identification of latent class models for G classes
- Say we have d variables with levels $(\ell_1, \ldots, \ell_d)$ and we wish to know if we can fit a G-class, latent class model to the data.

$$\text{Identifiable if: } \prod_{j=1}^{d} \ell_j - 1 > (\sum_{j=1}^{d} \ell_j - d)G + G - 1$$

$$\text{Equivalently if: } \prod_{j=1}^{d} \ell_j > (\sum_{j=1}^{d} \ell_j - d + 1)G$$

Identification of Latent Class Models

- Problem: Have a limited amount of information and need to be able to check that there is enough information to fit models for certain numbers of clusters/classes
- Goodman (1978) provided a necessary condition for the identification of latent class models for G classes
- Say we have d variables with levels $(\ell_1, \ldots, \ell_d)$ and we wish to know if we can fit a G-class, latent class model to the data.

$$\text{Identifiable if: } \prod_{j=1}^{d} \ell_j - 1 > (\sum_{j=1}^{d} \ell_j - d)G + G - 1$$

$$\text{Equivalently if: } \prod_{j=1}^{d} \ell_j > (\sum_{j=1}^{d} \ell_j - d + 1)G$$

Identification of Latent Class Models

- Problem: Have a limited amount of information and need to be able to check that there is enough information to fit models for certain numbers of clusters/classes
- Goodman (1978) provided a necessary condition for the identification of latent class models for G classes
- Say we have d variables with levels $(\ell_1, \ldots, \ell_d)$ and we wish to know if we can fit a G-class, latent class model to the data.

$$\text{Identifiable if: } \prod_{j=1}^{d} \ell_j - 1 > (\sum_{j=1}^{d} \ell_j - d)G + G - 1$$

$$\text{Equivalently if: } \prod_{j=1}^{d} \ell_j > (\sum_{j=1}^{d} \ell_j - d + 1)G$$

Outline

Why do Variable Selection?

- Both substantive and model selection issues
  - We may be as interested in which variables separate the clusters as the clusters found, e.g. medical settings, future datasets
  - As mentioned previously, BIC may not be consistent for choosing the number of clusters in the presence of noise variables

Why do Variable Selection?

- Both substantive and model selection issues
  - We may be as interested in which variables separate the clusters as the clusters found, e.g. medical settings, future datasets
  - As mentioned previously, BIC may not be consistent for choosing the number of clusters in the presence of noise variables

Why do Variable Selection?

- Both substantive and model selection issues
  - We may be as interested in which variables separate the clusters as the clusters found, e.g. medical settings, future datasets
  - As mentioned previously, BIC may not be consistent for choosing the number of clusters in the presence of noise variables

How do we do Variable Selection?

- If we knew the clustering we could use this to pick out the variables which best define the clustering
- If we knew the variables which best define the clustering we could use these to cluster the data
- We don't know either!
- We propose to iteratively estimate both.

How do we do Variable Selection?

- If we knew the clustering we could use this to pick out the variables which best define the clustering
- If we knew the variables which best define the clustering we could use these to cluster the data
- We don't know either!
- We propose to iteratively estimate both.

How do we do Variable Selection?

- If we knew the clustering we could use this to pick out the variables which best define the clustering
- If we knew the variables which best define the clustering we could use these to cluster the data
- We don't know either!
- We propose to iteratively estimate both.

How do we do Variable Selection?

- If we knew the clustering we could use this to pick out the variables which best define the clustering
- If we knew the variables which best define the clustering we could use these to cluster the data
- We don't know either!
- We propose to iteratively estimate both.

How do we do Variable Selection?

- First we propose two models for our current data, where we are examining one variable for its usefulness in clustering
- One model assumes that the variable is useful for clustering given the other current clustering variables
- The other model assumes that the variable is *not* useful for clustering given the other current clustering variables
- More formally, at each point in the procedure we can partition our data Y into 3 disjoint subsets $Y^{(clust)}$, $Y^{(?)}$ and $Y^{(other)}$ where
  - $Y^{(clust)}$ is the set of (other) currently selected clustering variables
  - $Y^{(?)}$ is the variable under consideration for inclusion (from $Y^{(other)}$) into/exclusion from $Y^{(clust)}$
  - $Y^{(other)}$ is the set of all other variables

How do we do Variable Selection?

- First we propose two models for our current data, where we are examining one variable for its usefulness in clustering
- One model assumes that the variable is useful for clustering given the other current clustering variables
- The other model assumes that the variable is *not* useful for clustering given the other current clustering variables
- More formally, at each point in the procedure we can partition our data Y into 3 disjoint subsets $Y^{(clust)}$, $Y^{(?)}$ and $Y^{(other)}$ where
  - $Y^{(clust)}$ is the set of (other) currently selected clustering variables
  - $Y^{(?)}$ is the variable under consideration for inclusion (from $Y^{(other)}$) into/exclusion from $Y^{(clust)}$
  - $Y^{(other)}$ is the set of all other variables

How do we do Variable Selection?

- First we propose two models for our current data, where we are examining one variable for its usefulness in clustering
- One model assumes that the variable is useful for clustering given the other current clustering variables
- The other model assumes that the variable is *not* useful for clustering given the other current clustering variables
- More formally, at each point in the procedure we can partition our data Y into 3 disjoint subsets $Y^{(clust)}$, $Y^{(?)}$ and $Y^{(other)}$ where
  - $Y^{(clust)}$ is the set of (other) currently selected clustering variables
  - $Y^{(?)}$ is the variable under consideration for inclusion (from $Y^{(other)}$) into/exclusion from $Y^{(clust)}$
  - $Y^{(other)}$ is the set of all other variables

How do we do Variable Selection?

- First we propose two models for our current data, where we are examining one variable for its usefulness in clustering
- One model assumes that the variable is useful for clustering given the other current clustering variables
- The other model assumes that the variable is *not* useful for clustering given the other current clustering variables
- More formally, at each point in the procedure we can partition our data Y into 3 disjoint subsets $Y^{(clust)}$, $Y^{(?)}$ and $Y^{(other)}$ where
  - $Y^{(clust)}$ is the set of (other) currently selected clustering variables
  - $Y^{(?)}$ is the variable under consideration for inclusion (from $Y^{(other)}$) into/exclusion from $Y^{(clust)}$
  - $Y^{(other)}$ is the set of all other variables

Variable Selection for Model-Based Clustering

$$
\begin{aligned}
p(Y \mid M_1) &= p(Y^{(clust)}, Y^{(?)}, Y^{(other)} \mid Z) \\
&= p(Y^{(other)} \mid Y^{(clust)}, Y^{(?)}) \\
&\times\ p(Y^{(?)} \mid Y^{(clust)}) p(Y^{(clust)} \mid Z)
\end{aligned}
$$

$$
\begin{aligned}
p(Y \mid M_2) &= p(Y^{(clust)}, Y^{(?)}, Y^{(other)} \mid Z) \\
&= p(Y^{(other)} \mid Y^{(clust)}, Y^{(?)}) \\
&\times\ p(Y^{(?)}, Y^{(clust)} \mid Z)
\end{aligned}
$$

Variable Selection for Model-Based Clustering

$$
\begin{aligned}
p(Y \mid M_1) &= p(Y^{(clust)}, Y^{(?)}, Y^{(other)} \mid Z) \\
&= p(Y^{(other)} \mid Y^{(clust)}, Y^{(?)}) \\
&\times\ p(Y^{(?)} \mid Y^{(clust)}) p(Y^{(clust)} \mid Z)
\end{aligned}
$$

$$
\begin{aligned}
p(Y \mid M_2) &= p(Y^{(clust)}, Y^{(?)}, Y^{(other)} \mid Z) \\
&= p(Y^{(other)} \mid Y^{(clust)}, Y^{(?)}) \\
&\times\ p(Y^{(?)}, Y^{(clust)} \mid Z)
\end{aligned}
$$

Variable Selection for Model-Based Clustering

$$
\begin{aligned}
p(Y \mid M_1) &= p(Y^{(clust)}, Y^{(?)}, Y^{(other)} \mid Z) \\
&= p(Y^{(other)} \mid Y^{(clust)}, Y^{(?)}) \\
&\times \; p(Y^{(?)} \mid Y^{(clust)}) p(Y^{(clust)} \mid Z)
\end{aligned}
$$

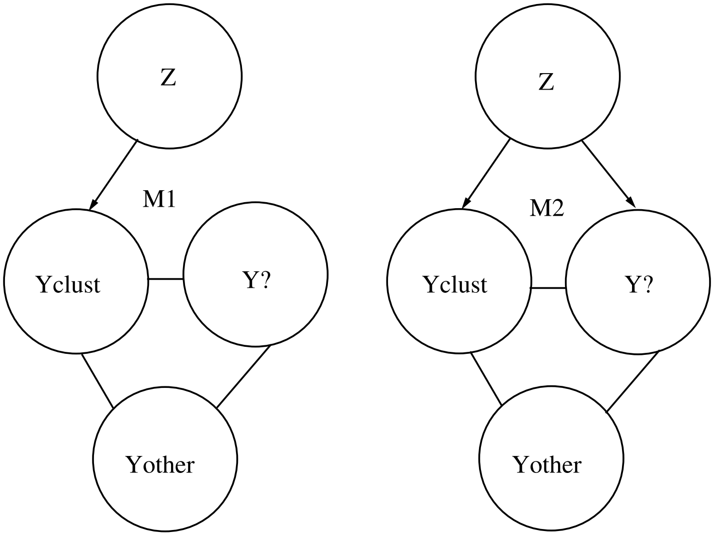
$$
\begin{aligned}
p(Y \mid M_2) &= p(Y^{(clust)}, Y^{(?)}, Y^{(other)} \mid Z) \\
&= p(Y^{(other)} \mid Y^{(clust)}, Y^{(?)}) \\
&\times \; p(Y^{(?)}, Y^{(clust)} \mid Z)
\end{aligned}
$$

Variable Selection for Model-Based Clustering

$$\begin{aligned} p(Y \mid M_1) &= p(Y^{(clust)}, Y^{(?)}, Y^{(other)} \mid Z) \\ &= p(Y^{(other)} \mid Y^{(clust)}, Y^{(?)}) \\ &\times \; p(Y^{(?)} \mid Y^{(clust)}) p(Y^{(clust)} \mid Z) \end{aligned}$$

$$\begin{aligned} p(Y \mid M_2) &= p(Y^{(clust)}, Y^{(?)}, Y^{(other)} \mid Z) \\ &= p(Y^{(other)} \mid Y^{(clust)}, Y^{(?)}) \\ &\times \; p(Y^{(?)}, Y^{(clust)} \mid Z) \end{aligned}$$

Variable Selection for Model-Based Clustering



Variable Selection for Latent Class Analysis

$$
\begin{aligned}
p(Y \mid M_1) &= p(Y^{(clust)}, Y^{(?)}, Y^{(other)} \mid Z) \\
&= p(Y^{(other)} \mid Y^{(clust)}, Y^{(?)}) \\
&\times \quad p(Y^{(?)}) \qquad p(Y^{(clust)} \mid Z)
\end{aligned}
$$

$$
\begin{aligned}
p(Y \mid M_2) &= p(Y^{(clust)}, Y^{(?)}, Y^{(other)} \mid Z) \\
&= p(Y^{(other)} \mid Y^{(clust)}, Y^{(?)}) \\
&\times \quad p(Y^{(?)}, Y^{(clust)} \mid Z)
\end{aligned}
$$

Variable Selection for Latent Class Analysis

$$
\begin{aligned}
p(Y \mid M_1) &= p(Y^{(clust)}, Y^{(?)}, Y^{(other)} \mid Z) \\
&= p(Y^{(other)} \mid Y^{(clust)}, Y^{(?)}) \\
&\times \; p(Y^{(?)}) \qquad p(Y^{(clust)} \mid Z)
\end{aligned}
$$

$$
\begin{aligned}
p(Y \mid M_2) &= p(Y^{(clust)}, Y^{(?)}, Y^{(other)} \mid Z) \\
&= p(Y^{(other)} \mid Y^{(clust)}, Y^{(?)}) \\
&\times \; p(Y^{(?)}, Y^{(clust)} \mid Z)
\end{aligned}
$$

Variable Selection for Latent Class Analysis

$$
\begin{aligned}
p(Y \mid M_1) &= p(Y^{(clust)}, Y^{(?)}, Y^{(other)} \mid Z) \\
&= p(Y^{(other)} \mid Y^{(clust)}, Y^{(?)}) \\
&\times\ p(Y^{(?)}) \qquad p(Y^{(clust)} \mid Z)
\end{aligned}
$$

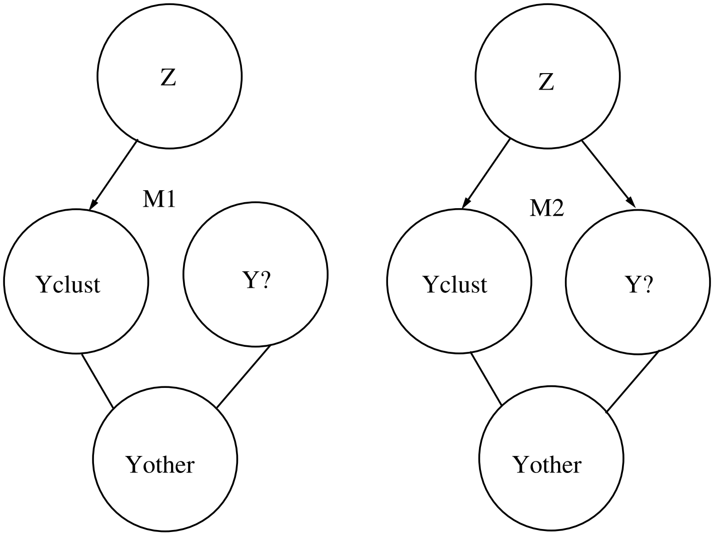
$$
\begin{aligned}
p(Y \mid M_2) &= p(Y^{(clust)}, Y^{(?)}, Y^{(other)} \mid Z) \\
&= p(Y^{(other)} \mid Y^{(clust)}, Y^{(?)}) \\
&\times\ p(Y^{(?)}, Y^{(clust)} \mid Z)
\end{aligned}
$$

Variable Selection for Latent Class Analysis

$$\begin{aligned} p(Y \mid M_1) &= p(Y^{(clust)}, Y^{(?)}, Y^{(other)} \mid Z) \\ &= p(Y^{(other)} \mid Y^{(clust)}, Y^{(?)}) \\ &\times \; p(Y^{(?)}) \qquad p(Y^{(clust)} \mid Z) \end{aligned}$$

$$\begin{aligned} p(Y \mid M_2) &= p(Y^{(clust)}, Y^{(?)}, Y^{(other)} \mid Z) \\ &= p(Y^{(other)} \mid Y^{(clust)}, Y^{(?)}) \\ &\times \; p(Y^{(?)}, Y^{(clust)} \mid Z) \end{aligned}$$

Variable Selection for Latent Class Analysis



Implementation of Variable Selection

- For Model-Based Clustering:
  - If $Y^{(?)}$ is a single variable

$$\Rightarrow \quad E(Y^{(?)} \mid Y^{(clust)}) = \alpha + Y^{(clust)}\beta$$
$$\Rightarrow \quad p(Y^{(?)} \mid Y^{(clust)}) = \text{regression model}$$

- For Latent Class Analysis:
  - If $Y^{(?)}$ is a single variable

$$\Rightarrow p(Y^{(?)} \mid Y^{(clust)}) = p(Y^{(?)}) = Mult(p_1, \ldots, p_\ell)$$

Implementation of Variable Selection

- For Model-Based Clustering:
  - If $Y^{(?)}$ is a single variable

$$\Rightarrow \quad E(Y^{(?)} \mid Y^{(clust)}) = \alpha + Y^{(clust)}\beta$$
$$\Rightarrow \quad p(Y^{(?)} \mid Y^{(clust)}) = \text{regression model}$$

- For Latent Class Analysis:
  - If $Y^{(?)}$ is a single variable

$$\Rightarrow p(Y^{(?)} \mid Y^{(clust)}) = p(Y^{(?)}) = Mult(p_1, \ldots, p_\ell)$$

Implementation of Variable Selection Models

- Given the partition and the two models we would like to make a decision based on the Bayes factor B_{21}.
- Recall: this is not available in closed form.
- Instead we use the BIC approximation

$$2 \log B_{21} \approx BIC(M_2) - BIC(M_1)$$

- With certain assumptions about the models' parameter priors each Bayes factor decomposes into separate mixture model and regression components.
- Thus each BIC is the sum of BICs for mixture models and possibly regression models.

Implementation of Variable Selection Models

- Given the partition and the two models we would like to make a decision based on the Bayes factor B_{21}.
- Recall: this is not available in closed form.
- Instead we use the BIC approximation

$$2 \log B_{21} \approx BIC(M_2) - BIC(M_1)$$

- With certain assumptions about the models' parameter priors each Bayes factor decomposes into separate mixture model and regression components.
- Thus each BIC is the sum of BICs for mixture models and possibly regression models.

Implementation of Variable Selection Models

- Given the partition and the two models we would like to make a decision based on the Bayes factor B_{21}.
- Recall: this is not available in closed form.
- Instead we use the BIC approximation

$$2 \log B_{21} \approx BIC(M_2) - BIC(M_1)$$

- With certain assumptions about the models' parameter priors each Bayes factor decomposes into separate mixture model and regression components.
- Thus each BIC is the sum of BICs for mixture models and possibly regression models.

Implementation of Variable Selection Models

- Given the partition and the two models we would like to make a decision based on the Bayes factor B_{21}.
- Recall: this is not available in closed form.
- Instead we use the BIC approximation

$$2 \log B_{21} \approx BIC(M_2) - BIC(M_1)$$

- With certain assumptions about the models' parameter priors each Bayes factor decomposes into separate mixture model and regression components.

$$\frac{p(Y \mid M_2)}{p(Y \mid M_1)} = \frac{p(Y^{(other)}, Y^{(?)}, Y^{(clust)} \mid M_2)}{p(Y^{(other)}, Y^{(?)}, Y^{(clust)} \mid M_1)}$$

- Thus each BIC is the sum of BICs for mixture models and possibly regression models.

Implementation of Variable Selection Models

- Given the partition and the two models we would like to make a decision based on the Bayes factor B_{21}.
- Recall: this is not available in closed form.
- Instead we use the BIC approximation

$$2 \log B_{21} \approx BIC(M_2) - BIC(M_1)$$

- With certain assumptions about the models' parameter priors each Bayes factor decomposes into separate mixture model and regression components.

$$\frac{p(Y \mid M_2)}{p(Y \mid M_1)} = \frac{p(Y^{(other)} \mid Y^{(?)}, Y^{(clust)})}{p(Y^{(other)} \mid Y^{(?)}, Y^{(clust)})} \times \frac{p(Y^{(?)}, Y^{(clust)} \mid M_2)}{p(Y^{(?)} \mid Y^{(clust)}, M_1) p(Y^{(clust)} \mid M_1)}$$

- Thus each BIC is the sum of BICs for mixture models and

Implementation of Variable Selection Models

- Given the partition and the two models we would like to make a decision based on the Bayes factor B_{21}.
- Recall: this is not available in closed form.
- Instead we use the BIC approximation

$$2 \log B_{21} \approx BIC(M_2) - BIC(M_1)$$

- With certain assumptions about the models' parameter priors each Bayes factor decomposes into separate mixture model and regression components.

$$\frac{p(Y \mid M_2)}{p(Y \mid M_1)} = \frac{p(Y^{(clust)}, Y^{(?)} \mid M_2)}{p(Y^{(?)} \mid Y^{(clust)}, M_1) p(Y^{(clust)} \mid M_1)}$$

- Thus each BIC is the sum of BICs for mixture models and possibly regression models.

Implementation of Variable Selection Models

- Given the partition and the two models we would like to make a decision based on the Bayes factor B_{21}.
- Recall: this is not available in closed form.
- Instead we use the BIC approximation

$$2 \log B_{21} \approx BIC(M_2) - BIC(M_1)$$

- With certain assumptions about the models' parameter priors each Bayes factor decomposes into separate mixture model and regression components.

$$\frac{p(Y \mid M_2)}{p(Y \mid M_1)} = \frac{p(Y^{(clust)}, Y^{(?)} \mid M_2)}{p(Y^{(?)} \mid M_1) p(Y^{(clust)} \mid M_1)}$$

- Thus each BIC is the sum of BICs for mixture models and possibly regression models.

Implementation of Variable Selection Models

- Given the partition and the two models we would like to make a decision based on the Bayes factor B_{21}.
- Recall: this is not available in closed form.
- Instead we use the BIC approximation

$$2 \log B_{21} \approx BIC(M_2) - BIC(M_1)$$

- With certain assumptions about the models' parameter priors each Bayes factor decomposes into separate mixture model and regression components.

- Thus each BIC is the sum of BICs for mixture models and possibly regression models.

Implementation of Variable Selection Models

$$BIC_{diff}(Y^{(?)}) = BIC_{clust}(Y^{(?)}) - BIC_{not\ clust}(Y^{(?)})$$

with

$$BIC_{clust}(Y^{(?)}) = BIC(p(Y^{(clust)}, Y^{(?)} \mid z))$$

MBC: $BIC_{not\ clust}(Y^{(?)}) = BIC(p(Y^{(?)} \mid Y^{(clust)})) + BIC(p(Y^{(clust)} \mid z))$

LCA: $BIC_{not\ clust}(Y^{(?)}) = BIC(p(Y^{(?)})) + BIC(p(Y^{(clust)} \mid z))$

where z are the (unknown) cluster memberships.

- When the BIC difference is positive this is taken as evidence for the variable $Y^{(?)}$'s usefulness in clustering
- When the BIC difference is negative this is taken as evidence against the variable $Y^{(?)}$'s usefulness in clustering

Implementation of Variable Selection Models

$$BIC_{diff}(Y^{(?)}) = BIC_{clust}(Y^{(?)}) - BIC_{not\ clust}(Y^{(?)})$$

with

$$BIC_{clust}(Y^{(?)}) = BIC(p(Y^{(clust)}, Y^{(?)} \mid z))$$

MBC: $BIC_{not\ clust}(Y^{(?)}) = BIC(p(Y^{(?)} \mid Y^{(clust)})) + BIC(p(Y^{(clust)} \mid z))$

LCA: $BIC_{not\ clust}(Y^{(?)}) = BIC(p(Y^{(?)})) + BIC(p(Y^{(clust)} \mid z))$

where z are the (unknown) cluster memberships.

- When the BIC difference is positive this is taken as evidence for the variable $Y^{(?)}$'s usefulness in clustering
- When the BIC difference is negative this is taken as evidence against the variable $Y^{(?)}$'s usefulness in clustering

Implementation of Variable Selection Models

$$BIC_{diff}(Y^{(?)}) = BIC_{clust}(Y^{(?)}) - BIC_{not\ clust}(Y^{(?)})$$

with

$$BIC_{clust}(Y^{(?)}) = BIC(p(Y^{(clust)}, Y^{(?)} \mid z))$$

MBC: $BIC_{not\ clust}(Y^{(?)}) = BIC(p(Y^{(?)} \mid Y^{(clust)})) + BIC(p(Y^{(clust)} \mid z))$

LCA: $BIC_{not\ clust}(Y^{(?)}) = BIC(p(Y^{(?)})) + BIC(p(Y^{(clust)} \mid z))$

where z are the (unknown) cluster memberships.

- When the BIC difference is positive this is taken as evidence for the variable $Y^{(?)}$'s usefulness in clustering
- When the BIC difference is negative this is taken as evidence against the variable $Y^{(?)}$'s usefulness in clustering

Implementation of Variable Selection Models

$$BIC_{diff}(Y^{(?)}) = BIC_{clust}(Y^{(?)}) - BIC_{not\ clust}(Y^{(?)})$$

with

$$BIC_{clust}(Y^{(?)}) = BIC(p(Y^{(clust)}, Y^{(?)} \mid z))$$

MBC: $BIC_{not\ clust}(Y^{(?)}) = BIC(p(Y^{(?)} \mid Y^{(clust)})) + BIC(p(Y^{(clust)} \mid z))$

LCA: $BIC_{not\ clust}(Y^{(?)}) = BIC(p(Y^{(?)})) + BIC(p(Y^{(clust)} \mid z))$

where z are the (unknown) cluster memberships.

- When the BIC difference is positive this is taken as evidence for the variable $Y^{(?)}$'s usefulness in clustering
- When the BIC difference is negative this is taken as evidence against the variable $Y^{(?)}$'s usefulness in clustering

Implementation of Variable Selection Models

$$BIC_{diff}(Y^{(?)}) = BIC_{clust}(Y^{(?)}) - BIC_{not\ clust}(Y^{(?)})$$

with

$$BIC_{clust}(Y^{(?)}) = BIC(p(Y^{(clust)}, Y^{(?)} \mid z))$$

MBC: $BIC_{not\ clust}(Y^{(?)}) = BIC(p(Y^{(?)} \mid Y^{(clust)})) + BIC(p(Y^{(clust)} \mid z))$

LCA: $BIC_{not\ clust}(Y^{(?)}) = BIC(p(Y^{(?)})) + BIC(p(Y^{(clust)} \mid z))$

where z are the (unknown) cluster memberships.

- When the BIC difference is positive this is taken as evidence for the variable $Y^{(?)}$'s usefulness in clustering
- When the BIC difference is negative this is taken as evidence against the variable $Y^{(?)}$'s usefulness in clustering

Implementation of Variable Selection Models

$$BIC_{diff}(Y^{(?)}) = BIC_{clust}(Y^{(?)}) - BIC_{not\ clust}(Y^{(?)})$$

with

$$BIC_{clust}(Y^{(?)}) = BIC(p(Y^{(clust)}, Y^{(?)} \mid z))$$

MBC: $BIC_{not\ clust}(Y^{(?)}) = BIC(p(Y^{(?)} \mid Y^{(clust)})) + BIC(p(Y^{(clust)} \mid z))$

LCA: $BIC_{not\ clust}(Y^{(?)}) = BIC(p(Y^{(?)})) + BIC(p(Y^{(clust)} \mid z))$

where z are the (unknown) cluster memberships.

- When the BIC difference is positive this is taken as evidence for the variable $Y^{(?)}$'s usefulness in clustering
- When the BIC difference is negative this is taken as evidence against the variable $Y^{(?)}$'s usefulness in clustering

Outline

General Search Algorithm

- In order to explore all of the model space (create different partitions of the variables to check) we need a search algorithm.
- Approach is to iterate inclusion and exclusion steps
  - Inclusion steps test new variables for inclusion into the set of clustering variables
  - Exclusion steps test variables currently in the set of clustering variables for exclusion from that set
- Regardless of the type of step, for the variable being looked at, we will always fit models M_1 and M_2 to the partition involving that variable and make decisions based on that.

General Search Algorithm

- In order to explore all of the model space (create different partitions of the variables to check) we need a search algorithm.
- Approach is to iterate inclusion and exclusion steps
  - Inclusion steps test new variables for inclusion into the set of clustering variables
  - Exclusion steps test variables currently in the set of clustering variables for exclusion from that set
- Regardless of the type of step, for the variable being looked at, we will always fit models M_1 and M_2 to the partition involving that variable and make decisions based on that.

General Search Algorithm

- In order to explore all of the model space (create different partitions of the variables to check) we need a search algorithm.
- Approach is to iterate inclusion and exclusion steps
  - Inclusion steps test new variables for inclusion into the set of clustering variables
  - Exclusion steps test variables currently in the set of clustering variables for exclusion from that set
- Regardless of the type of step, for the variable being looked at, we will always fit models M_1 and M_2 to the partition involving that variable and make decisions based on that.

General Search Algorithm

- In order to explore all of the model space (create different partitions of the variables to check) we need a search algorithm.
- Approach is to iterate inclusion and exclusion steps
  - Inclusion steps test new variables for inclusion into the set of clustering variables
  - Exclusion steps test variables currently in the set of clustering variables for exclusion from that set
- Regardless of the type of step, for the variable being looked at, we will always fit models M_1 and M_2 to the partition involving that variable and make decisions based on that.

Greedy Search Algorithm
Inclusion Step

- Basic idea:
  - Exhaustively check all other variables not currently included in the set of clustering variables singly for evidence of usefulness for clustering
  - Propose the variable with the strongest evidence of usefulness for clustering (variable with largest BIC difference between M_2 and M_1)
  - If $BIC_{diff} > 0$ include the proposed variable in the current set of clustering variables
  - If $BIC_{diff} < 0$ do not include any new variable in the current set of clustering variables

$$(Y^{(?)})^{k+1} \in (Y^{(other)})^{k}$$

$$(Y^{(clust)})^{k+1} = (Y^{(clust)})^{k}$$

Greedy Search Algorithm
Inclusion Step

- Basic idea:
  - Exhaustively check all other variables not currently included in the set of clustering variables singly for evidence of usefulness for clustering
  - Propose the variable with the strongest evidence of usefulness for clustering (variable with largest BIC difference between M_2 and M_1)
  - If $BIC_{diff} > 0$ include the proposed variable in the current set of clustering variables
  - If $BIC_{diff} < 0$ do not include any new variable in the current set of clustering variables

$$(Y^{(?)})^{k+1} \in (Y^{(other)})^{k}$$

$$(Y^{(clust)})^{k+1} = (Y^{(clust)})^{k}$$

Greedy Search Algorithm
Inclusion Step

- Basic idea:
  - Exhaustively check all other variables not currently included in the set of clustering variables singly for evidence of usefulness for clustering
  - Propose the variable with the strongest evidence of usefulness for clustering (variable with largest BIC difference between M_2 and M_1)
  - If $BIC_{diff} > 0$ include the proposed variable in the current set of clustering variables
  - If $BIC_{diff} < 0$ do not include any new variable in the current set of clustering variables

$$(Y^{(?)})^{k+1} \in (Y^{(other)})^{k}$$

$$(Y^{(clust)})^{k+1} = (Y^{(clust)})^{k}$$

Greedy Search Algorithm
Inclusion Step

- Basic idea:
  - Exhaustively check all other variables not currently included in the set of clustering variables singly for evidence of usefulness for clustering
  - Propose the variable with the strongest evidence of usefulness for clustering (variable with largest BIC difference between M_2 and M_1)
  - If $BIC_{diff} > 0$ include the proposed variable in the current set of clustering variables
  - If $BIC_{diff} < 0$ do not include any new variable in the current set of clustering variables

$$(Y^{(?)})^{k+1} \in (Y^{(other)})^{k}$$

$$(Y^{(clust)})^{k+1} = (Y^{(clust)})^{k}$$

Greedy Search Algorithm
Inclusion Step

- Basic idea:
  - Exhaustively check all other variables not currently included in the set of clustering variables singly for evidence of usefulness for clustering
  - Propose the variable with the strongest evidence of usefulness for clustering (variable with largest BIC difference between M_2 and M_1)
  - If $BIC_{diff} > 0$ include the proposed variable in the current set of clustering variables
  - If $BIC_{diff} < 0$ do not include any new variable in the current set of clustering variables

$$(Y^{(?)})^{k+1} \in (Y^{(other)})^k$$

$$(Y^{(clust)})^{k+1} = (Y^{(clust)})^k$$

Greedy Search Algorithm
Exclusion Step

- Basic idea:
  - Exhaustively check all variables currently included in the set of clustering variables singly for evidence of usefulness for clustering
  - Propose the variable with the weakest evidence of usefulness for clustering (variable with smallest BIC difference between M_2 and M_1)
  - If $BIC_{diff} < 0$ remove the proposed variable from the current set of clustering variables
  - If $BIC_{diff} > 0$ do not remove any variable from the current set of clustering variables

$$(Y^{(?)})^{k+1} \in (Y^{(clust)})^{k}$$

$$(Y^{(clust)})^{k+1} = (Y^{(clust)})^{k} / (Y^{(?)})^{k+1}$$

Greedy Search Algorithm

Exclusion Step

- Basic idea:
  - Exhaustively check all variables currently included in the set of clustering variables singly for evidence of usefulness for clustering
  - Propose the variable with the weakest evidence of usefulness for clustering (variable with smallest BIC difference between M_2 and M_1)
  - If $BIC_{diff} < 0$ remove the proposed variable from the current set of clustering variables
  - If $BIC_{diff} > 0$ do not remove any variable from the current set of clustering variables

$$(Y^{(?)})^{k+1} \in (Y^{(clust)})^{k}$$

$$(Y^{(clust)})^{k+1} = (Y^{(clust)})^{k} / (Y^{(?)})^{k+1}$$

- Basic idea:
  - Exhaustively check all variables currently included in the set of clustering variables singly for evidence of usefulness for clustering
  - Propose the variable with the weakest evidence of usefulness for clustering (variable with smallest BIC difference between M_2 and M_1)
  - If $BIC_{diff} < 0$ remove the proposed variable from the current set of clustering variables
  - If $BIC_{diff} > 0$ do not remove any variable from the current set of clustering variables

$$(Y^{(?)})^{k+1} \in (Y^{(clust)})^{k}$$

$$(Y^{(clust)})^{k+1} = (Y^{(clust)})^{k} / (Y^{(?)})^{k+1}$$

Greedy Search Algorithm
Exclusion Step

- Basic idea:
  - Exhaustively check all variables currently included in the set of clustering variables singly for evidence of usefulness for clustering
  - Propose the variable with the weakest evidence of usefulness for clustering (variable with smallest BIC difference between M_2 and M_1)
  - If $BIC_{diff} < 0$ remove the proposed variable from the current set of clustering variables
  - If $BIC_{diff} > 0$ do not remove any variable from the current set of clustering variables

$$(Y^{(?)})^{k+1} \in (Y^{(clust)})^{k}$$

$$(Y^{(clust)})^{k+1} = (Y^{(clust)})^{k} / (Y^{(?)})^{k+1}$$

Greedy Search Algorithm

Exclusion Step

- Basic idea:
  - Exhaustively check all variables currently included in the set of clustering variables singly for evidence of usefulness for clustering
  - Propose the variable with the weakest evidence of usefulness for clustering (variable with smallest BIC difference between M_2 and M_1)
  - If $BIC_{diff} < 0$ remove the proposed variable from the current set of clustering variables
  - If $BIC_{diff} > 0$ do not remove any variable from the current set of clustering variables

$$(Y^{(?)})^{k+1} \in (Y^{(clust)})^{k}$$

$$(Y^{(clust)})^{k+1} = (Y^{(clust)})^{k} / (Y^{(?)})^{k+1}$$

Headlong Search Algorithm
Inclusion Step

- Basic idea - similar to Greedy Search:
  - Check, in order, each variable not currently included in the set of clustering variables for evidence of usefulness for clustering
  - Once a variable has $BIC_{diff} > upper$ include this variable in the current set of clustering variables and stop the inclusion step
  - If any variable checked has $BIC_{diff} < lower$ remove this variable from consideration for the rest of the search
  - If no variable has $BIC_{diff} > upper$ no variable is included in the current set of clustering variables

$$(Y^{(?)})^{k+1} \in (Y^{(other)})^k$$

$$(Y^{(clust)})^{k+1} = (Y^{(clust)})^k$$

Headlong Search Algorithm
Inclusion Step

- Basic idea - similar to Greedy Search:
  - Check, in order, each variable not currently included in the set of clustering variables for evidence of usefulness for clustering
  - Once a variable has $BIC_{diff} > upper$ include this variable in the current set of clustering variables and stop the inclusion step
  - If any variable checked has $BIC_{diff} < lower$ remove this variable from consideration for the rest of the search
  - If no variable has $BIC_{diff} > upper$ no variable is included in the current set of clustering variables

$$(Y^{(?)})^{k+1} \in (Y^{(other)})^{k}$$

$$(Y^{(clust)})^{k+1} = (Y^{(clust)})^{k}$$

Headlong Search Algorithm
Inclusion Step

- Basic idea - similar to Greedy Search:
  - Check, in order, each variable not currently included in the set of clustering variables for evidence of usefulness for clustering
  - Once a variable has $BIC_{diff} > upper$ include this variable in the current set of clustering variables and stop the inclusion step
  - If any variable checked has $BIC_{diff} < lower$ remove this variable from consideration for the rest of the search
  - If no variable has $BIC_{diff} > upper$ no variable is included in the current set of clustering variables

$$(Y^{(?)})^{k+1} \in (Y^{(other)})^k$$

$$(Y^{(clust)})^{k+1} = (Y^{(clust)})^k$$

Headlong Search Algorithm
Inclusion Step

- Basic idea - similar to Greedy Search:
  - Check, in order, each variable not currently included in the set of clustering variables for evidence of usefulness for clustering
  - Once a variable has $BIC_{diff} > upper$ include this variable in the current set of clustering variables and stop the inclusion step
  - If any variable checked has $BIC_{diff} < lower$ remove this variable from consideration for the rest of the search
  - If no variable has $BIC_{diff} > upper$ no variable is included in the current set of clustering variables

$$(Y^{(?)})^{k+1} \in (Y^{(other)})^{k}$$

$$(Y^{(clust)})^{k+1} = (Y^{(clust)})^{k}$$

Headlong Search Algorithm

Inclusion Step

- Basic idea - similar to Greedy Search:
  - Check, in order, each variable not currently included in the set of clustering variables for evidence of usefulness for clustering
  - Once a variable has $BIC_{diff} > upper$ include this variable in the current set of clustering variables and stop the inclusion step
  - If any variable checked has $BIC_{diff} < lower$ remove this variable from consideration for the rest of the search
  - If no variable has $BIC_{diff} > upper$ no variable is included in the current set of clustering variables

$$(Y^{(?)})^{k+1} \in (Y^{(other)})^{k}$$

$$(Y^{(clust)})^{k+1} = (Y^{(clust)})^{k}$$

Headlong Search Algorithm
Exclusion Step

- Basic idea - similar to Greedy Search:
  - Check, in order, each variable currently included in the set of clustering variables for evidence of usefulness for clustering
  - Once a variable has $BIC_{diff} < upper$ remove this variable from the current set of clustering variables and stop the exclusion step
  - If the variable removed also has $BIC_{diff} < lower$ remove this variable from consideration for the rest of the search
  - If no variable has $BIC_{diff} < upper$ no variable is removed from the current set of clustering variables

$$(Y^{(?)})^{k+1} \in (Y^{(clust)})^{k}$$

$$(Y^{(clust)})^{k+1} = (Y^{(clust)})^{k} / (Y^{(?)})^{k+1}$$

Headlong Search Algorithm

Exclusion Step

- Basic idea - similar to Greedy Search:
  - Check, in order, each variable currently included in the set of clustering variables for evidence of usefulness for clustering
  - Once a variable has $BIC_{diff} < upper$ remove this variable from the current set of clustering variables and stop the exclusion step
  - If the variable removed also has $BIC_{diff} < lower$ remove this variable from consideration for the rest of the search
  - If no variable has $BIC_{diff} < upper$ no variable is removed from the current set of clustering variables

$$(Y^{(?)})^{k+1} \in (Y^{(clust)})^{k}$$

$$(Y^{(clust)})^{k+1} = (Y^{(clust)})^{k} / (Y^{(?)})^{k+1}$$

Headlong Search Algorithm
Exclusion Step

- Basic idea - similar to Greedy Search:
  - Check, in order, each variable currently included in the set of clustering variables for evidence of usefulness for clustering
  - Once a variable has $BIC_{diff} < upper$ remove this variable from the current set of clustering variables and stop the exclusion step
  - If the variable removed also has $BIC_{diff} < lower$ remove this variable from consideration for the rest of the search
  - If no variable has $BIC_{diff} < upper$ no variable is removed from the current set of clustering variables

$$(Y^{(?)})^{k+1} \in (Y^{(clust)})^{k}$$

$$(Y^{(clust)})^{k+1} = (Y^{(clust)})^{k} / (Y^{(?)})^{k+1}$$

Headlong Search Algorithm
Exclusion Step

- Basic idea - similar to Greedy Search:
  - Check, in order, each variable currently included in the set of clustering variables for evidence of usefulness for clustering
  - Once a variable has $BIC_{diff} < upper$ remove this variable from the current set of clustering variables and stop the exclusion step
  - If the variable removed also has $BIC_{diff} < lower$ remove this variable from consideration for the rest of the search
  - If no variable has $BIC_{diff} < upper$ no variable is removed from the current set of clustering variables

$$(Y^{(?)})^{k+1} \in (Y^{(clust)})^k$$

$$(Y^{(clust)})^{k+1} = (Y^{(clust)})^k / (Y^{(?)})^{k+1}$$

- Basic idea - similar to Greedy Search:
  - Check, in order, each variable currently included in the set of clustering variables for evidence of usefulness for clustering
  - Once a variable has $BIC_{diff} < upper$ remove this variable from the current set of clustering variables and stop the exclusion step
  - If the variable removed also has $BIC_{diff} < lower$ remove this variable from consideration for the rest of the search
  - If no variable has $BIC_{diff} < upper$ no variable is removed from the current set of clustering variables

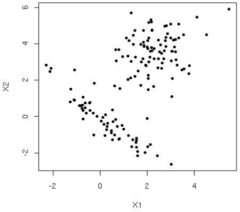
$$(Y^{(?)})^{k+1} \in (Y^{(clust)})^k$$

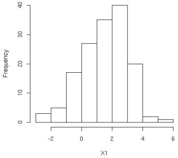
$$(Y^{(clust)})^{k+1} = (Y^{(clust)})^k / (Y^{(?)})^{k+1}$$

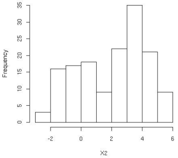
Outline

Simulated Data: 2 clusters
No noise variables

- First we look at an example where there are no noise variables present
- Have two variables with clustering information
- 150 observations
- The clusters are well separated with different variances



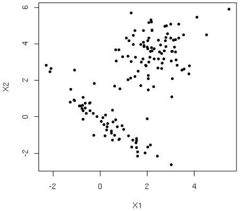


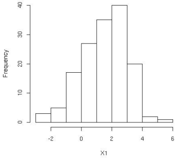


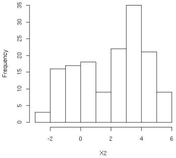
Simulated Data: 2 clusters

No noise variables

- First we look at an example where there are no noise variables present
- Have two variables with clustering information
- 150 observations
- The clusters are well separated with different variances

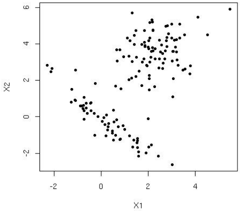


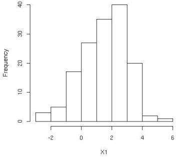


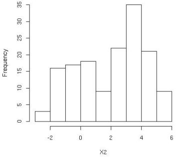


Simulated Data: 2 clusters
No noise variables

- First we look at an example where there are no noise variables present
- Have two variables with clustering information
- 150 observations
- The clusters are well separated with different variances



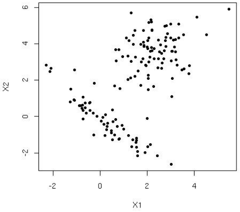


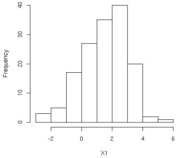


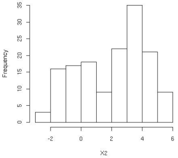
Simulated Data: 2 clusters

No noise variables

- First we look at an example where there are no noise variables present
- Have two variables with clustering information
- 150 observations
- The clusters are well separated with different variances







Greedy Search Results

- First, check both variables for clustering versus not clustering?

Variable set	Step	BIC cluster	BIC no cluster	BIC difference
-	Select Y1?	-544	-542	-2
-	Select Y2?	-634	-668	34

- First variable selected is Y2
- Second, check Y1 for evidence of bivariate clustering
- Y1 is also selected

Greedy Search Results

- First, check both variables for clustering versus not clustering?

Variable set	Step	BIC cluster	BIC no cluster	BIC difference
-	Select Y1?	-544	-542	-2
-	Select Y2?	-634	-668	34

- First variable selected is Y2
- Second, check Y1 for evidence of bivariate clustering
- Y1 is also selected

Greedy Search Results

- First, check both variables for clustering versus not clustering?

Variable set	Step	BIC cluster	BIC no cluster	BIC difference
-	Select Y1?	-544	-542	-2
-	Select Y2?	-634	-668	34

- First variable selected is Y2
- Second, check Y1 for evidence of bivariate clustering
- Y1 is also selected

Greedy Search Results

- First, check both variables for clustering versus not clustering?

Variable set	Step	BIC cluster	BIC no cluster	BIC difference
-	Select Y1?	-544	-542	-2
-	Select Y2?	-634	-668	34

- First variable selected is Y2
- Second, check Y1 for evidence of bivariate clustering
- Y1 is also selected

Greedy Search Results

- First, check both variables for clustering versus not clustering?

Variable set	Step	BIC cluster	BIC no cluster	BIC difference
-	Select Y1?	-544	-542	-2
-	Select Y2?	-634	-668	34

- First variable selected is Y2
- Second, check Y1 for evidence of bivariate clustering
- Y1 is also selected

Greedy Search Results

- First, check both variables for clustering versus not clustering?

Variable set	Step	BIC cluster	BIC no cluster	BIC difference
-	Select Y1?	-544	-542	-2
-	Select Y2?	-634	-668	34

- First variable selected is Y2
- Second, check Y1 for evidence of bivariate clustering

Variable set	Step	BIC cluster	BIC no cluster	BIC difference
Y2	Include Y1?	-1023	-1141	118

- Y1 is also selected

Greedy Search Results

- First, check both variables for clustering versus not clustering?

Variable set	Step	BIC cluster	BIC no cluster	BIC difference
-	Select Y1?	-544	-542	-2
-	Select Y2?	-634	-668	34

- First variable selected is Y2
- Second, check Y1 for evidence of bivariate clustering

Variable set	Step	BIC cluster	BIC no cluster	BIC difference
Y2	Include Y1?	-1023	-1141	118

- Y1 is also selected

Greedy Search Results

- Next check if a variable should be removed
- Neither variable is removed
- Final selection is (Y1, Y2)

Greedy Search Results

- Next check if a variable should be removed

Variable set	Step	BIC cluster	BIC no cluster	BIC difference
Y1, Y2	Exclude Y2?	-1023	-1177	154
Y1, Y2	Exclude Y1?	-1023	-1141	118

- Neither variable is removed
- Final selection is (Y1, Y2)

Greedy Search Results

- Next check if a variable should be removed

Variable set	Step	BIC cluster	BIC no cluster	BIC difference
Y1, Y2	Exclude Y2?	-1023	-1177	154
Y1, Y2	Exclude Y1?	-1023	-1141	118

- Neither variable is removed
- Final selection is (Y1, Y2)

Greedy Search Results

- Next check if a variable should be removed

Variable set	Step	BIC cluster	BIC no cluster	BIC difference
Y1, Y2	Exclude Y2?	-1023	-1177	154
Y1, Y2	Exclude Y1?	-1023	-1141	118

- Neither variable is removed
- Final selection is (Y1, Y2)

Greedy Search Results

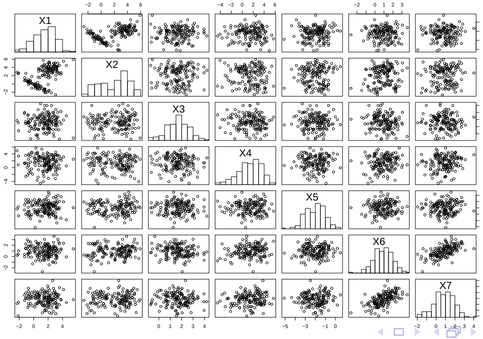
- Next check if a variable should be removed

Variable set	Step	BIC cluster	BIC no cluster	BIC difference
Y1, Y2	Exclude Y2?	-1023	-1177	154
Y1, Y2	Exclude Y1?	-1023	-1141	118

- Neither variable is removed
- Final selection is (Y1, Y2)

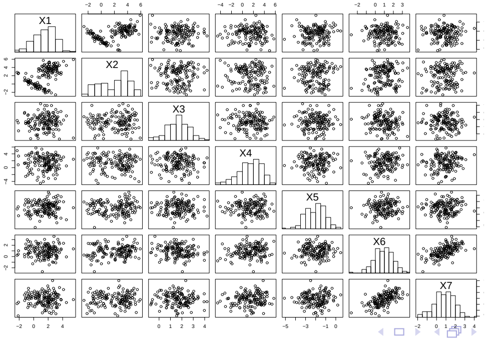
Simulated Data: 2 Clusters
5 noise variables

- Now we add noise variables to the 2 clustering variables
- Add 5 noise variables
  - Y3, Y4 and Y5 are independent normally distributed variables
  - Y6 and Y7 are dependent multivariate normally distributed variables



Simulated Data: 2 Clusters
5 noise variables

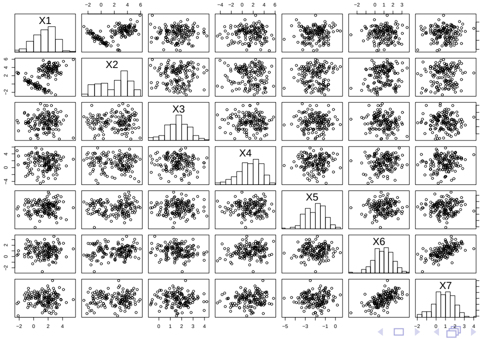
- Now we add noise variables to the 2 clustering variables
- Add 5 noise variables
  - Y3, Y4 and Y5 are independent normally distributed variables
  - Y6 and Y7 are dependent multivariate normally distributed variables



Simulated Data: 2 Clusters

5 noise variables

- Now we add noise variables to the 2 clustering variables
- Add 5 noise variables
  - Y3, Y4 and Y5 are independent normally distributed variables
  - Y6 and Y7 are dependent multivariate normally distributed variables



Greedy Search Results

Variable set	Step	BIC difference	Accepted?
-	Select Y2?	34	Yes
Y2	Include Y1?	118	Yes
Y1, Y2	Include Y6?	-12	No
Y1, Y2	Exclude Y1?	118	No

Greedy Search Results

Variable set	Step	BIC difference	Accepted?
-	Select Y2?	34	Yes
Y2	Include Y1?	118	Yes
Y1, Y2	Include Y6?	-12	No
Y1, Y2	Exclude Y1?	118	No

Greedy Search Results

Variable set	Step	BIC difference	Accepted?
-	Select Y2?	34	Yes
Y2	Include Y1?	118	Yes
Y1, Y2	Include Y6?	-12	No
Y1, Y2	Exclude Y1?	118	No

Greedy Search Results

Variable set	Step	BIC difference	Accepted?
-	Select Y2?	34	Yes
Y2	Include Y1?	118	Yes
Y1, Y2	Include Y6?	-12	No
Y1, Y2	Exclude Y1?	118	No

Compare Clustering Results

# of Variables	# of Groups	Error rate	Rand Index
All 7	5	44.7%	0.69
All 7	2 (constrained)	3.3%	0.94
Selected 2	2	0%	1

- Error Rate denotes the misclassification rate from the optimal matching of one cluster to one group
- The Rand Index is the sum of the number discordant and concordant matching pairs of observations across clusters and groups divided the total number of possible pairs of observations. 0 indicates poor matching of the clusters to groups, 1 indicates perfect matching.

Compare Clustering Results

# of Variables	# of Groups	Error rate	Rand Index
All 7	5	44.7%	0.69
All 7	2 (constrained)	3.3%	0.94
Selected 2	2	0%	1

- Error Rate denotes the misclassification rate from the optimal matching of one cluster to one group
- The Rand Index is the sum of the number discordant and concordant matching pairs of observations across clusters and groups divided the total number of possible pairs of observations. 0 indicates poor matching of the clusters to groups, 1 indicates perfect matching.

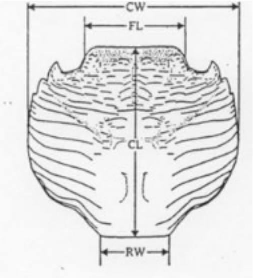
Compare Clustering Results

# of Variables	# of Groups	Error rate	Rand Index
All 7	5	44.7%	0.69
All 7	2 (constrained)	3.3%	0.94
Selected 2	2	0%	1

- Error Rate denotes the misclassification rate from the optimal matching of one cluster to one group
- The Rand Index is the sum of the number discordant and concordant matching pairs of observations across clusters and groups divided the total number of possible pairs of observations. 0 indicates poor matching of the clusters to groups, 1 indicates perfect matching.

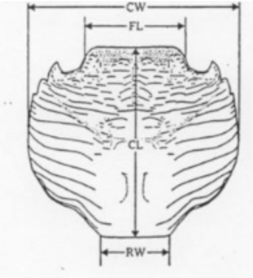
Crabs Data

- Crabs data has theoretically 4 groups: male orange, female orange, male blue and female blue
- 200 observations (50 per group)
- 5 variables measuring size
  - Width of frontal lip (FL)
  - Rear width (RW)
  - Length along mid-line of carapace (CL)
  - Maximum width of carapace (CW)
  - Body depth (BD)



Crabs Data

- Crabs data has theoretically 4 groups: male orange, female orange, male blue and female blue
- 200 observations (50 per group)
- 5 variables measuring size
  - Width of frontal lip (FL)
  - Rear width (RW)
  - Length along mid-line of carapace (CL)
  - Maximum width of carapace (CW)
  - Body depth (BD)



Compare Clustering Results

- Variables selected: all variables except length along mid-line of carapace (CL)

Compare Clustering Results

- Variables selected: all variables except length along mid-line of carapace (CL)

# of Variables	# of Groups	Error rate	Rand Index
All 5	7	42.5%	0.77
All 5	4 (constrained)	7.5%	0.93
Selected 4	4	7.5%	0.93

Compare Clustering Results

- Variables selected: all variables except length along mid-line of carapace (CL)

# of Variables	# of Groups	Error rate	Rand Index
All 5	7	42.5%	0.77
All 5	4 (constrained)	7.5%	0.93
Selected 4	4	7.5%	0.93

Compare Clustering Results

- Variables selected: all variables except length along mid-line of carapace (CL)

# of Variables	# of Groups	Error rate	Rand Index
All 5	7	42.5%	0.77
All 5	4 (constrained)	7.5%	0.93
Selected 4	4	7.5%	0.93

Outline

Simulated Data: 2 Classes
No noise variables

- We have 6 binary variables with success probabilities:

	Var1	Var2	Var3	Var4	Var5	Var6
P(Var_i =1|Group 1)	0.9	0.2	0.1	0.8	0.7	0.6
P(Var_i =1|Group 2)	0.2	0.9	0.8	0.1	0.2	0.3

Greedy Search Results

- First check all subsets of 4 variables
- Add/remove single variables
- Allow number of groups estimated to increase or decrease as the number of clustering variables does

Greedy Search Results

- First check all subsets of 4 variables
- Add/remove single variables
- Allow number of groups estimated to increase or decrease as the number of clustering variables does

Greedy Search Results

- First check all subsets of 4 variables
- Add/remove single variables
- Allow number of groups estimated to increase or decrease as the number of clustering variables does

Greedy Search Results

- First check all subsets of 4 variables
- Add/remove single variables
- Allow number of groups estimated to increase or decrease as the number of clustering variables does

Variable set	Step	BIC difference	Accepted?
-	Select Y1, Y2, Y3 & Y4?	2068	Yes
Y1, Y2, Y3, & Y4	Include Y5?	516	Yes
Y1, Y2, Y3, Y4 & Y5	Include Y6?	196	Yes
Y1, Y2, Y3, Y4, Y5 & Y6	Exclude Y6?	196	No

Greedy Search Results

- First check all subsets of 4 variables
- Add/remove single variables
- Allow number of groups estimated to increase or decrease as the number of clustering variables does

Variable set	Step	BIC difference	Accepted?
-	Select Y1, Y2, Y3 & Y4?	2068	Yes
Y1, Y2, Y3, & Y4	Include Y5?	516	Yes
Y1, Y2, Y3, Y4 & Y5	Include Y6?	196	Yes
Y1, Y2, Y3, Y4, Y5 & Y6	Exclude Y6?	196	No

Greedy Search Results

- First check all subsets of 4 variables
- Add/remove single variables
- Allow number of groups estimated to increase or decrease as the number of clustering variables does

Variable set	Step	BIC difference	Accepted?
-	Select Y1, Y2, Y3 & Y4?	2068	Yes
Y1, Y2, Y3, & Y4	Include Y5?	516	Yes
Y1, Y2, Y3, Y4 & Y5	Include Y6?	196	Yes
Y1, Y2, Y3, Y4, Y5 & Y6	Exclude Y6?	196	No

Greedy Search Results

- First check all subsets of 4 variables
- Add/remove single variables
- Allow number of groups estimated to increase or decrease as the number of clustering variables does

Variable set	Step	BIC difference	Accepted?
-	Select Y1, Y2, Y3 & Y4?	2068	Yes
Y1, Y2, Y3, & Y4	Include Y5?	516	Yes
Y1, Y2, Y3, Y4 & Y5	Include Y6?	196	Yes
Y1, Y2, Y3, Y4, Y5 & Y6	Exclude Y6?	196	No

Simulated Data: 2 Classes
4 noise variables

- To the previous data we add four noise variables
- Each of the noise variables has the same success probabilities in each of the groups
  - P(Var_7=1) = 0.5
  - P(Var_8=1) = 0.9
  - P(Var_9=1) = 0.1
  - P(Var_{10}=1) = 0.8

Simulated Data: 2 Classes
4 noise variables

- To the previous data we add four noise variables
- Each of the noise variables has the same success probabilities in each of the groups
  - P(Var_7=1) = 0.5
  - P(Var_8=1) = 0.9
  - P(Var_9=1) = 0.1
  - P(Var_{10}=1) = 0.8

Simulated Data: 2 Classes

4 noise variables

- To the previous data we add four noise variables
- Each of the noise variables has the same success probabilities in each of the groups
  - P(Var_7=1) = 0.5
  - P(Var_8=1) = 0.9
  - P(Var_9=1) = 0.1
  - P(Var_{10}=1) = 0.8

Simulated Data: 2 Classes

4 noise variables

- To the previous data we add four noise variables
- Each of the noise variables has the same success probabilities in each of the groups
  - P(Var_7=1) = 0.5
  - P(Var_8=1) = 0.9
  - P(Var_9=1) = 0.1
  - P(Var_{10}=1) = 0.8

Simulated Data: 2 Classes
4 noise variables

- To the previous data we add four noise variables
- Each of the noise variables has the same success probabilities in each of the groups
  - P(Var_7=1) = 0.5
  - P(Var_8=1) = 0.9
  - P(Var_9=1) = 0.1
  - P(Var_{10}=1) = 0.8

Variable Selection Results

Variable set	Step	BIC difference	Accepted?
-	Select Y1, Y2, Y3 & Y4?	2068	Yes
Y1, Y2, Y3 & Y4	Include Y5?	516	Yes
Y1, Y2, Y3, Y4 & Y5	Include Y6?	196	Yes
Y1, Y2, Y3, Y4, Y5 & Y6	Exclude Y6?	196	No
Y1, Y2, Y3, Y4, Y5 & Y6	Include Y9?	-3	No

Variable Selection Results

Variable set	Step	BIC difference	Accepted?
-	Select Y1, Y2, Y3 & Y4?	2068	Yes
Y1, Y2, Y3 & Y4	Include Y5?	516	Yes
Y1, Y2, Y3, Y4 & Y5	Include Y6?	196	Yes
Y1, Y2, Y3, Y4, Y5 & Y6	Exclude Y6?	196	No
Y1, Y2, Y3, Y4, Y5 & Y6	Include Y9?	-3	No

Variable Selection Results

Variable set	Step	BIC difference	Accepted?
-	Select Y1, Y2, Y3 & Y4?	2068	Yes
Y1, Y2, Y3 & Y4	Include Y5?	516	Yes
Y1, Y2, Y3, Y4 & Y5	Include Y6?	196	Yes
Y1, Y2, Y3, Y4, Y5 & Y6	Exclude Y6?	196	No
Y1, Y2, Y3, Y4, Y5 & Y6	Include Y9?	-3	No

Variable Selection Results

Variable set	Step	BIC difference	Accepted?
-	Select Y1, Y2, Y3 & Y4?	2068	Yes
Y1, Y2, Y3 & Y4	Include Y5?	516	Yes
Y1, Y2, Y3, Y4 & Y5	Include Y6?	196	Yes
Y1, Y2, Y3, Y4, Y5 & Y6	Exclude Y6?	196	No
Y1, Y2, Y3, Y4, Y5 & Y6	Include Y9?	-3	No

Variable Selection Results

Variable set	Step	BIC difference	Accepted?
-	Select Y1, Y2, Y3 & Y4?	2068	Yes
Y1, Y2, Y3 & Y4	Include Y5?	516	Yes
Y1, Y2, Y3, Y4 & Y5	Include Y6?	196	Yes
Y1, Y2, Y3, Y4, Y5 & Y6	Exclude Y6?	196	No
Y1, Y2, Y3, Y4, Y5 & Y6	Include Y9?	-3	No

Comparing Clustering Results

- Correct 6 variables selected (Y1, Y2, Y3, Y4, Y5 & Y6) out of 10

Comparing Clustering Results

- Correct 6 variables selected (Y1, Y2, Y3, Y4, Y5 & Y6) out of 10

# of Variables	# of Groups	Error rate
All 10	2	2.85%
Selected 6	2	2.75%

Comparing Clustering Results

- Correct 6 variables selected (Y1, Y2, Y3, Y4, Y5 & Y6) out of 10

# of Variables	# of Groups	Error rate
All 10	2	2.85%
Selected 6	2	2.75%

Outline

HapMap Data

- HapMap Project: international effort to identify and catalog genetic similarities and differences in human beings, started in October 2002
- Goal: to compare the genetic sequences of different individuals to identify chromosomal regions where genetic variants are shared
- On average, one in every 1,200 bases will differ between individuals
- Most common difference: single nucleotide polymorphism (SNP)
- An estimated 10 million SNPs commonly occurring in the human genome

HapMap Data

- HapMap Project: international effort to identify and catalog genetic similarities and differences in human beings, started in October 2002
- Goal: to compare the genetic sequences of different individuals to identify chromosomal regions where genetic variants are shared
- On average, one in every 1,200 bases will differ between individuals
- Most common difference: single nucleotide polymorphism (SNP)
- An estimated 10 million SNPs commonly occurring in the human genome

HapMap Data

- HapMap Project: international effort to identify and catalog genetic similarities and differences in human beings, started in October 2002
- Goal: to compare the genetic sequences of different individuals to identify chromosomal regions where genetic variants are shared
- On average, one in every 1,200 bases will differ between individuals
- Most common difference: single nucleotide polymorphism (SNP)
- An estimated 10 million SNPs commonly occurring in the human genome

HapMap Data

- HapMap Project: international effort to identify and catalog genetic similarities and differences in human beings, started in October 2002
- Goal: to compare the genetic sequences of different individuals to identify chromosomal regions where genetic variants are shared
- On average, one in every 1,200 bases will differ between individuals
- Most common difference: single nucleotide polymorphism (SNP)
- An estimated 10 million SNPs commonly occurring in the human genome

HapMap Data

- HapMap Project: international effort to identify and catalog genetic similarities and differences in human beings, started in October 2002
- Goal: to compare the genetic sequences of different individuals to identify chromosomal regions where genetic variants are shared
- On average, one in every 1,200 bases will differ between individuals
- Most common difference: single nucleotide polymorphism (SNP)
- An estimated 10 million SNPs commonly occurring in the human genome

HapMap Data

- 210 subjects from different ethnic populations
- 3 or 4 possible groups
  - European (60 Utah residents with ancestry from Northern and Western Europe)
  - African (60 Yoruban in Ibadan, Nigeria, West Africa)
  - Asian (can be split into 45 Japanese in Tokyo, Japan and 45 Han Chinese in Beijing, China)
- 639 variables: SNPs
- Most of the variables are binary but some have 3 categories

HapMap Data

- 210 subjects from different ethnic populations
- 3 or 4 possible groups
  - European (60 Utah residents with ancestry from Northern and Western Europe)
  - African (60 Yoruban in Ibadan, Nigeria, West Africa)
  - Asian (can be split into 45 Japanese in Tokyo, Japan and 45 Han Chinese in Beijing, China)
- 639 variables: SNPs
- Most of the variables are binary but some have 3 categories

HapMap Data

- 210 subjects from different ethnic populations
- 3 or 4 possible groups
  - European (60 Utah residents with ancestry from Northern and Western Europe)
  - African (60 Yoruban in Ibadan, Nigeria, West Africa)
  - Asian (can be split into 45 Japanese in Tokyo, Japan and 45 Han Chinese in Beijing, China)
- 639 variables: SNPs
- Most of the variables are binary but some have 3 categories

HapMap Data

- 210 subjects from different ethnic populations
- 3 or 4 possible groups
  - European (60 Utah residents with ancestry from Northern and Western Europe)
  - African (60 Yoruban in Ibadan, Nigeria, West Africa)
  - Asian (can be split into 45 Japanese in Tokyo, Japan and 45 Han Chinese in Beijing, China)
- 639 variables: SNPs
- Most of the variables are binary but some have 3 categories

HapMap Data

- 210 subjects from different ethnic populations
- 3 or 4 possible groups
  - European (60 Utah residents with ancestry from Northern and Western Europe)
  - African (60 Yoruban in Ibadan, Nigeria, West Africa)
  - Asian (can be split into 45 Japanese in Tokyo, Japan and 45 Han Chinese in Beijing, China)
- 639 variables: SNPs
- Most of the variables are binary but some have 3 categories

HapMap Data

- 210 subjects from different ethnic populations
- 3 or 4 possible groups
  - European (60 Utah residents with ancestry from Northern and Western Europe)
  - African (60 Yoruban in Ibadan, Nigeria, West Africa)
  - Asian (can be split into 45 Japanese in Tokyo, Japan and 45 Han Chinese in Beijing, China)
- 639 variables: SNPs
- Most of the variables are binary but some have 3 categories

HapMap Data

- 210 subjects from different ethnic populations
- 3 or 4 possible groups
  - European (60 Utah residents with ancestry from Northern and Western Europe)
  - African (60 Yoruban in Ibadan, Nigeria, West Africa)
  - Asian (can be split into 45 Japanese in Tokyo, Japan and 45 Han Chinese in Beijing, China)
- 639 variables: SNPs
- Most of the variables are binary but some have 3 categories

Results

- For the Latent Class models on all variables: 3 class model selected (BIC -141418)
- Difference in BIC values from other models:
  - 2 Class Model: -1293
  - 4 Class Model: -5244
- 413 variables are selected with a 3 class model (BIC -91147)
- Difference in BIC values from other models:
  - 2 Class Model: -2324
  - 4 Class Model: -3344

Results

- For the Latent Class models on all variables: 3 class model selected (BIC -141418)
- Difference in BIC values from other models:
  - 2 Class Model: -1293
  - 4 Class Model: -5244
- 413 variables are selected with a 3 class model (BIC -91147)
- Difference in BIC values from other models:
  - 2 Class Model: -2324
  - 4 Class Model: -3344

Results

- For the Latent Class models on all variables: 3 class model selected (BIC -141418)
- Difference in BIC values from other models:
  - 2 Class Model: -1293
  - 4 Class Model: -5244
- 413 variables are selected with a 3 class model (BIC -91147)
- Difference in BIC values from other models:
  - 2 Class Model: -2324
  - 4 Class Model: -3344

Results

- For the Latent Class models on all variables: 3 class model selected (BIC -141418)
- Difference in BIC values from other models:
  - 2 Class Model: -1293
  - 4 Class Model: -5244
- 413 variables are selected with a 3 class model (BIC -91147)
- Difference in BIC values from other models:
  - 2 Class Model: -2324
  - 4 Class Model: -3344

Results

- For the Latent Class models on all variables: 3 class model selected (BIC -141418)
- Difference in BIC values from other models:
  - 2 Class Model: -1293
  - 4 Class Model: -5244
- 413 variables are selected with a 3 class model (BIC -91147)
- Difference in BIC values from other models:
  - 2 Class Model: -2324
  - 4 Class Model: -3344

Results

- For the Latent Class models on all variables: 3 class model selected (BIC -141418)
- Difference in BIC values from other models:
  - 2 Class Model: -1293
  - 4 Class Model: -5244
- 413 variables are selected with a 3 class model (BIC -91147)
- Difference in BIC values from other models:
  - 2 Class Model: -2324
  - 4 Class Model: -3344

Outline

Summary

- We introduced a simple stepwise method of variable selection specifically tailored to the mixture model clustering setting
- In the simulated examples this method was shown to select the correct variables
- In both the real and simulated examples shown the method improved both the estimate of the number of groups and the misclassification rate
- Possible Advantages of this approach in practice:
  - Decrease the number of variables being modeled or collected
  - Improve estimate of the number of groups in the data
  - Decrease the misclassification rate

Summary

- We introduced a simple stepwise method of variable selection specifically tailored to the mixture model clustering setting
- In the simulated examples this method was shown to select the correct variables
- In both the real and simulated examples shown the method improved both the estimate of the number of groups and the misclassification rate
- Possible Advantages of this approach in practice:
  - Decrease the number of variables being modeled or collected
  - Improve estimate of the number of groups in the data
  - Decrease the misclassification rate

Summary

- We introduced a simple stepwise method of variable selection specifically tailored to the mixture model clustering setting
- In the simulated examples this method was shown to select the correct variables
- In both the real and simulated examples shown the method improved both the estimate of the number of groups and the misclassification rate
- Possible Advantages of this approach in practice:
  - Decrease the number of variables being modeled or collected
  - Improve estimate of the number of groups in the data
  - Decrease the misclassification rate

Summary

- We introduced a simple stepwise method of variable selection specifically tailored to the mixture model clustering setting
- In the simulated examples this method was shown to select the correct variables
- In both the real and simulated examples shown the method improved both the estimate of the number of groups and the misclassification rate
- Possible Advantages of this approach in practice:
  - Decrease the number of variables being modeled or collected
  - Improve estimate of the number of groups in the data
  - Decrease the misclassification rate

Summary

- We introduced a simple stepwise method of variable selection specifically tailored to the mixture model clustering setting
- In the simulated examples this method was shown to select the correct variables
- In both the real and simulated examples shown the method improved both the estimate of the number of groups and the misclassification rate
- Possible Advantages of this approach in practice:
    - Decrease the number of variables being modeled or collected
    - Improve estimate of the number of groups in the data
    - Decrease the misclassification rate

Summary

- We introduced a simple stepwise method of variable selection specifically tailored to the mixture model clustering setting
- In the simulated examples this method was shown to select the correct variables
- In both the real and simulated examples shown the method improved both the estimate of the number of groups and the misclassification rate
- Possible Advantages of this approach in practice:
  - Decrease the number of variables being modeled or collected
  - Improve estimate of the number of groups in the data
  - Decrease the misclassification rate

Outline

Other Work

- Variable Selection model (along with incorporation of unlabelled data for estimation) applied to Model-Based Discriminant Analysis
- Variable Selection in Mixture of Experts models
- Incoporating dependence in Variable Selection for LCA models

Acknowledgements

- Adrian Raftery
- NIH Grant 8 R01 EB002137-02
- Model-based Clustering Working Group in Seattle

References

- `clustvarsel`. R package for Variable Selection with `mclust02`. *http://www.stats.bris.ac.uk/R/src/contrib/Descriptions/clustvarsel.html*, 2006
- A. E. Raftery and N. Dean. Variable Selection for Model-Based Clustering. *Journal of the American Statistical Association*, 473: 168–178 , 2006.
- C. Keribin. Consistent Estimate of the Order of Mixture Models. *Comptes Rendues de l'Academie des Sciences, Série I-Mathématiques*, 326:243–248, 1998.
- D. Rusakov and D. Geiger. Asymptotic Model Selection for Naive Bayesian Networks. *Journal of Machine Learning Research*, 6 : 1–35, 2005.
- W.C. Chang. On Using Principal Components Before Separating a Mixture of Two Multivariate Normal Distributions. *Applied Statistics*, 32 : 267–275, 1983.

Search Algorithms Issues

- The search is stopped after consecutive inclusion and exclusion steps fail to change the set of clustering variables
- Need to specify *lower* and *upper* for the headlong algorithm
  - *upper* is the minimum BIC_{diff} which we consider evidence for a variable's inclusion/exclusion (default=0)
  - *lower* is the level of BIC_{diff} for which we believe a variable will never be included in subsequent steps
- Neither search algorithm is guaranteed to find the overall optimal set of clustering variables (only a local optimum)
- For each variable checked in the inclusion/exclusion steps, clustering models need to be fitted to two different datasets $(Y^{(clust)}, Y^{(?)})$ and $Y^{(clust)}$.
- Clustering models for various numbers of clusters and different model restrictions are fit and the models with the best BIC scores are used

Search Algorithms Issues

- The search is stopped after consecutive inclusion and exclusion steps fail to change the set of clustering variables
- Need to specify *lower* and *upper* for the headlong algorithm
  - *upper* is the minimum BIC_{diff} which we consider evidence for a variable's inclusion/exclusion (default=0)
  - *lower* is the level of BIC_{diff} for which we believe a variable will never be included in subsequent steps
- Neither search algorithm is guaranteed to find the overall optimal set of clustering variables (only a local optimum)
- For each variable checked in the inclusion/exclusion steps, clustering models need to be fitted to two different datasets $(Y^{(clust)}, Y^{(?)})$ and $Y^{(clust)}$.
- Clustering models for various numbers of clusters and different model restrictions are fit and the models with the best BIC scores are used

Search Algorithms Issues

- The search is stopped after consecutive inclusion and exclusion steps fail to change the set of clustering variables
- Need to specify *lower* and *upper* for the headlong algorithm
  - *upper* is the minimum BIC_{diff} which we consider evidence for a variable's inclusion/exclusion (default=0)
  - *lower* is the level of BIC_{diff} for which we believe a variable will never be included in subsequent steps
- Neither search algorithm is guaranteed to find the overall optimal set of clustering variables (only a local optimum)
- For each variable checked in the inclusion/exclusion steps, clustering models need to be fitted to two different datasets ($Y^{(clust)}, Y^{(?)}$) and $Y^{(clust)}$.
- Clustering models for various numbers of clusters and different model restrictions are fit and the models with the best BIC scores are used

Search Algorithms Issues

- The search is stopped after consecutive inclusion and exclusion steps fail to change the set of clustering variables
- Need to specify *lower* and *upper* for the headlong algorithm
  - *upper* is the minimum BIC_{diff} which we consider evidence for a variable's inclusion/exclusion (default=0)
  - *lower* is the level of BIC_{diff} for which we believe a variable will never be included in subsequent steps
- Neither search algorithm is guaranteed to find the overall optimal set of clustering variables (only a local optimum)
- For each variable checked in the inclusion/exclusion steps, clustering models need to be fitted to two different datasets ($Y^{(clust)}$, $Y^{(?)}$) and $Y^{(clust)}$.
- Clustering models for various numbers of clusters and different model restrictions are fit and the models with the best BIC scores are used

Search Algorithms Issues

- The search is stopped after consecutive inclusion and exclusion steps fail to change the set of clustering variables
- Need to specify *lower* and *upper* for the headlong algorithm
  - *upper* is the minimum BIC_{diff} which we consider evidence for a variable's inclusion/exclusion (default=0)
  - *lower* is the level of BIC_{diff} for which we believe a variable will never be included in subsequent steps
- Neither search algorithm is guaranteed to find the overall optimal set of clustering variables (only a local optimum)
- For each variable checked in the inclusion/exclusion steps, clustering models need to be fitted to two different datasets ($Y^{(clust)}, Y^{(?)}$) and $Y^{(clust)}$.
- Clustering models for various numbers of clusters and different model restrictions are fit and the models with the best BIC scores are used

Search Algorithms Issues

- The search is stopped after consecutive inclusion and exclusion steps fail to change the set of clustering variables
- Need to specify *lower* and *upper* for the headlong algorithm
  - *upper* is the minimum BIC_{diff} which we consider evidence for a variable's inclusion/exclusion (default=0)
  - *lower* is the level of BIC_{diff} for which we believe a variable will never be included in subsequent steps
- Neither search algorithm is guaranteed to find the overall optimal set of clustering variables (only a local optimum)
- For each variable checked in the inclusion/exclusion steps, clustering models need to be fitted to two different datasets $(Y^{(clust)}, Y^{(?)})$ and $Y^{(clust)}$.
- Clustering models for various numbers of clusters and different model restrictions are fit and the models with the best BIC scores are used

Search Algorithms Issues

- The search is stopped after consecutive inclusion and exclusion steps fail to change the set of clustering variables
- Need to specify *lower* and *upper* for the headlong algorithm
  - *upper* is the minimum BIC_{diff} which we consider evidence for a variable's inclusion/exclusion (default=0)
  - *lower* is the level of BIC_{diff} for which we believe a variable will never be included in subsequent steps
- Neither search algorithm is guaranteed to find the overall optimal set of clustering variables (only a local optimum)
- For each variable checked in the inclusion/exclusion steps, clustering models need to be fitted to two different datasets $(Y^{(clust)}, Y^{(?)})$ and $Y^{(clust)}$.
- Clustering models for various numbers of clusters and different model restrictions are fit and the models with the best BIC scores are used

Search Algorithms Issues

- Within each clustering model (for each dataset , each number of clusters) starting values are needed
  - In MBC we can use hierarchical clustering to give a single set of good values to use for starting the clustering algorithm
  - In LCA we need to generate multiple sets of starting values, run the algorithm and use the model with the highest BIC/likelihood

  $\Rightarrow$ Possibly huge number of clustering runs (depending on whether the range of numbers of clusters allowed overall is large)
- Restricting the range of number of clusters could cause errors/omissions in the variables selected
- In LCA, the number of clusters/classes allowed will depend on the size of the set of clustering variables

Search Algorithms Issues

- Within each clustering model (for each dataset , each number of clusters) starting values are needed
  - In MBC we can use hierarchical clustering to give a single set of good values to use for starting the clustering algorithm
  - In LCA we need to generate multiple sets of starting values, run the algorithm and use the model with the highest BIC/likelihood

  $\Rightarrow$ Possibly huge number of clustering runs (depending on whether the range of numbers of clusters allowed overall is large)
- Restricting the range of number of clusters could cause errors/omissions in the variables selected
- In LCA, the number of clusters/classes allowed will depend on the size of the set of clustering variables

Search Algorithms Issues

- Within each clustering model (for each dataset , each number of clusters) starting values are needed
  - In MBC we can use hierarchical clustering to give a single set of good values to use for starting the clustering algorithm
  - In LCA we need to generate multiple sets of starting values, run the algorithm and use the model with the highest BIC/likelihood

  ⇒ Possibly huge number of clustering runs (depending on whether the range of numbers of clusters allowed overall is large)
- Restricting the range of number of clusters could cause errors/omissions in the variables selected
- In LCA, the number of clusters/classes allowed will depend on the size of the set of clustering variables

Search Algorithms Issues

- Within each clustering model (for each dataset , each number of clusters) starting values are needed
  - In MBC we can use hierarchical clustering to give a single set of good values to use for starting the clustering algorithm
  - In LCA we need to generate multiple sets of starting values, run the algorithm and use the model with the highest BIC/likelihood

  $\Rightarrow$ Possibly huge number of clustering runs (depending on whether the range of numbers of clusters allowed overall is large)
- Restricting the range of number of clusters could cause errors/omissions in the variables selected
- In LCA, the number of clusters/classes allowed will depend on the size of the set of clustering variables

Search Algorithms Issues

- Within each clustering model (for each dataset , each number of clusters) starting values are needed
  - In MBC we can use hierarchical clustering to give a single set of good values to use for starting the clustering algorithm
  - In LCA we need to generate multiple sets of starting values, run the algorithm and use the model with the highest BIC/likelihood

  $\Rightarrow$ Possibly huge number of clustering runs (depending on whether the range of numbers of clusters allowed overall is large)
- Restricting the range of number of clusters could cause errors/omissions in the variables selected
- In LCA, the number of clusters/classes allowed will depend on the size of the set of clustering variables

Search Algorithms Issues

- Within each clustering model (for each dataset , each number of clusters) starting values are needed
  - In MBC we can use hierarchical clustering to give a single set of good values to use for starting the clustering algorithm
  - In LCA we need to generate multiple sets of starting values, run the algorithm and use the model with the highest BIC/likelihood

  $\Rightarrow$ Possibly huge number of clustering runs (depending on whether the range of numbers of clusters allowed overall is large)
- Restricting the range of number of clusters could cause errors/omissions in the variables selected
- In LCA, the number of clusters/classes allowed will depend on the size of the set of clustering variables

Adjustment of Search Methods

Reducing the number of clusters checked

- We would like to start the number of clusters allowed to be all possible for the first selection inclusion step
- For subsequent steps we would like to center the number of clusters checked around the best number of clusters found in the previous step and grow the number of clusters allowed gradually
- Define $G_{current}$ as the best number of clusters, in terms of BIC, for the previous step and $G_{max\ allowed}$ as the maximum number of clusters allowed for the entire algorithm
- We allow the number of clusters checked for datasets $(Y^{(clust)}, Y^{(?)})$ and $(Y^{(clust)})$ to range from $\max(2, G_{current} - 1)$ to $\min(G_{current} + 1, G_{max\ allowed})$

Adjustment of Search Methods

Reducing the number of clusters checked

- We would like to start the number of clusters allowed to be all possible for the first selection inclusion step
- For subsequent steps we would like to center the number of clusters checked around the best number of clusters found in the previous step and grow the number of clusters allowed gradually
- Define $G_{current}$ as the best number of clusters, in terms of BIC, for the previous step and $G_{max\ allowed}$ as the maximum number of clusters allowed for the entire algorithm
- We allow the number of clusters checked for datasets $(Y^{(clust)}, Y^{(?)})$ and $(Y^{(clust)})$ to range from $\max(2, G_{current} - 1)$ to $\min(G_{current} + 1, G_{max\ allowed})$

Adjustment of Search Methods
Reducing the number of clusters checked

- We would like to start the number of clusters allowed to be all possible for the first selection inclusion step
- For subsequent steps we would like to center the number of clusters checked around the best number of clusters found in the previous step and grow the number of clusters allowed gradually
- Define $G_{current}$ as the best number of clusters, in terms of BIC, for the previous step and $G_{max\ allowed}$ as the maximum number of clusters allowed for the entire algorithm
- We allow the number of clusters checked for datasets $(Y^{(clust)}, Y^{(?)})$ and $(Y^{(clust)})$ to range from $\max(2, G_{current} - 1)$ to $\min(G_{current} + 1, G_{max\ allowed})$

Adjustment of Search Methods

Reducing the number of clusters checked

- We would like to start the number of clusters allowed to be all possible for the first selection inclusion step
- For subsequent steps we would like to center the number of clusters checked around the best number of clusters found in the previous step and grow the number of clusters allowed gradually
- Define $G_{current}$ as the best number of clusters, in terms of BIC, for the previous step and $G_{max\ allowed}$ as the maximum number of clusters allowed for the entire algorithm
- We allow the number of clusters checked for datasets $(Y^{(clust)}, Y^{(?)})$ and $(Y^{(clust)})$ to range from $\max(2, G_{current} - 1)$ to $\min(G_{current} + 1, G_{max\ allowed})$

Adjustment of Search Methods

Getting quick reasonable starting points

$$z_{ij} = P(\text{Observation } i \text{ being in cluster } j)$$

- We need to get good starting posterior probability membership matrices z for (at most) $G_{current} \pm 1$ clusters
- For $G_{current}$, use z matrix saved from last clustering
- For $G_{current} - 1$ merge 2 closest clusters from last clustering (add corresponding columns in z matrix for $G_{current}$ from last clustering)
- For $G_{current} + 1$ split largest cluster from last clustering into 2 (estimate 2 cluster model using weighted mixture model clustering with weights from z column corresponding to largest cluster)

Adjustment of Search Methods

Getting quick reasonable starting points

$$z_{ij} = P(\text{Observation } i \text{ being in cluster } j)$$

- We need to get good starting posterior probability membership matrices z for (at most) $G_{current} \pm 1$ clusters
- For $G_{current}$, use z matrix saved from last clustering
- For $G_{current} - 1$ merge 2 closest clusters from last clustering (add corresponding columns in z matrix for $G_{current}$ from last clustering)
- For $G_{current} + 1$ split largest cluster from last clustering into 2 (estimate 2 cluster model using weighted mixture model clustering with weights from z column corresponding to largest cluster)

Adjustment of Search Methods

Getting quick reasonable starting points

$$z_{ij} = P(\text{Observation } i \text{ being in cluster } j)$$

- We need to get good starting posterior probability membership matrices z for (at most) $G_{current} \pm 1$ clusters
- For $G_{current}$, use z matrix saved from last clustering
- For $G_{current} - 1$ merge 2 closest clusters from last clustering (add corresponding columns in z matrix for $G_{current}$ from last clustering)
- For $G_{current} + 1$ split largest cluster from last clustering into 2 (estimate 2 cluster model using weighted mixture model clustering with weights from z column corresponding to largest cluster)

Adjustment of Search Methods
Getting quick reasonable starting points

$$z_{ij} = P(\text{Observation } i \text{ being in cluster } j)$$

- We need to get good starting posterior probability membership matrices z for (at most) $G_{current} \pm 1$ clusters
- For $G_{current}$, use z matrix saved from last clustering
- For $G_{current} - 1$ merge 2 closest clusters from last clustering (add corresponding columns in z matrix for $G_{current}$ from last clustering)
- For $G_{current} + 1$ split largest cluster from last clustering into 2 (estimate 2 cluster model using weighted mixture model clustering with weights from z column corresponding to largest cluster)

Adjustment of Search Methods

Getting quick reasonable starting points

$$z_{ij} = P(\text{Observation } i \text{ being in cluster } j)$$

- We need to get good starting posterior probability membership matrices z for (at most) $G_{current} \pm 1$ clusters
- For $G_{current}$, use z matrix saved from last clustering
- For $G_{current} - 1$ merge 2 closest clusters from last clustering (add corresponding columns in z matrix for $G_{current}$ from last clustering)
- For $G_{current} + 1$ split largest cluster from last clustering into 2 (estimate 2 cluster model using weighted mixture model clustering with weights from z column corresponding to largest cluster)