Recent Advances in Model-Based Clustering: Image Segmentation and Variable Selection

Adrian E. Raftery
University of Washington, Seattle, and UTIA Dept of Adaptive Systems, Prague
www.stat.washington.edu/raftery

Joint work with Nema Dean, Chris Fraley, Florence Forbes and Nathalie Peyrard
Supported by the US National Institutes of Health

2nd International Workshop on Data – Algorithms – Decision Making
Třešť, ČR
December 10, 2006

Outline

Outline

- Model-based clustering: Basic ideas

Outline

- Model-based clustering: Basic ideas
- Image segmentation applications of model-based clustering

Outline

- Model-based clustering: Basic ideas
- Image segmentation applications of model-based clustering
- Variable/feature selection for model-based clustering

Cluster Analysis

Cluster Analysis

- Automatic numerical methods for finding groups in data that are

Cluster Analysis

- Automatic numerical methods for finding groups in data that are
  - separated

Cluster Analysis

- Automatic numerical methods for finding groups in data that are
  - separated
  - internally cohesive

Cluster Analysis

- Automatic numerical methods for finding groups in data that are
  - separated
  - internally cohesive
- Invented in the 1950s by Sokal, Sneath and others motivated by

Cluster Analysis

- Automatic numerical methods for finding groups in data that are
  - separated
  - internally cohesive
- Invented in the 1950s by Sokal, Sneath and others motivated by
  - biological taxonomy

Cluster Analysis

- Automatic numerical methods for finding groups in data that are
  - separated
  - internally cohesive
- Invented in the 1950s by Sokal, Sneath and others motivated by
  - biological taxonomy
  - market segmentation

Cluster Analysis

- Automatic numerical methods for finding groups in data that are
  - separated
  - internally cohesive
- Invented in the 1950s by Sokal, Sneath and others motivated by
  - biological taxonomy
  - market segmentation
- Interest now driven by new types of data:

Cluster Analysis

- Automatic numerical methods for finding groups in data that are
  - separated
  - internally cohesive
- Invented in the 1950s by Sokal, Sneath and others motivated by
  - biological taxonomy
  - market segmentation
- Interest now driven by new types of data:
  - Gene expression microarray data

Cluster Analysis

- Automatic numerical methods for finding groups in data that are
  - separated
  - internally cohesive
- Invented in the 1950s by Sokal, Sneath and others motivated by
  - biological taxonomy
  - market segmentation
- Interest now driven by new types of data:
  - Gene expression microarray data
  - finding groups and patterns in retail barcode data

Cluster Analysis

- Automatic numerical methods for finding groups in data that are
  - separated
  - internally cohesive
- Invented in the 1950s by Sokal, Sneath and others motivated by
  - biological taxonomy
  - market segmentation
- Interest now driven by new types of data:
  - Gene expression microarray data
  - finding groups and patterns in retail barcode data
  - datamining more generally

Cluster Analysis

- Automatic numerical methods for finding groups in data that are
  - separated
  - internally cohesive
- Invented in the 1950s by Sokal, Sneath and others motivated by
  - biological taxonomy
  - market segmentation
- Interest now driven by new types of data:
  - Gene expression microarray data
  - finding groups and patterns in retail barcode data
  - datamining more generally
  - analysis of Web data (finding groups of users and sites) $\longrightarrow$ Collaborative filtering

Cluster Analysis

- Automatic numerical methods for finding groups in data that are
  - separated
  - internally cohesive
- Invented in the 1950s by Sokal, Sneath and others motivated by
  - biological taxonomy
  - market segmentation
- Interest now driven by new types of data:
  - Gene expression microarray data
  - finding groups and patterns in retail barcode data
  - datamining more generally
  - analysis of Web data (finding groups of users and sites) $\longrightarrow$ Collaborative filtering
  - medical image segmentation, e.g. for finding tumors. Here, cluster = group of pixels

Cluster Analysis

- Automatic numerical methods for finding groups in data that are
  - separated
  - internally cohesive
- Invented in the 1950s by Sokal, Sneath and others motivated by
  - biological taxonomy
  - market segmentation
- Interest now driven by new types of data:
  - Gene expression microarray data
  - finding groups and patterns in retail barcode data
  - datamining more generally
  - analysis of Web data (finding groups of users and sites) $\longrightarrow$ Collaborative filtering
  - medical image segmentation, e.g. for finding tumors. Here, cluster = group of pixels
  - color image quantization (e.g. for the Internet on mobile phones)

Cluster Analysis

- Automatic numerical methods for finding groups in data that are
  - separated
  - internally cohesive
- Invented in the 1950s by Sokal, Sneath and others motivated by
  - biological taxonomy
  - market segmentation
- Interest now driven by new types of data:
  - Gene expression microarray data
  - finding groups and patterns in retail barcode data
  - datamining more generally
  - analysis of Web data (finding groups of users and sites) $\longrightarrow$ Collaborative filtering
  - medical image segmentation, e.g. for finding tumors. Here, cluster = group of pixels
  - color image quantization (e.g. for the Internet on mobile phones)
  - automatic document clustering for technical documents and Web sites

Cluster Analysis Methods

Cluster Analysis Methods

- Most methods heuristic or algorithmic, not statistical, for example:

Cluster Analysis Methods

- Most methods heuristic or algorithmic, not statistical, for example:
  - complete link clustering

Cluster Analysis Methods

- Most methods heuristic or algorithmic, not statistical, for example:
  - complete link clustering
  - average link clustering

Cluster Analysis Methods

- Most methods heuristic or algorithmic, not statistical, for example:
  - complete link clustering
  - average link clustering
  - single link clustering

Cluster Analysis Methods

- Most methods heuristic or algorithmic, not statistical, for example:
  - complete link clustering
  - average link clustering
  - single link clustering
  - k means

Cluster Analysis Methods

- Most methods heuristic or algorithmic, not statistical, for example:
  - complete link clustering
  - average link clustering
  - single link clustering
  - k means
  - Ward's sum of squares

Cluster Analysis Methods

- Most methods heuristic or algorithmic, not statistical, for example:
  - complete link clustering
  - average link clustering
  - single link clustering
  - k means
  - Ward's sum of squares
- Difficulties: No principled basis for answering:

Cluster Analysis Methods

- Most methods heuristic or algorithmic, not statistical, for example:
  - complete link clustering
  - average link clustering
  - single link clustering
  - k means
  - Ward's sum of squares
- Difficulties: No principled basis for answering:
  - Which method to use?

Cluster Analysis Methods

- Most methods heuristic or algorithmic, not statistical, for example:
  - complete link clustering
  - average link clustering
  - single link clustering
  - k means
  - Ward's sum of squares
- Difficulties: No principled basis for answering:
  - Which method to use?
  - How many groups are there?

Cluster Analysis Methods

- Most methods heuristic or algorithmic, not statistical, for example:
  - complete link clustering
  - average link clustering
  - single link clustering
  - k means
  - Ward's sum of squares
- Difficulties: No principled basis for answering:
  - Which method to use?
  - How many groups are there?
  - What's the uncertainty in the results?

Cluster Analysis Methods

- Most methods heuristic or algorithmic, not statistical, for example:
  - complete link clustering
  - average link clustering
  - single link clustering
  - k means
  - Ward's sum of squares
- Difficulties: No principled basis for answering:
  - Which method to use?
  - How many groups are there?
  - What's the uncertainty in the results?
  - How to deal with outliers?

Cluster Analysis Methods

- Most methods heuristic or algorithmic, not statistical, for example:
  - complete link clustering
  - average link clustering
  - single link clustering
  - k means
  - Ward's sum of squares
- Difficulties: No principled basis for answering:
  - Which method to use?
  - How many groups are there?
  - What's the uncertainty in the results?
  - How to deal with outliers?
- Model-based clustering:

Cluster Analysis Methods

- Most methods heuristic or algorithmic, not statistical, for example:
  - complete link clustering
  - average link clustering
  - single link clustering
  - k means
  - Ward's sum of squares
- Difficulties: No principled basis for answering:
  - Which method to use?
  - How many groups are there?
  - What's the uncertainty in the results?
  - How to deal with outliers?
- Model-based clustering:
  - A *framework* for cluster analysis

Cluster Analysis Methods

- Most methods heuristic or algorithmic, not statistical, for example:
  - complete link clustering
  - average link clustering
  - single link clustering
  - k means
  - Ward's sum of squares
- Difficulties: No principled basis for answering:
  - Which method to use?
  - How many groups are there?
  - What's the uncertainty in the results?
  - How to deal with outliers?
- Model-based clustering:
  - A *framework* for cluster analysis
  - Bases cluster analysis on a statistical (mixture) model: $y \sim \sum_{g=1}^{G} \tau_g f_g(y)$, where y is data and $f_g(\cdot)$ are distributions

Cluster Analysis Methods

- Most methods heuristic or algorithmic, not statistical, for example:
  - complete link clustering
  - average link clustering
  - single link clustering
  - k means
  - Ward's sum of squares
- Difficulties: No principled basis for answering:
  - Which method to use?
  - How many groups are there?
  - What's the uncertainty in the results?
  - How to deal with outliers?
- Model-based clustering:
  - A *framework* for cluster analysis
  - Bases cluster analysis on a statistical (mixture) model: $y \sim \sum_{g=1}^{G} \tau_g f_g(y)$, where y is data and $f_g(\cdot)$ are distributions
  - Gives answers to questions based on standard statistical principles

Cluster Analysis Methods

- Most methods heuristic or algorithmic, not statistical, for example:
  - complete link clustering
  - average link clustering
  - single link clustering
  - k means
  - Ward's sum of squares
- Difficulties: No principled basis for answering:
  - Which method to use?
  - How many groups are there?
  - What's the uncertainty in the results?
  - How to deal with outliers?
- Model-based clustering:
  - A *framework* for cluster analysis
  - Bases cluster analysis on a statistical (mixture) model: $y \sim \sum_{g=1}^{G} \tau_g f_g(y)$, where y is data and $f_g(\cdot)$ are distributions
  - Gives answers to questions based on standard statistical principles
  - Here we focus on continuous data and take $f_g \sim$ multivariate normal

Basic Ideas of Model-Based Clustering

Basic Ideas of Model-Based Clustering

- Based on a finite mixture of multivariate normal distributions:

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

Basic Ideas of Model-Based Clustering

- Based on a finite mixture of multivariate normal distributions:

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

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

Basic Ideas of Model-Based Clustering

- Based on a finite mixture of multivariate normal distributions:

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

- where $\Sigma_g = \lambda_g D_g A_g D_g^T$
- λ_g = 1st eigenvalue of Σ_g: controls the *volume* of the gth cluster

Basic Ideas of Model-Based Clustering

- Based on a finite mixture of multivariate normal distributions:

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

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

Basic Ideas of Model-Based Clustering

- Based on a finite mixture of multivariate normal distributions:

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

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

Basic Ideas of Model-Based Clustering

- Based on a finite mixture of multivariate normal distributions:

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

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

Basic Ideas of Model-Based Clustering

- Based on a finite mixture of multivariate normal distributions:

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

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

Basic Ideas of Model-Based Clustering

- Based on a finite mixture of multivariate normal distributions:

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

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

Basic Ideas of Model-Based Clustering

- Based on a finite mixture of multivariate normal distributions:

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

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

Basic Ideas of Model-Based Clustering

- Based on a finite mixture of multivariate normal distributions:

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

- where $\Sigma_g = \lambda_g D_g A_g D_g^T$
- λ_g = 1st eigenvalue of Σ_g: controls the *volume* of the gth cluster
- $A_g = \mathrm{diag}\{1, \alpha_{2g}, \ldots, \alpha_{dg}\}$
  - controls the *shape* of the gth cluster
  - $(1 \geq \alpha_2 \geq \ldots \geq \alpha_d > 0)$
  - E.g. α_2 close to zero: Cluster g concentrated about a line.
  - E.g. $\alpha_{2g}, \ldots, \alpha_{dg}$ all close to 1: Cluster g nearly spherical.
- D_g = Eigenvectors: Control the *orientation* of the gth cluster
- Different clustering models can be obtained by constraining each of *volume*, *shape* and *orientation* to be constant across clusters, or by allowing them to vary (Banfield & Raftery, 1993, *Biometrics*)

Model-Based Clustering Strategy

Model-Based Clustering Strategy

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

Model-Based Clustering Strategy

- Maximum likelihood estimation for the mixture model parameters $\theta = (\tau, \mu, \Sigma)$, via the EM algorithm
- Initialization of EM via hierarchical agglomerative model-based clustering, in which the groups merged at each stage are those that minimize the decrease in likelihood.

Model-Based Clustering Strategy

- Maximum likelihood estimation for the mixture model parameters $\theta = (\tau, \mu, \Sigma)$, via the EM algorithm
- Initialization of EM via hierarchical agglomerative model-based clustering, in which the groups merged at each stage are those that minimize the decrease in likelihood.
- Choosing the Number of Clusters and the Clustering Method/Model:

Model-Based Clustering Strategy

- Maximum likelihood estimation for the mixture model parameters $\theta = (\tau, \mu, \Sigma)$, via the EM algorithm
- Initialization of EM via hierarchical agglomerative model-based clustering, in which the groups merged at each stage are those that minimize the decrease in likelihood.
- Choosing the Number of Clusters and the Clustering Method/Model:
  - Both are reduced to statistical model selection problems, and solved simultaneously.

Model-Based Clustering Strategy

- Maximum likelihood estimation for the mixture model parameters $\theta = (\tau, \mu, \Sigma)$, via the EM algorithm
- Initialization of EM via hierarchical agglomerative model-based clustering, in which the groups merged at each stage are those that minimize the decrease in likelihood.
- Choosing the Number of Clusters and the Clustering Method/Model:
  - Both are reduced to statistical model selection problems, and solved simultaneously.
  - Each combination of (Number of Clusters, Clustering Model) is viewed as a separate statistical model

Model-Based Clustering Strategy

- Maximum likelihood estimation for the mixture model parameters $\theta = (\tau, \mu, \Sigma)$, via the EM algorithm
- Initialization of EM via hierarchical agglomerative model-based clustering, in which the groups merged at each stage are those that minimize the decrease in likelihood.
- Choosing the Number of Clusters and the Clustering Method/Model:
  - Both are reduced to statistical model selection problems, and solved simultaneously.
  - Each combination of (Number of Clusters, Clustering Model) is viewed as a separate statistical model
  - We use the Bayes factor, i.e. the ratio of posterior to prior odds for one model against another.

Model-Based Clustering Strategy

- Maximum likelihood estimation for the mixture model parameters $\theta = (\tau, \mu, \Sigma)$, via the EM algorithm
- Initialization of EM via hierarchical agglomerative model-based clustering, in which the groups merged at each stage are those that minimize the decrease in likelihood.
- Choosing the Number of Clusters and the Clustering Method/Model:
  - Both are reduced to statistical model selection problems, and solved simultaneously.
  - Each combination of (Number of Clusters, Clustering Model) is viewed as a separate statistical model
  - We use the Bayes factor, i.e. the ratio of posterior to prior odds for one model against another.
  - This allows comparison of the multiple, nonnested models considered.

Model-Based Clustering Strategy

- Maximum likelihood estimation for the mixture model parameters $\theta = (\tau, \mu, \Sigma)$, via the EM algorithm
- Initialization of EM via hierarchical agglomerative model-based clustering, in which the groups merged at each stage are those that minimize the decrease in likelihood.
- Choosing the Number of Clusters and the Clustering Method/Model:
  - Both are reduced to statistical model selection problems, and solved simultaneously.
  - Each combination of (Number of Clusters, Clustering Model) is viewed as a separate statistical model
  - We use the Bayes factor, i.e. the ratio of posterior to prior odds for one model against another.
  - This allows comparison of the multiple, nonnested models considered.
  - We approximate the Bayes factors via

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

Model-Based Clustering Strategy

- Maximum likelihood estimation for the mixture model parameters $\theta = (\tau, \mu, \Sigma)$, via the EM algorithm
- Initialization of EM via hierarchical agglomerative model-based clustering, in which the groups merged at each stage are those that minimize the decrease in likelihood.
- Choosing the Number of Clusters and the Clustering Method/Model:
  - Both are reduced to statistical model selection problems, and solved simultaneously.
  - Each combination of (Number of Clusters, Clustering Model) is viewed as a separate statistical model
  - We use the Bayes factor, i.e. the ratio of posterior to prior odds for one model against another.
  - This allows comparison of the multiple, nonnested models considered.
  - We approximate the Bayes factors via

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

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

Example: Diabetes Diagnosis

Example: Diabetes Diagnosis

- Data: Glucose, insulin and SSPG measurements on 145 patients (Reuven and Miller 1979).

Example: Diabetes Diagnosis

- Data: Glucose, insulin and SSPG measurements on 145 patients (Reuven and Miller 1979).
- Goal: Use these to diagnose patients as one of "Normal," "Chemical Diabetes," or "Overt Diabetes."

Example: Diabetes Diagnosis

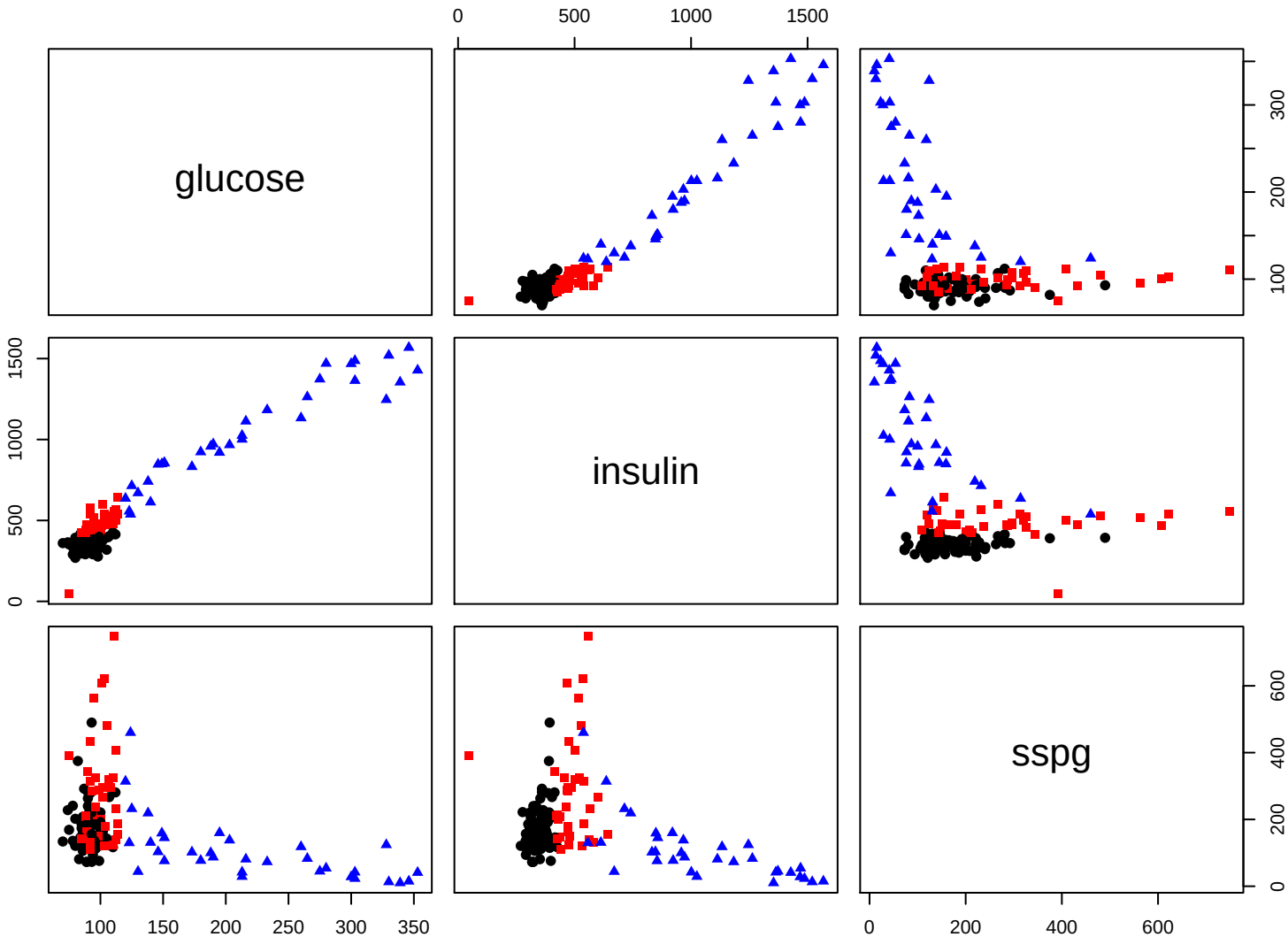
- Data: Glucose, insulin and SSPG measurements on 145 patients (Reuven and Miller 1979).
- Goal: Use these to diagnose patients as one of "Normal," "Chemical Diabetes," or "Overt Diabetes."
- There is a clinical classification that we will ignore in the clustering, but that we will use to evaluate it.

Example: Diabetes Diagnosis

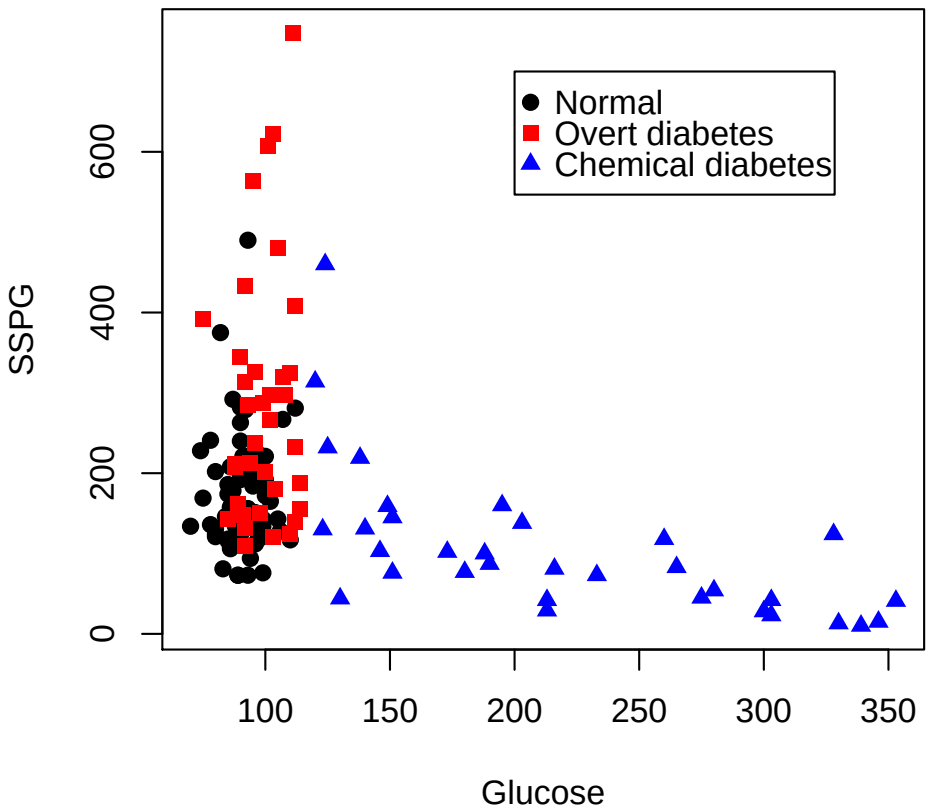
- Data: Glucose, insulin and SSPG measurements on 145 patients (Reuven and Miller 1979).
- Goal: Use these to diagnose patients as one of "Normal," "Chemical Diabetes," or "Overt Diabetes."
- There is a clinical classification that we will ignore in the clustering, but that we will use to evaluate it.
- Many clustering methods require that we "know" the number of clusters, but model-based clustering does not.

Diabetes Data

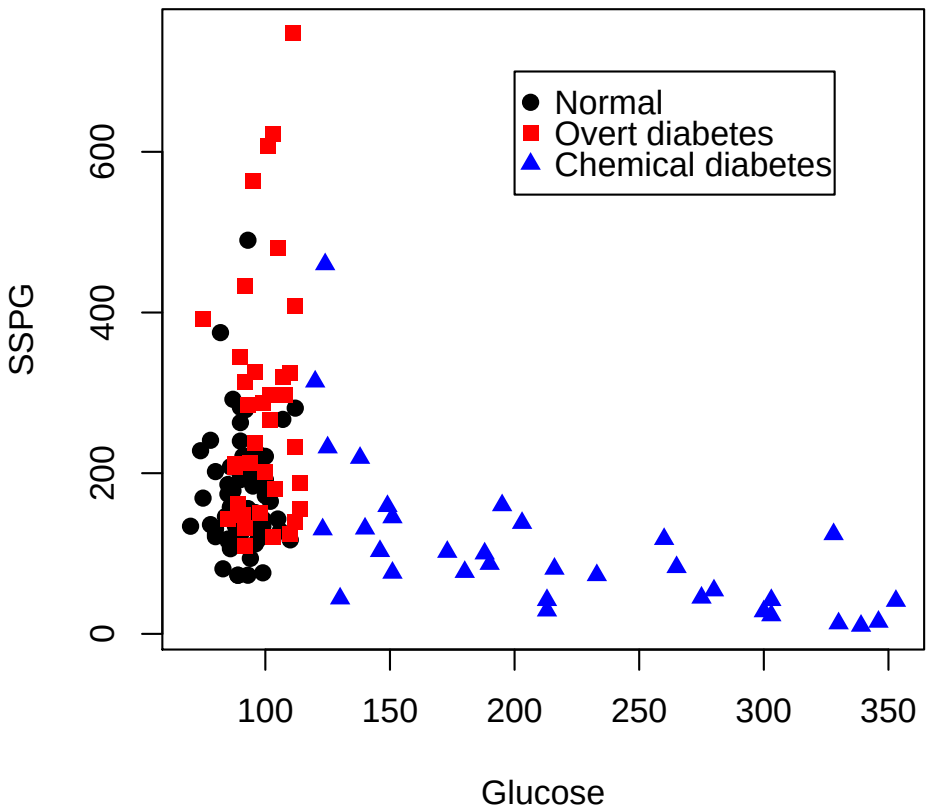
Diabetes Data



Diabetes Data

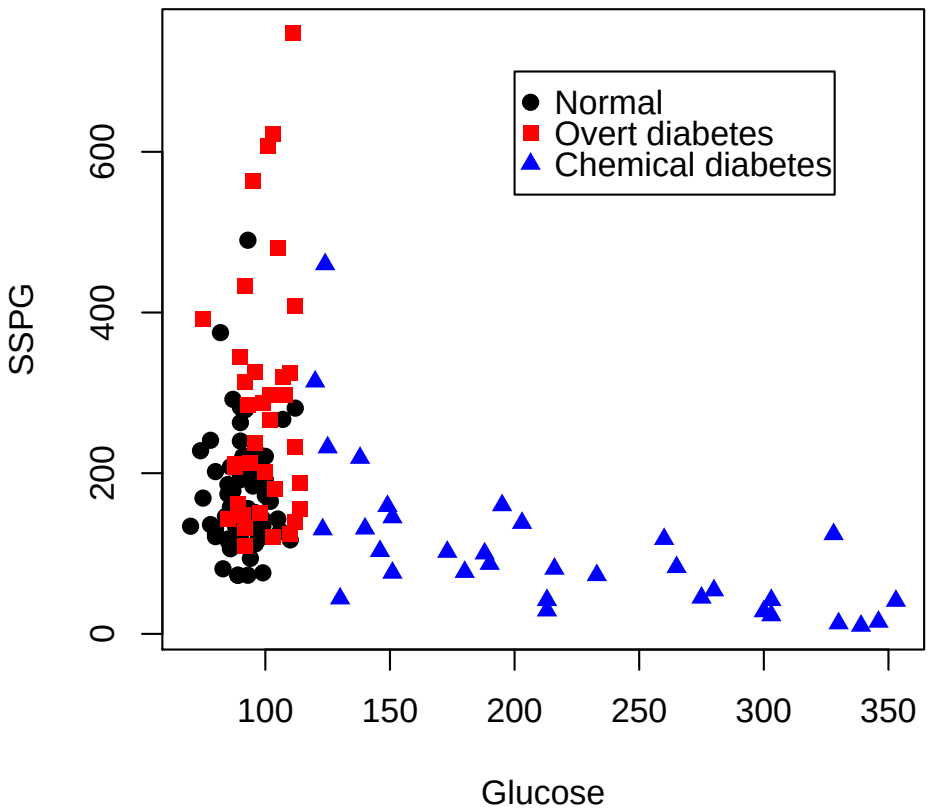


Diabetes Data



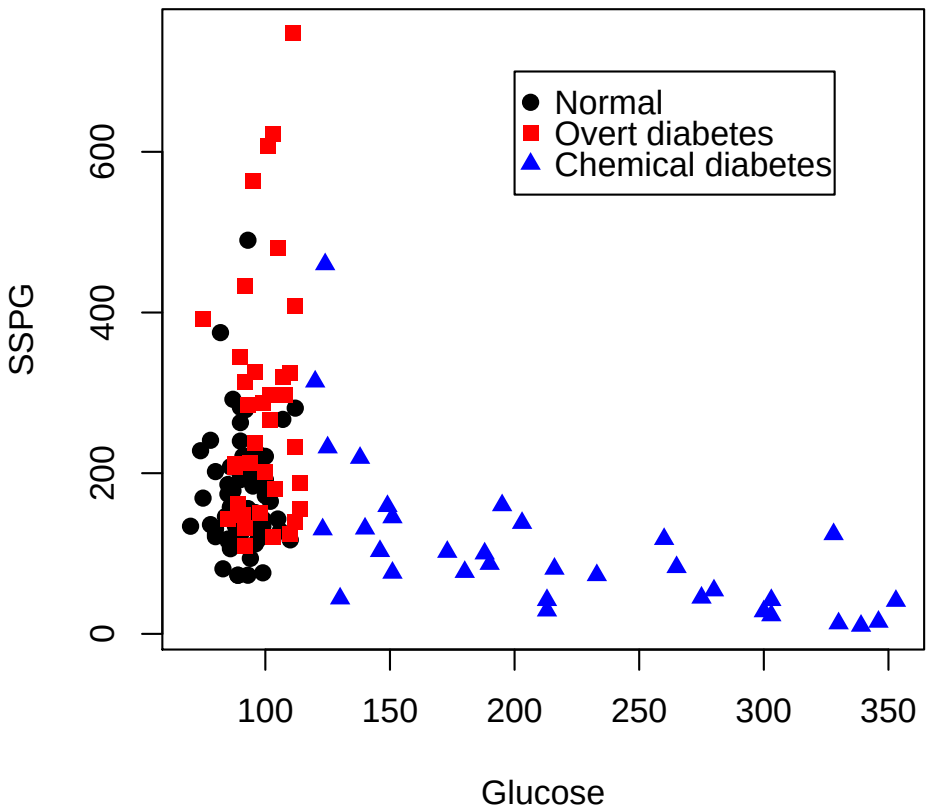
- We know the "right" model:

Diabetes Data



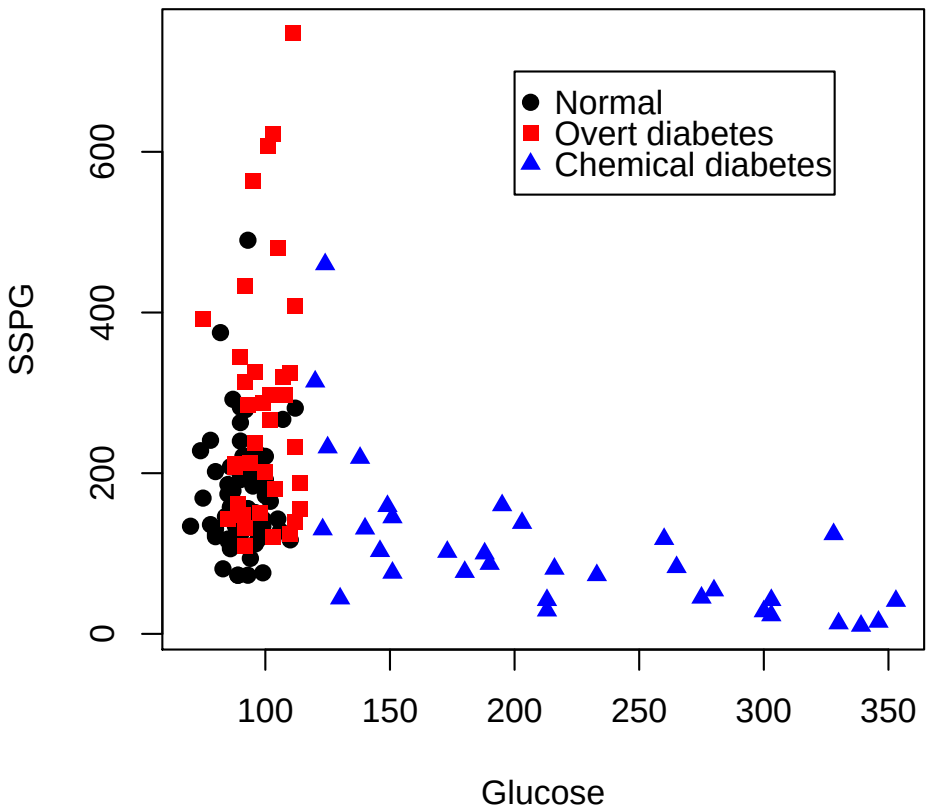
- We know the "right" model:
  - The volume (Normal: small; Diabetes clusters: bigger)

Diabetes Data



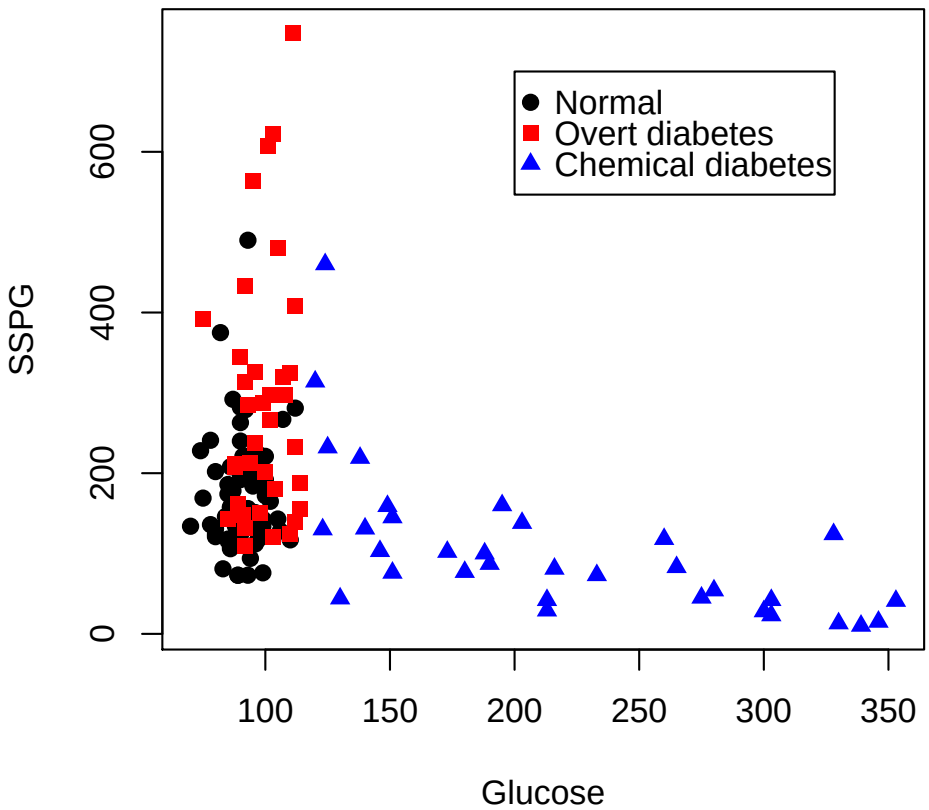
- We know the "right" model:
  - The volume (Normal: small; Diabetes clusters: bigger)
  - shape (Normal: spherical, Diabetes clusters: long and thin)

Diabetes Data



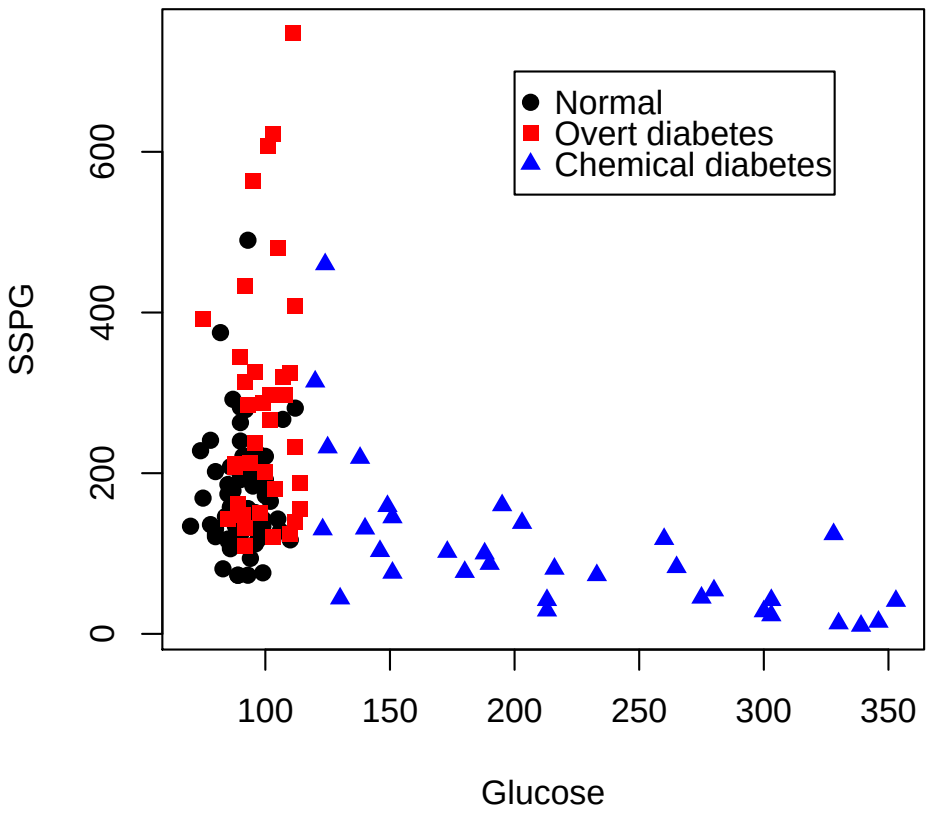
- We know the "right" model:
  - The volume (Normal: small; Diabetes clusters: bigger)
  - shape (Normal: spherical, Diabetes clusters: long and thin)
  - and orientation (Chemical and Overt: orthogonal)

Diabetes Data



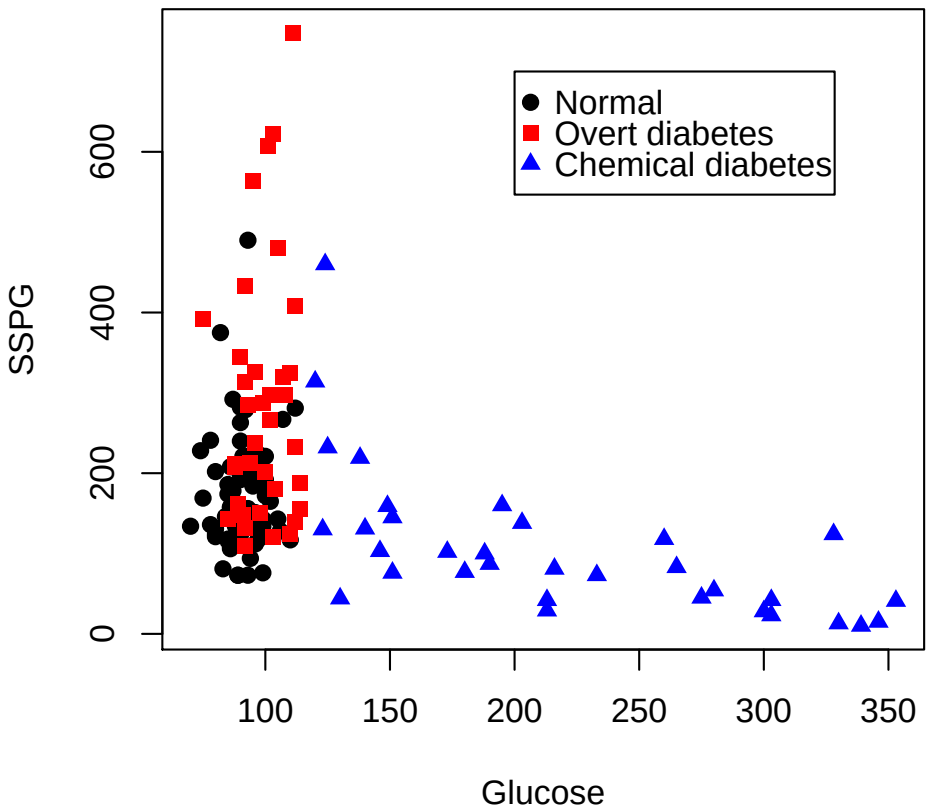
- We know the "right" model:
  - The volume (Normal: small; Diabetes clusters: bigger)
  - shape (Normal: spherical, Diabetes clusters: long and thin)
  - and orientation (Chemical and Overt: orthogonal)
  - are all different

Diabetes Data



- We know the "right" model:
  - The volume (Normal: small; Diabetes clusters: bigger)
  - shape (Normal: spherical, Diabetes clusters: long and thin)
  - and orientation (Chemical and Overt: orthogonal)
  - are all different
- $\Longrightarrow$ Model is Σ_g all different (unconstrained)

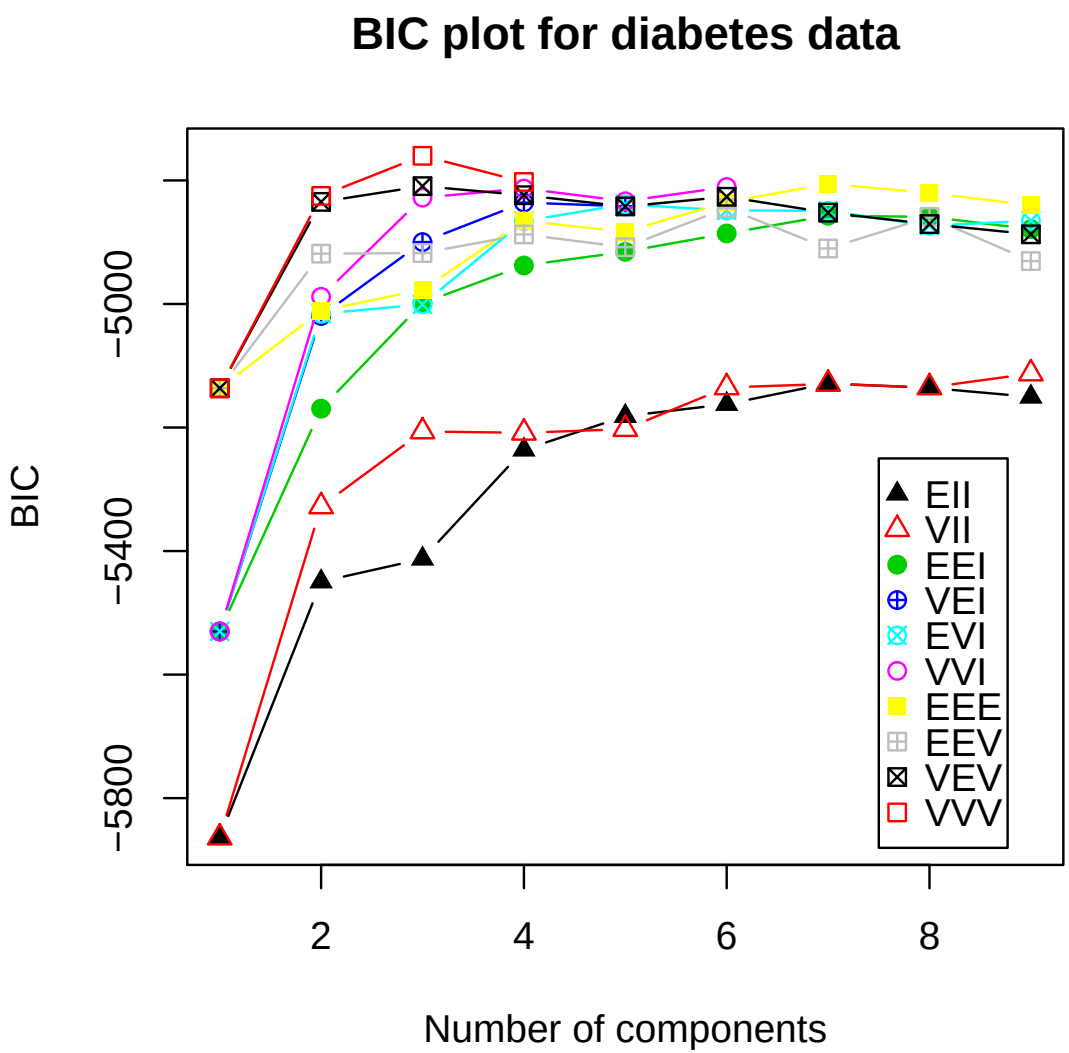
Diabetes Data



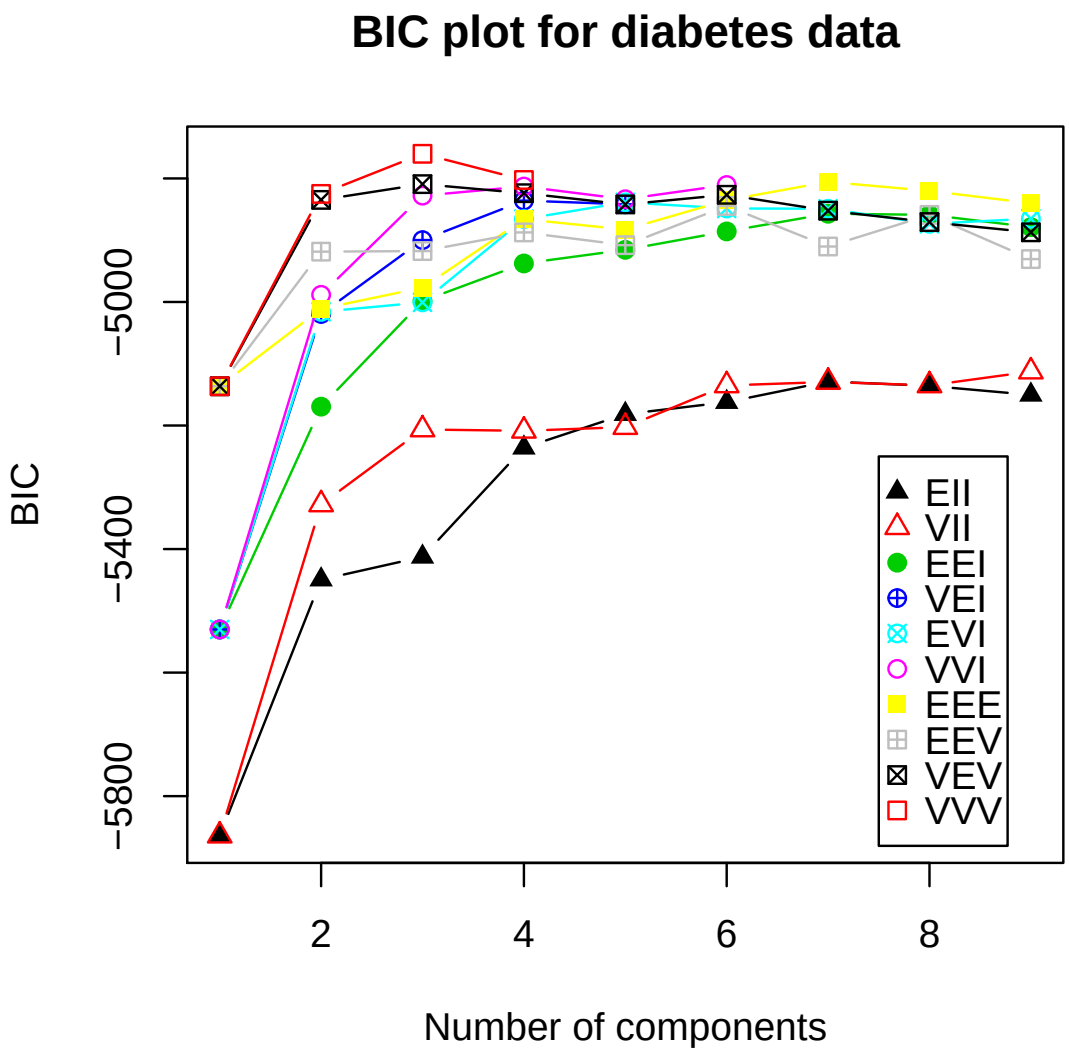
- We know the "right" model:
  - The volume (Normal: small; Diabetes clusters: bigger)
  - shape (Normal: spherical, Diabetes clusters: long and thin)
  - and orientation (Chemical and Overt: orthogonal)
  - are all different
- $\Longrightarrow$ Model is Σ_g all different (unconstrained)
- The `mclust` R package is used

Choosing the Number of Clusters and the Clustering Model

Choosing the Number of Clusters and the Clustering Model

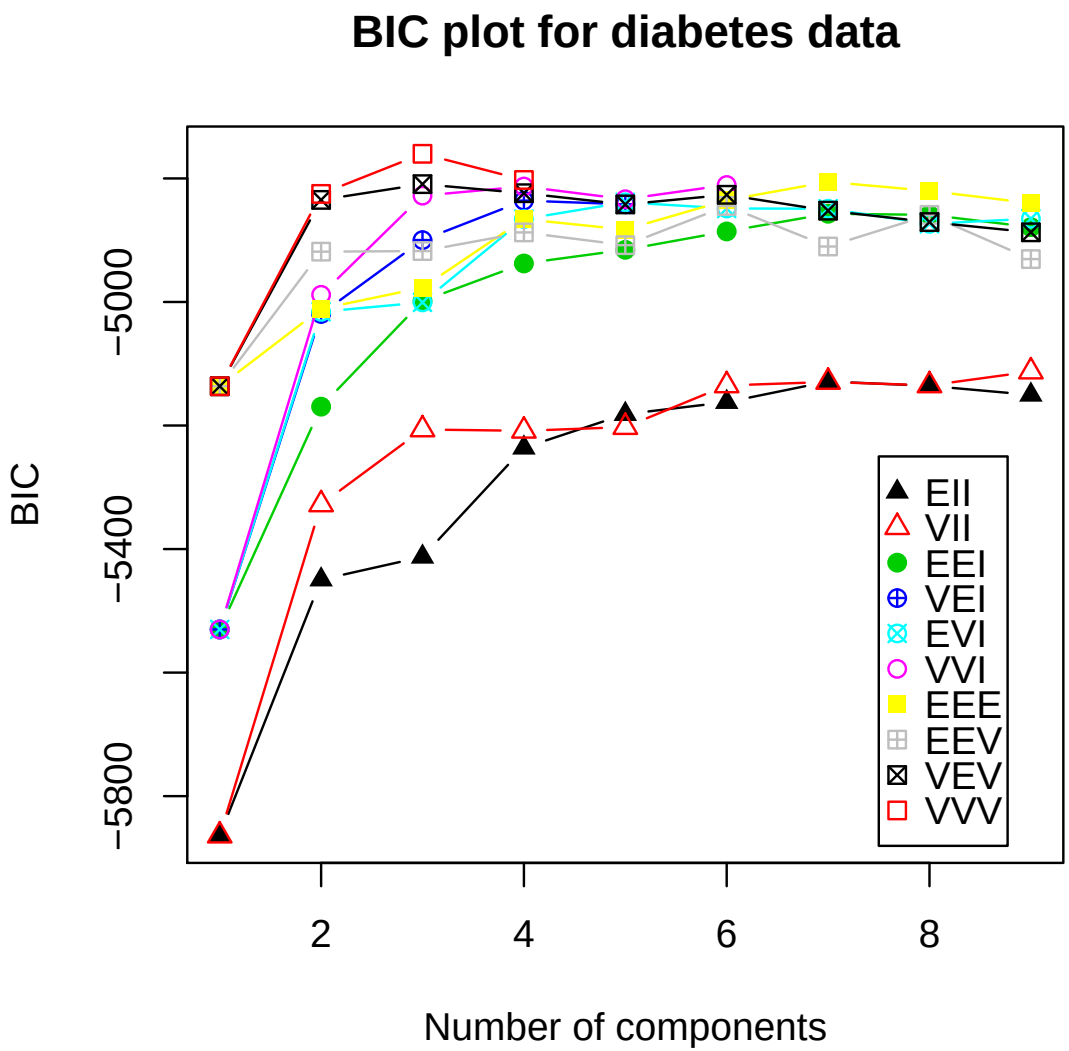


Choosing the Number of Clusters and the Clustering Model



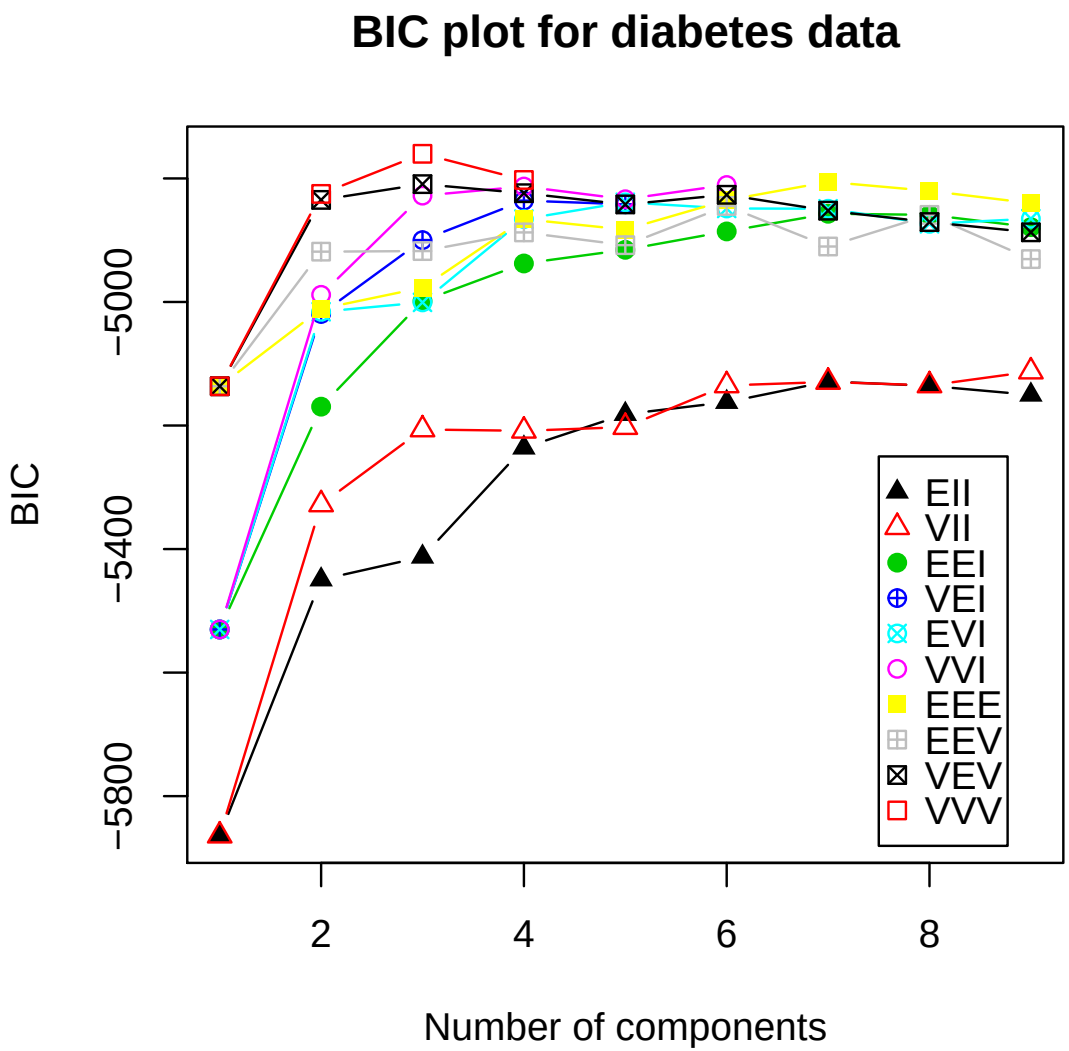
- Model code:

Choosing the Number of Clusters and the Clustering Model



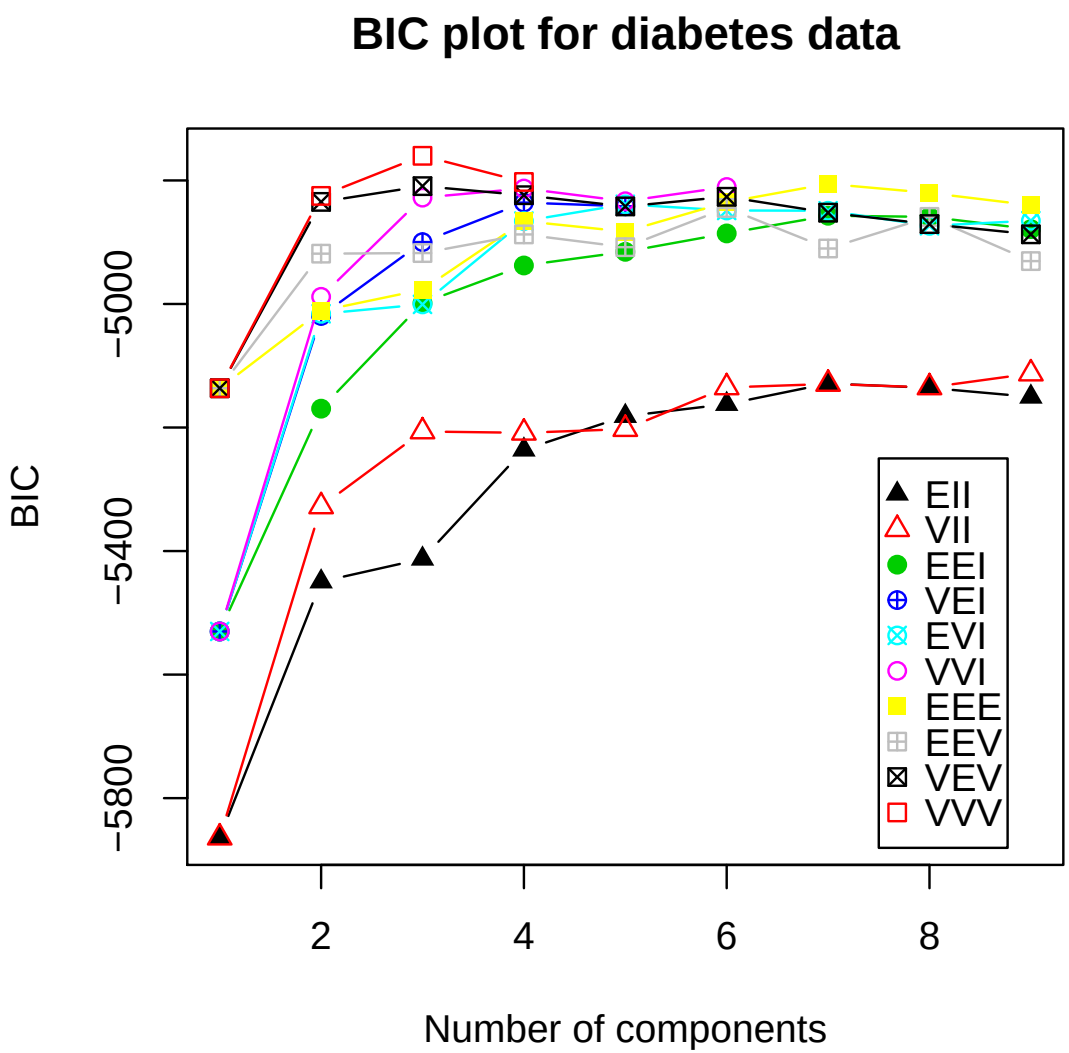
- Model code:
  - Letters refer to (Volume, Shape, Orientation):

Choosing the Number of Clusters and the Clustering Model



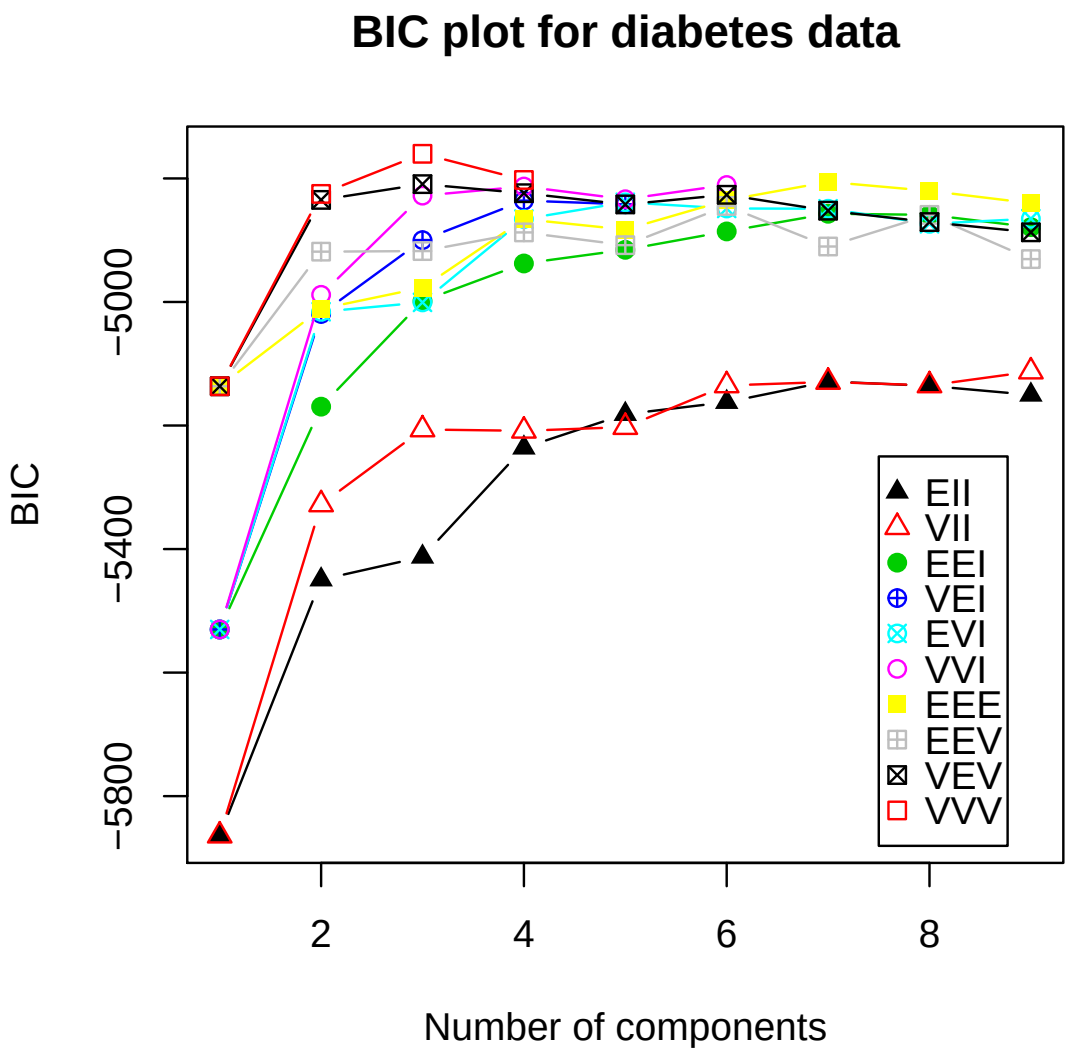
- Model code:
  - Letters refer to (Volume, Shape, Orientation):
    - **E**: **E**qual across clusters

Choosing the Number of Clusters and the Clustering Model



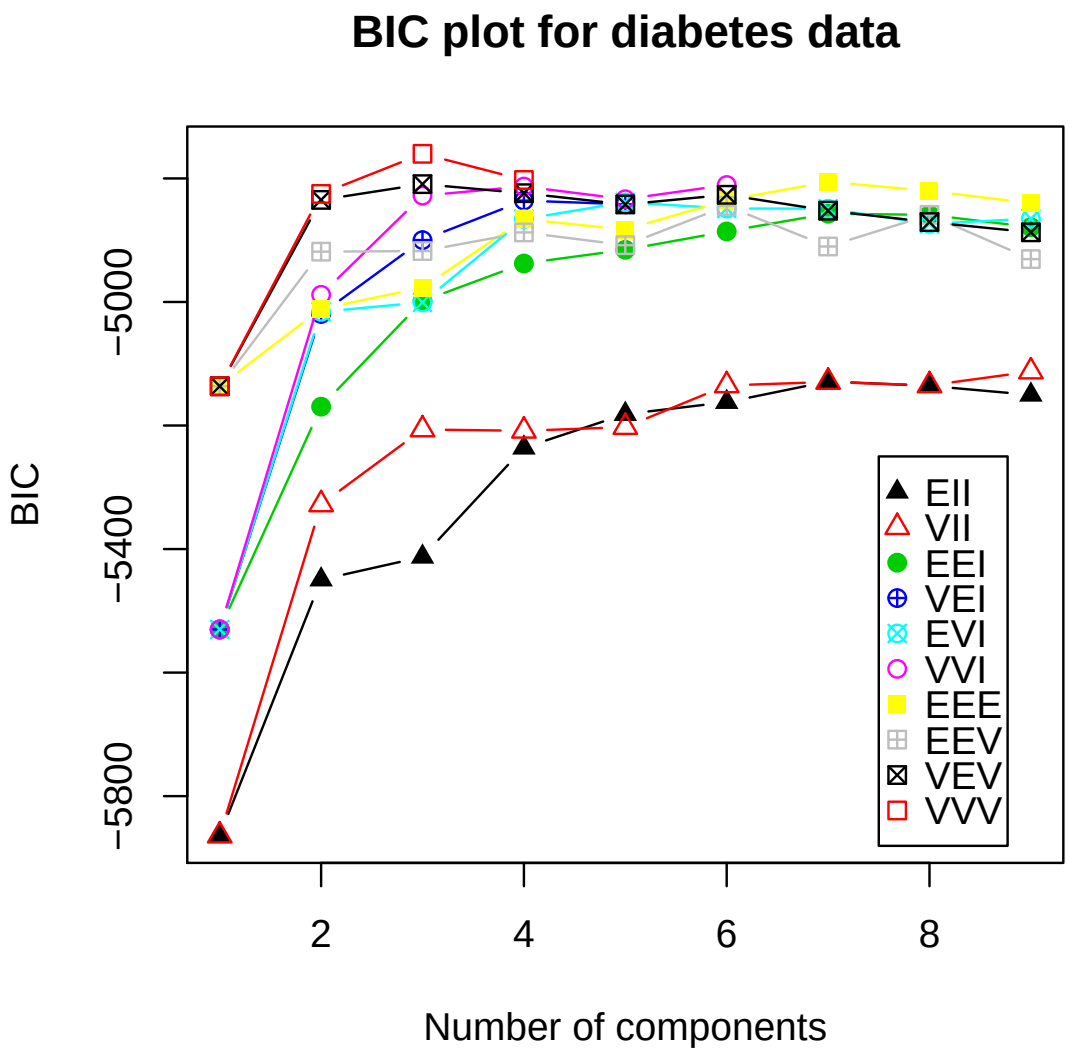
- Model code:
  - Letters refer to (Volume, Shape, Orientation):
    - **E**: **E**qual across clusters
    - **V**: **V**ary across clusters

Choosing the Number of Clusters and the Clustering Model



- Model code:
  - Letters refer to (Volume, Shape, Orientation):
    - **E**: **E**qual across clusters
    - **V**: **V**ary across clusters
    - **I**: **I**dentity (spherical) covariance matrix

Choosing the Number of Clusters and the Clustering Model

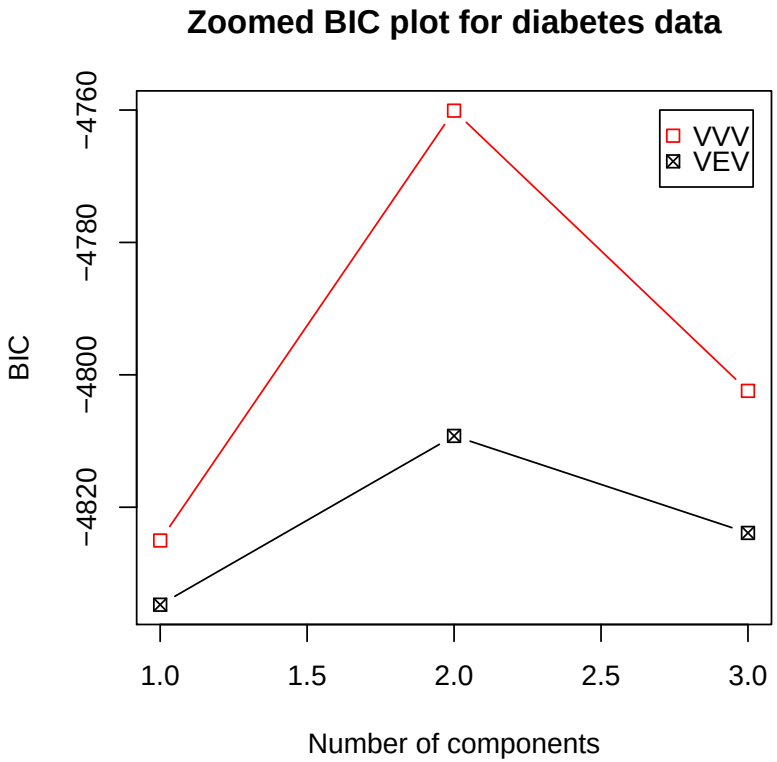


- Model code:
  - Letters refer to (Volume, Shape, Orientation):
    - **E**: **E**qual across clusters
    - **V**: **V**ary across clusters
    - **I**: **I**dentity (spherical) covariance matrix
  - Example: **EEV**: Equal volume, equal shape, varying orientations

The Choice of Number of Clusters and Clustering Model

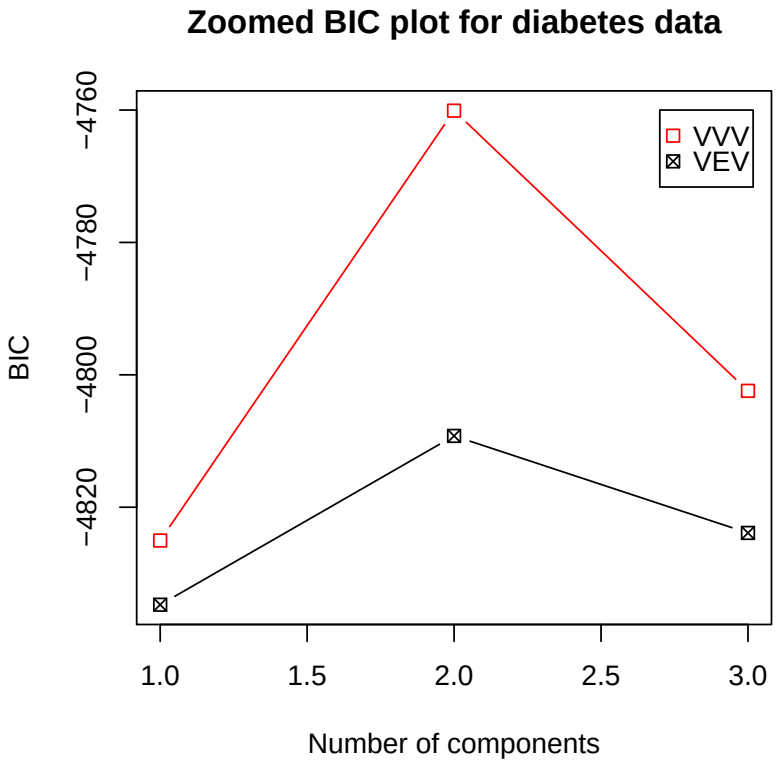
The Choice of Number of Clusters and Clustering Model

The Choice of Number of Clusters and Clustering Model



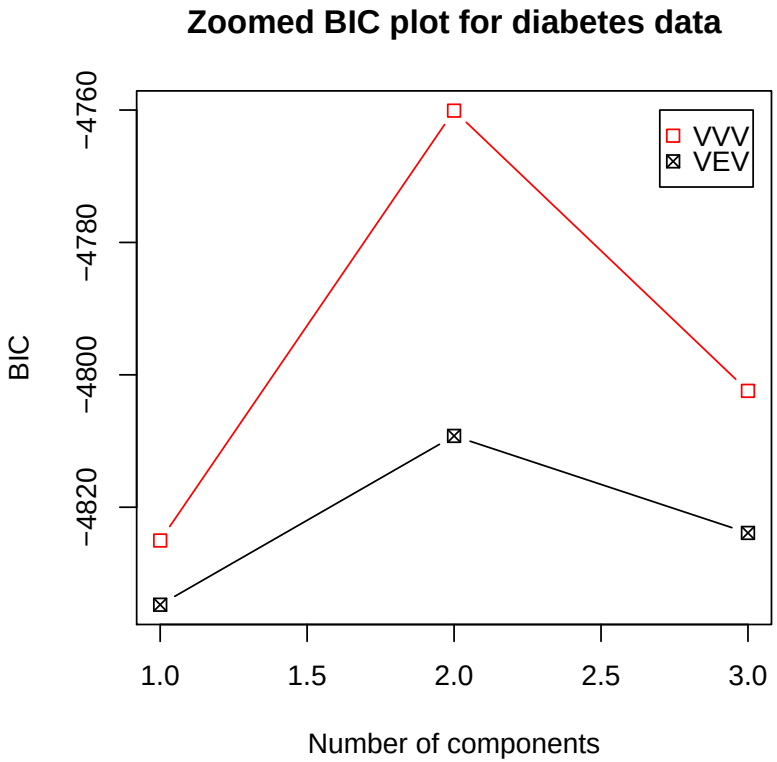
- BIC chooses the unconstrained (VVV) model with 3 clusters.

The Choice of Number of Clusters and Clustering Model



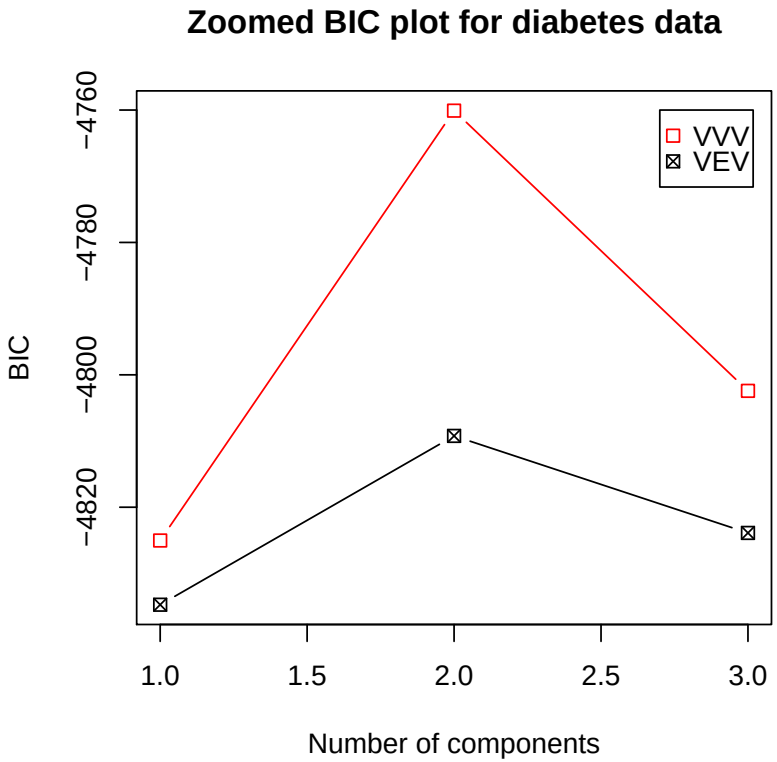
- BIC chooses the unconstrained (VVV) model with 3 clusters.
- The right answer!

The Choice of Number of Clusters and Clustering Model



- BIC chooses the unconstrained (VVV) model with 3 clusters.
- The right answer!

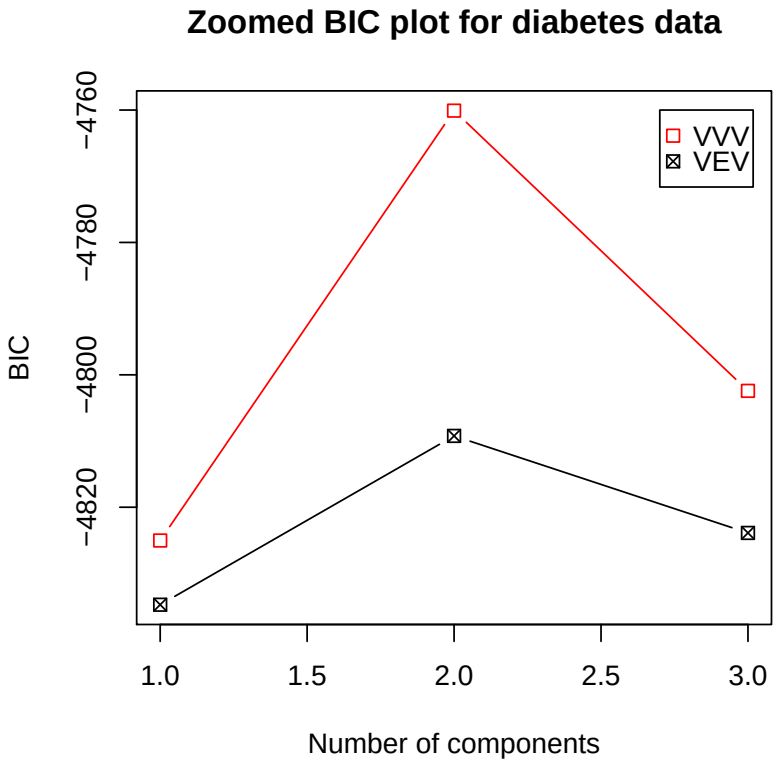
The Choice of Number of Clusters and Clustering Model



- The EII model, $\Sigma_g = \lambda I$ ($\approx k$ means) not good.

- BIC chooses the unconstrained (VVV) model with 3 clusters.
- The right answer!

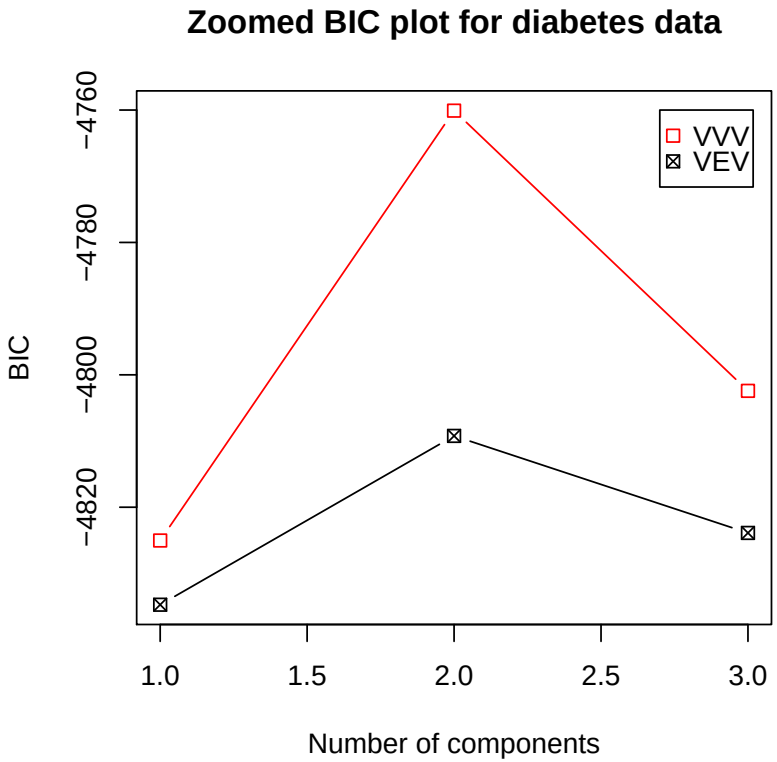
The Choice of Number of Clusters and Clustering Model



- The EII model, $\Sigma_g = \lambda I$ ($\approx k$ means) not good.
  - Thus k means would not be good for these data.

- BIC chooses the unconstrained (VVV) model with 3 clusters.
- The right answer!

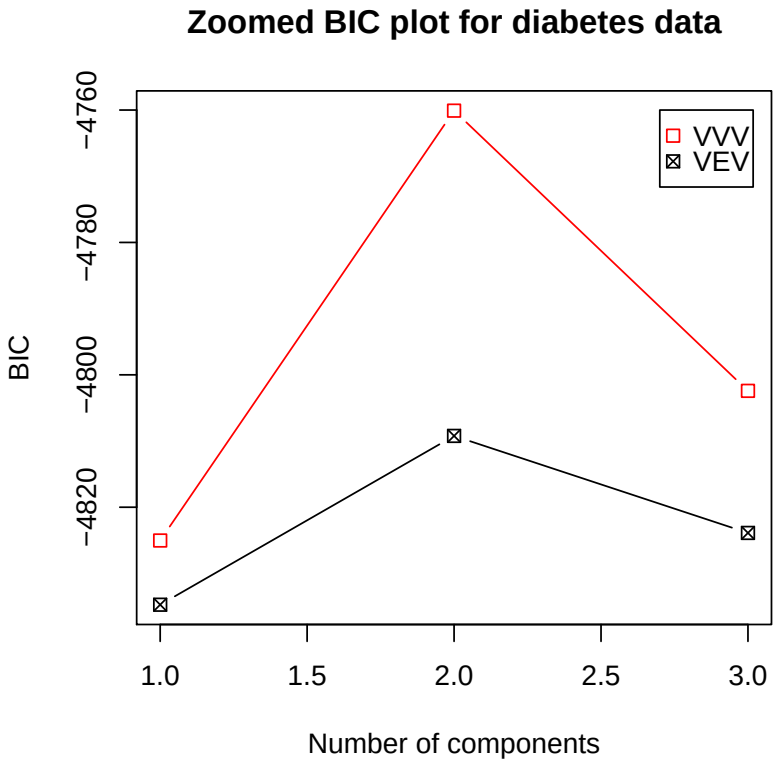
The Choice of Number of Clusters and Clustering Model



- The EII model, $\Sigma_g = \lambda I$ ($\approx k$ means) not good.
  - Thus k means would not be good for these data.
  - BIC allows us to assess when k means, or other methods, would work well.

- BIC chooses the unconstrained (VVV) model with 3 clusters.
- The right answer!

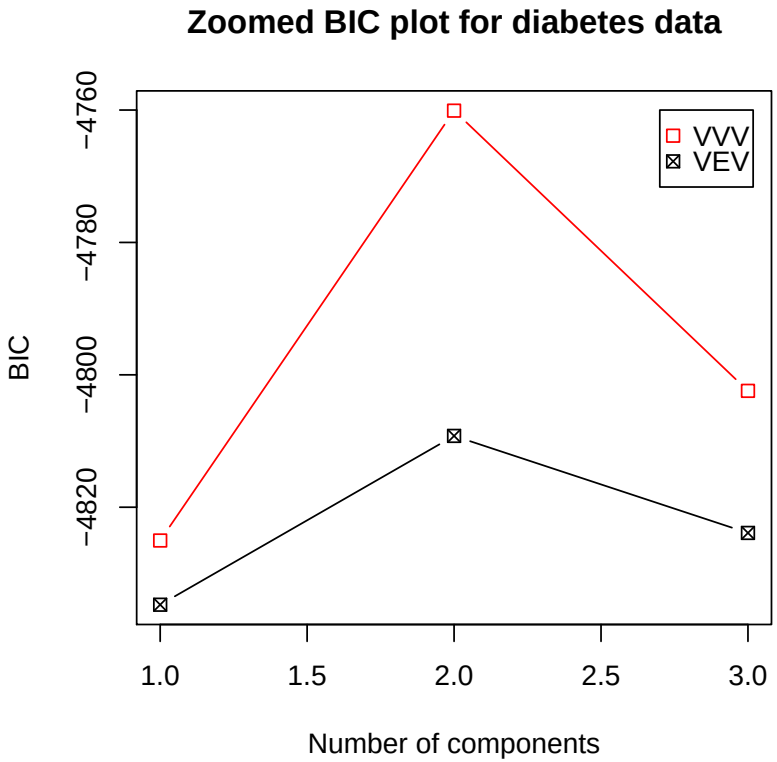
The Choice of Number of Clusters and Clustering Model



- The EII model, $\Sigma_g = \lambda I$ ($\approx k$ means) not good.
  - Thus k means would not be good for these data.
  - BIC allows us to assess when k means, or other methods, would work well.
- Tradeoff between the clustering model and the number of clusters:

- BIC chooses the unconstrained (VVV) model with 3 clusters.
- The right answer!

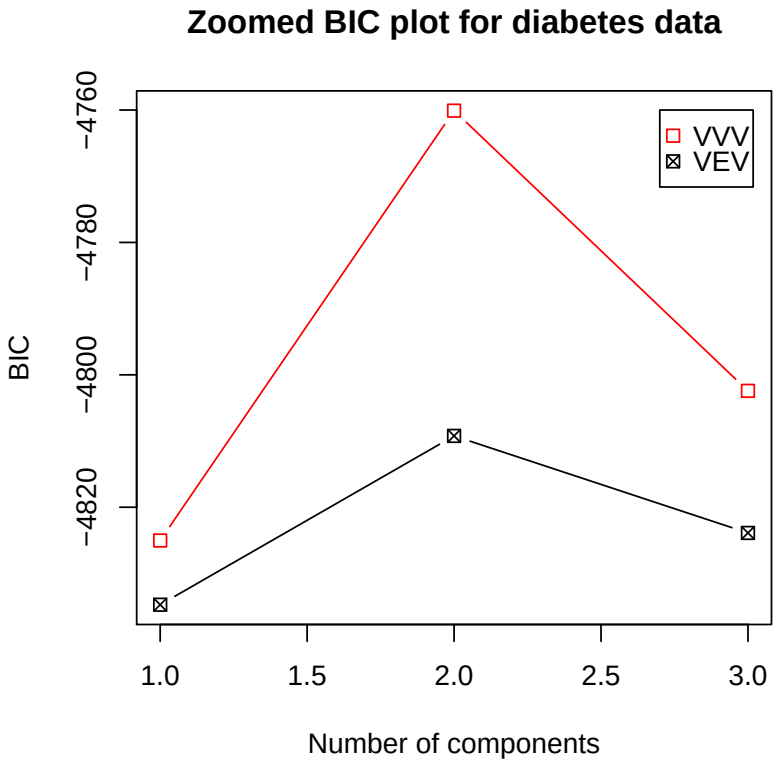
The Choice of Number of Clusters and Clustering Model



- BIC chooses the unconstrained (VVV) model with 3 clusters.
- The right answer!

- The EII model, $\Sigma_g = \lambda I$ ($\approx k$ means) not good.
  - Thus k means would not be good for these data.
  - BIC allows us to assess when k means, or other methods, would work well.
- Tradeoff between the clustering model and the number of clusters:
  - E.g. with the EII model (equal volume spherical), far more clusters are needed than with the VVV model (unconstrained ellipses).

The Choice of Number of Clusters and Clustering Model

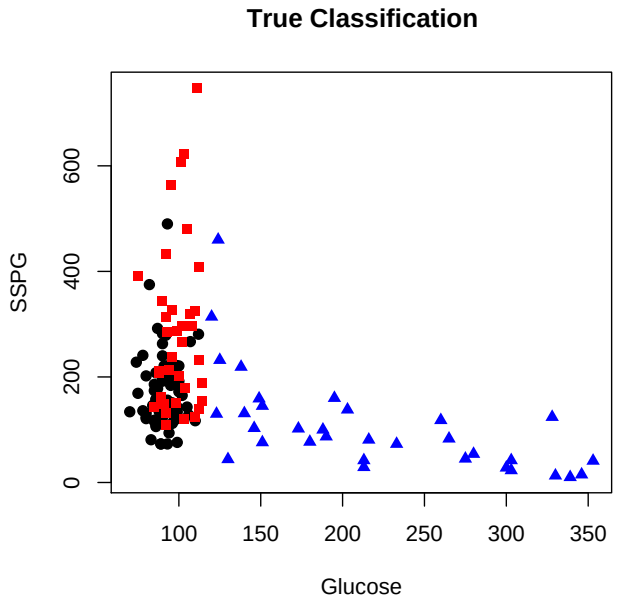


- BIC chooses the unconstrained (VVV) model with 3 clusters.
- The right answer!

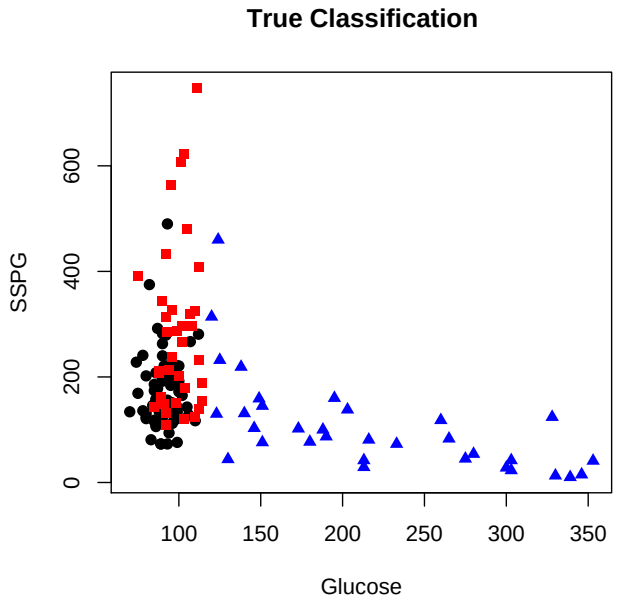
- The EII model, $\Sigma_g = \lambda I$ ($\approx k$ means) not good.
  - Thus k means would not be good for these data.
  - BIC allows us to assess when k means, or other methods, would work well.
- Tradeoff between the clustering model and the number of clusters:
  - E.g. with the EII model (equal volume spherical), far more clusters are needed than with the VVV model (unconstrained ellipses).
  - Thus BIC determines whether it is better to use the "peas" or the "pod."

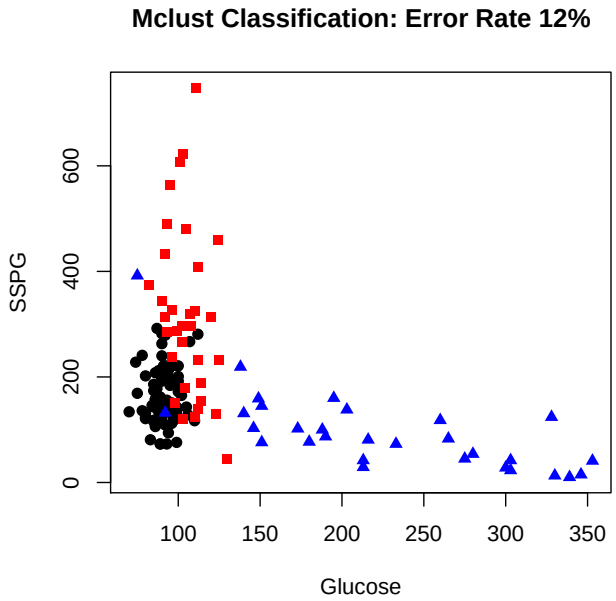
Comparisons Between Methods

Comparisons Between Methods



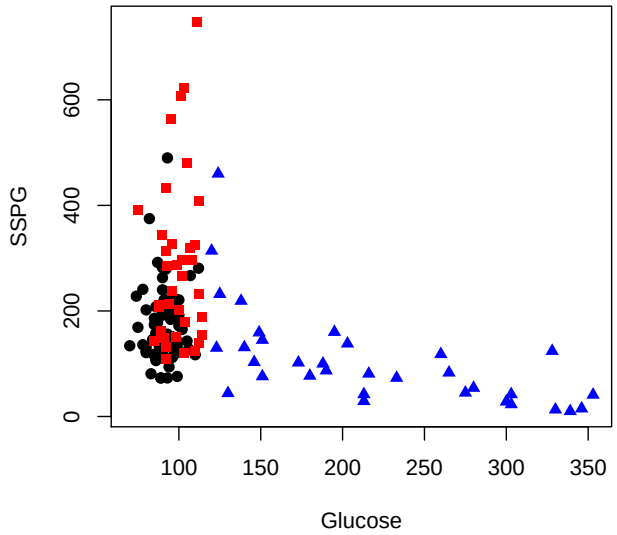
Comparisons Between Methods



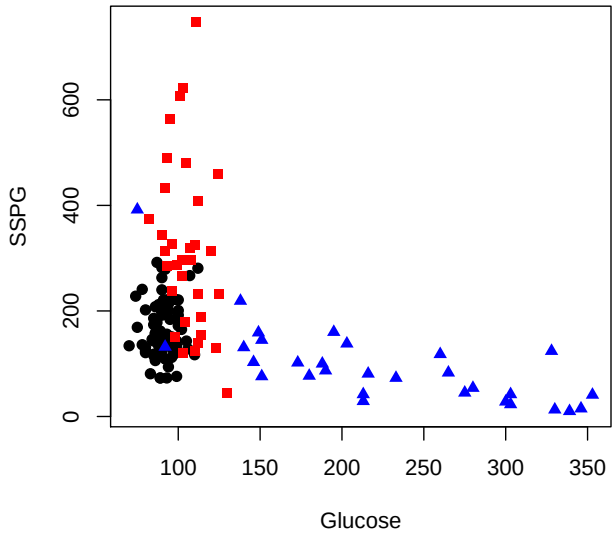


Comparisons Between Methods

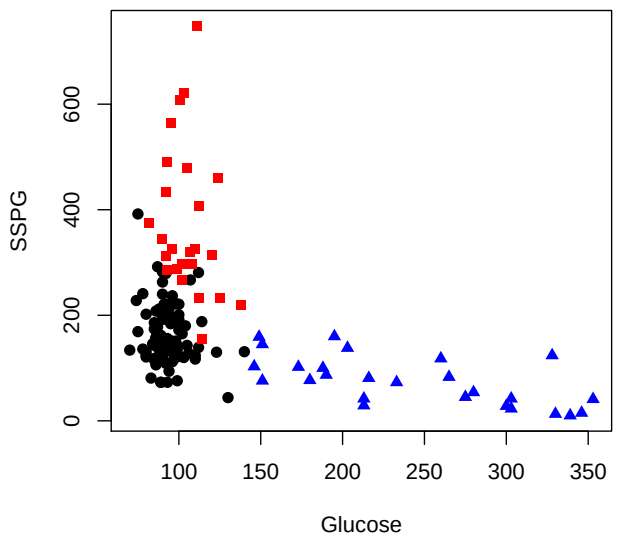
True Classification



Mclust Classification: Error Rate 12%

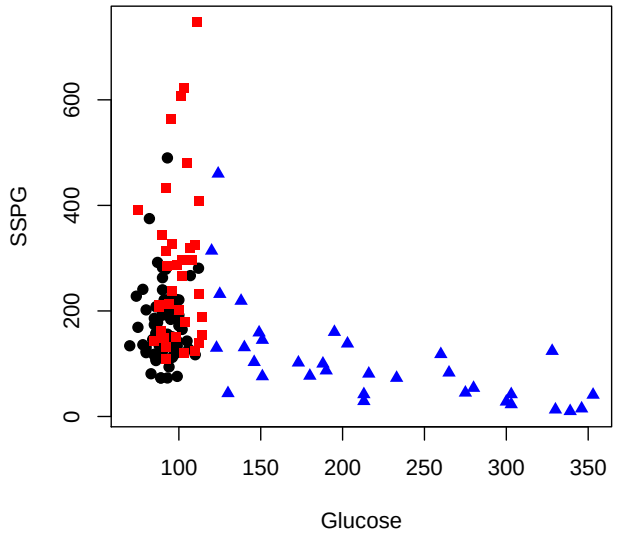


K means Classification: Error rate 18%

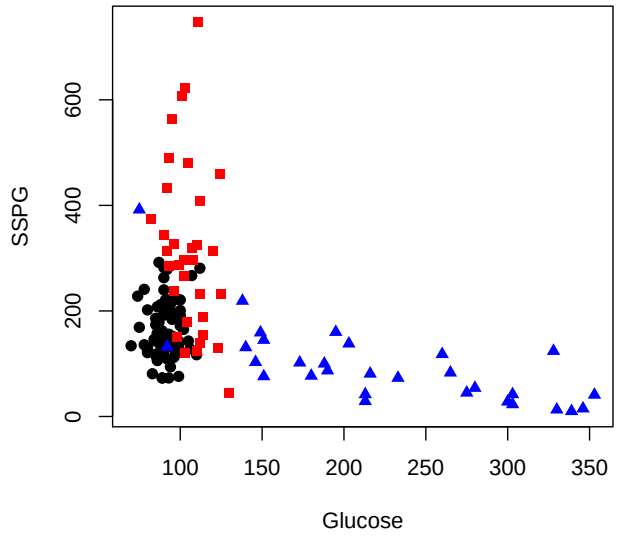


Comparisons Between Methods

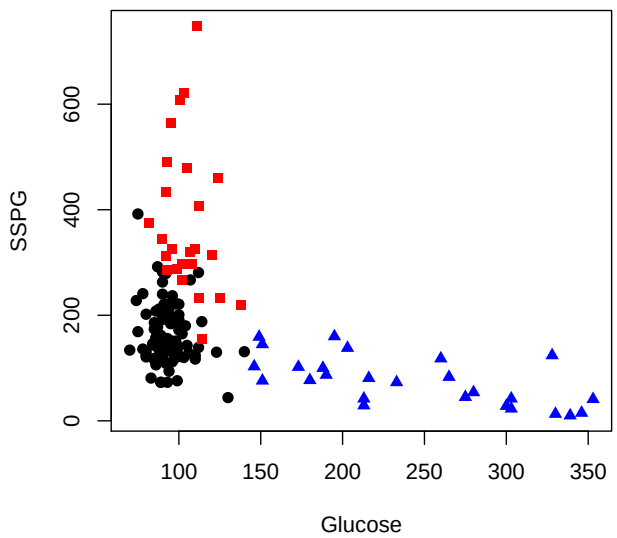
True Classification



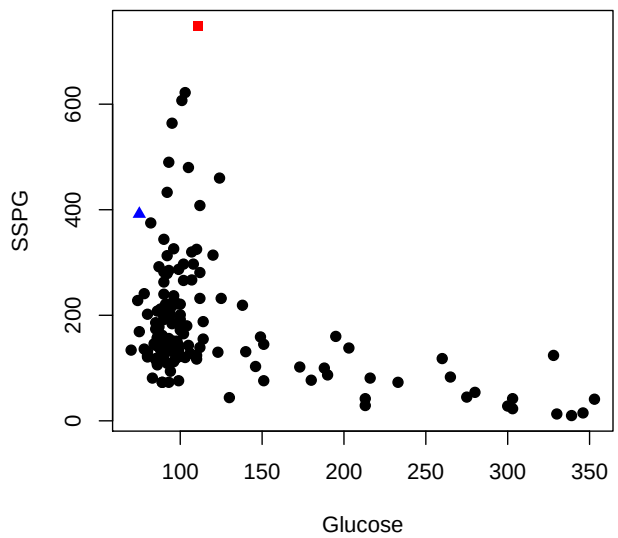
Mclust Classification: Error Rate 12%



K means Classification: Error rate 18%



Single Link Classification: Error rate 47%



Outliers in Model-Based Clustering

Outliers in Model-Based Clustering

- The model is expanded to explicitly include outliers:

Outliers in Model-Based Clustering

- The model is expanded to explicitly include outliers:
  - Outliers arise from a low-intensity Poisson process on the "data region," R.

Outliers in Model-Based Clustering

- The model is expanded to explicitly include outliers:
  - Outliers arise from a low-intensity Poisson process on the "data region," R.
  - $\Longrightarrow$ outliers are generated from a uniform distribution on the data region.

Outliers in Model-Based Clustering

- The model is expanded to explicitly include outliers:
  - Outliers arise from a low-intensity Poisson process on the "data region," R.
  - $\Longrightarrow$ outliers are generated from a uniform distribution on the data region.
- Expanded mixture model:

$$y_i \sim \sum_{g=1}^{G} \tau_g \mathrm{MVN}_d(\mu_g, \Sigma_g) + \tau_0 U(R)$$

Outliers in Model-Based Clustering

- The model is expanded to explicitly include outliers:
  - Outliers arise from a low-intensity Poisson process on the "data region," R.
  - $\Longrightarrow$ outliers are generated from a uniform distribution on the data region.
- Expanded mixture model:

$$y_i \sim \sum_{g=1}^{G} \tau_g \mathrm{MVN}_d(\mu_g, \Sigma_g) + \tau_0 U(R)$$

- Proceed as before with EM and BIC

Outliers in Model-Based Clustering

- The model is expanded to explicitly include outliers:
  - Outliers arise from a low-intensity Poisson process on the "data region," R.
  - $\Longrightarrow$ outliers are generated from a uniform distribution on the data region.
- Expanded mixture model:

$$y_i \sim \sum_{g=1}^{G} \tau_g \mathrm{MVN}_d(\mu_g, \Sigma_g) + \tau_0 U(R)$$

- Proceed as before with EM and BIC
- This has good robustness properties (Hennig 2004, Ann Stat)

Image Segmentation Application: Finding Regions of Interest in Dynamic Breast MRI

Image Segmentation Application: Finding Regions of Interest in Dynamic Breast MRI

- Breast tumor detection is usually done using X-ray mammography

Image Segmentation Application: Finding Regions of Interest in Dynamic Breast MRI

- Breast tumor detection is usually done using X-ray mammography
- This has a high false positive rate, leading to

Image Segmentation Application: Finding Regions of Interest in Dynamic Breast MRI

- Breast tumor detection is usually done using X-ray mammography
- This has a high false positive rate, leading to
  - many unnecessary biopsies

Image Segmentation Application: Finding Regions of Interest in Dynamic Breast MRI

- Breast tumor detection is usually done using X-ray mammography
- This has a high false positive rate, leading to
  - many unnecessary biopsies
  - unnecessary deaths

Image Segmentation Application: Finding Regions of Interest in Dynamic Breast MRI

- Breast tumor detection is usually done using X-ray mammography
- This has a high false positive rate, leading to
  - many unnecessary biopsies
  - unnecessary deaths
  - search for a better method

Image Segmentation Application: Finding Regions of Interest in Dynamic Breast MRI

- Breast tumor detection is usually done using X-ray mammography
- This has a high false positive rate, leading to
  - many unnecessary biopsies
  - unnecessary deaths
  - search for a better method
- An alternative: Dynamic Magnetic Resonance Imaging (MRI):

Image Segmentation Application: Finding Regions of Interest in Dynamic Breast MRI

- Breast tumor detection is usually done using X-ray mammography
- This has a high false positive rate, leading to
  - many unnecessary biopsies
  - unnecessary deaths
  - search for a better method
- An alternative: Dynamic Magnetic Resonance Imaging (MRI):
  - Patient injected with a contrast agent, *Gaudolinium*

Image Segmentation Application: Finding Regions of Interest in Dynamic Breast MRI

- Breast tumor detection is usually done using X-ray mammography
- This has a high false positive rate, leading to
  - many unnecessary biopsies
  - unnecessary deaths
  - search for a better method
- An alternative: Dynamic Magnetic Resonance Imaging (MRI):
  - Patient injected with a contrast agent, *Gaudolinium*
  - 3-d images made every 10 seconds for about 4 minutes (25 images)

Image Segmentation Application: Finding Regions of Interest in Dynamic Breast MRI

- Breast tumor detection is usually done using X-ray mammography
- This has a high false positive rate, leading to
  - many unnecessary biopsies
  - unnecessary deaths
  - search for a better method
- An alternative: Dynamic Magnetic Resonance Imaging (MRI):
  - Patient injected with a contrast agent, *Gaudolinium*
  - 3-d images made every 10 seconds for about 4 minutes (25 images)
- Our goal: Automatically find a region of interest that may contain the tumor. Method:

Image Segmentation Application: Finding Regions of Interest in Dynamic Breast MRI

- Breast tumor detection is usually done using X-ray mammography
- This has a high false positive rate, leading to
  - many unnecessary biopsies
  - unnecessary deaths
  - search for a better method
- An alternative: Dynamic Magnetic Resonance Imaging (MRI):
  - Patient injected with a contrast agent, *Gaudolinium*
  - 3-d images made every 10 seconds for about 4 minutes (25 images)
- Our goal: Automatically find a region of interest that may contain the tumor. Method:
  - Each slice analyzed separately; best results used

Image Segmentation Application: Finding Regions of Interest in Dynamic Breast MRI

- Breast tumor detection is usually done using X-ray mammography
- This has a high false positive rate, leading to
  - many unnecessary biopsies
  - unnecessary deaths
  - search for a better method
- An alternative: Dynamic Magnetic Resonance Imaging (MRI):
  - Patient injected with a contrast agent, *Gaudolinium*
  - 3-d images made every 10 seconds for about 4 minutes (25 images)
- Our goal: Automatically find a region of interest that may contain the tumor. Method:
  - Each slice analyzed separately; best results used
  - Each voxel has a 25-dimensional intensity measurement

Image Segmentation Application: Finding Regions of Interest in Dynamic Breast MRI

- Breast tumor detection is usually done using X-ray mammography
- This has a high false positive rate, leading to
  - many unnecessary biopsies
  - unnecessary deaths
  - search for a better method
- An alternative: Dynamic Magnetic Resonance Imaging (MRI):
  - Patient injected with a contrast agent, *Gaudolinium*
  - 3-d images made every 10 seconds for about 4 minutes (25 images)
- Our goal: Automatically find a region of interest that may contain the tumor. Method:
  - Each slice analyzed separately; best results used
  - Each voxel has a 25-dimensional intensity measurement
  - Reduced to 5 variables: Time to peak, Difference at peak, ...

Image Segmentation Application: Finding Regions of Interest in Dynamic Breast MRI

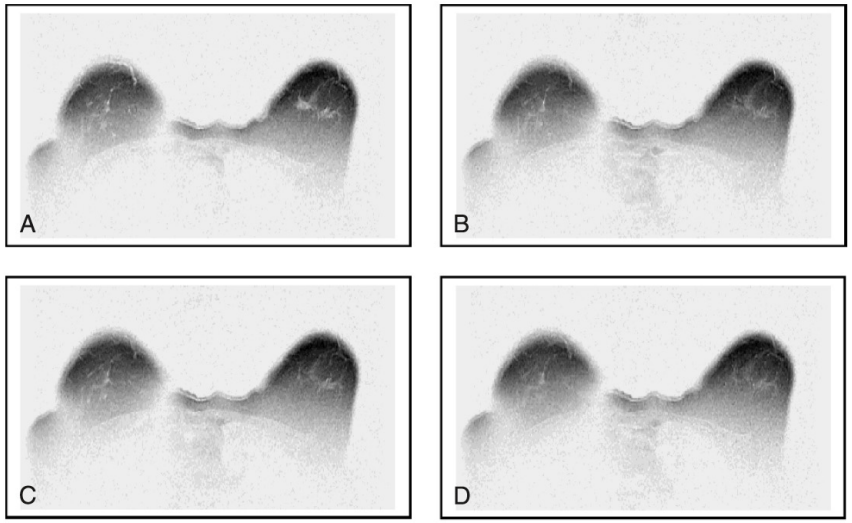
- Breast tumor detection is usually done using X-ray mammography
- This has a high false positive rate, leading to
  - many unnecessary biopsies
  - unnecessary deaths
  - search for a better method
- An alternative: Dynamic Magnetic Resonance Imaging (MRI):
  - Patient injected with a contrast agent, *Gaudolinium*
  - 3-d images made every 10 seconds for about 4 minutes (25 images)
- Our goal: Automatically find a region of interest that may contain the tumor. Method:
  - Each slice analyzed separately; best results used
  - Each voxel has a 25-dimensional intensity measurement
  - Reduced to 5 variables: Time to peak, Difference at peak, ...
  - Mclust with $G = 3$ (background, heart, skin) and $G = 4$ (same + tumor) groups

Image Segmentation Application: Finding Regions of Interest in Dynamic Breast MRI

- Breast tumor detection is usually done using X-ray mammography
- This has a high false positive rate, leading to
  - many unnecessary biopsies
  - unnecessary deaths
  - search for a better method
- An alternative: Dynamic Magnetic Resonance Imaging (MRI):
  - Patient injected with a contrast agent, *Gaudolinium*
  - 3-d images made every 10 seconds for about 4 minutes (25 images)
- Our goal: Automatically find a region of interest that may contain the tumor. Method:
  - Each slice analyzed separately; best results used
  - Each voxel has a 25-dimensional intensity measurement
  - Reduced to 5 variables: Time to peak, Difference at peak, ...
  - Mclust with $G = 3$ (background, heart, skin) and $G = 4$ (same + tumor) groups
  - Bayesian morphology: Fast Bayesian image restoration via mathematical morphology (Forbes and Raftery 1999, JASA)

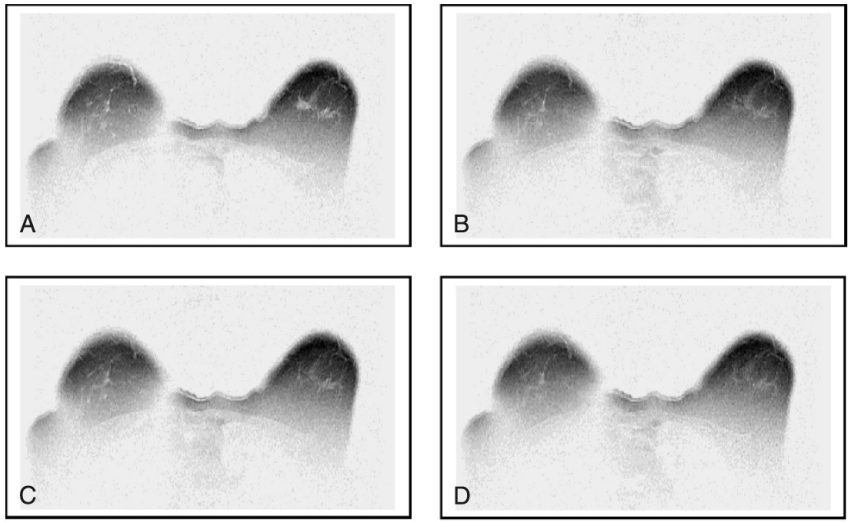
Breast MRI Example Results

Breast MRI Example Results

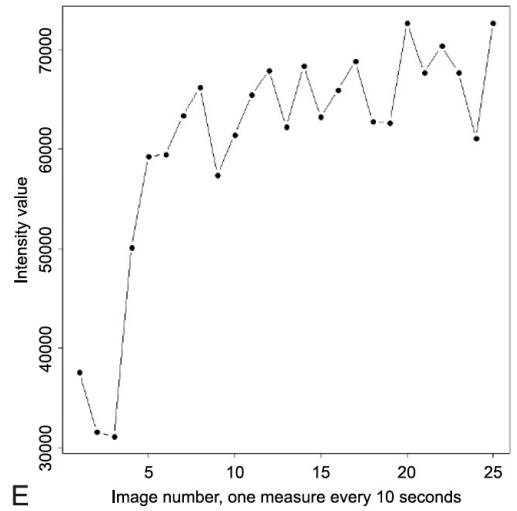


Images at 10, 70, 150, 250 seconds

Breast MRI Example Results

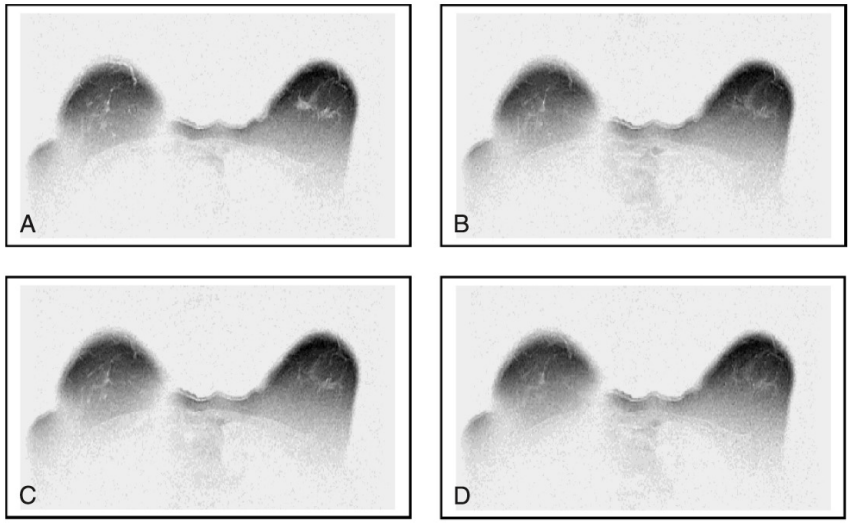


Images at 10, 70, 150, 250 seconds

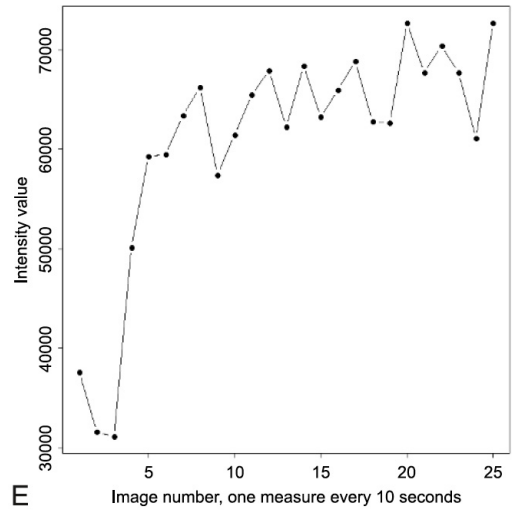


Intensity curve for one voxel

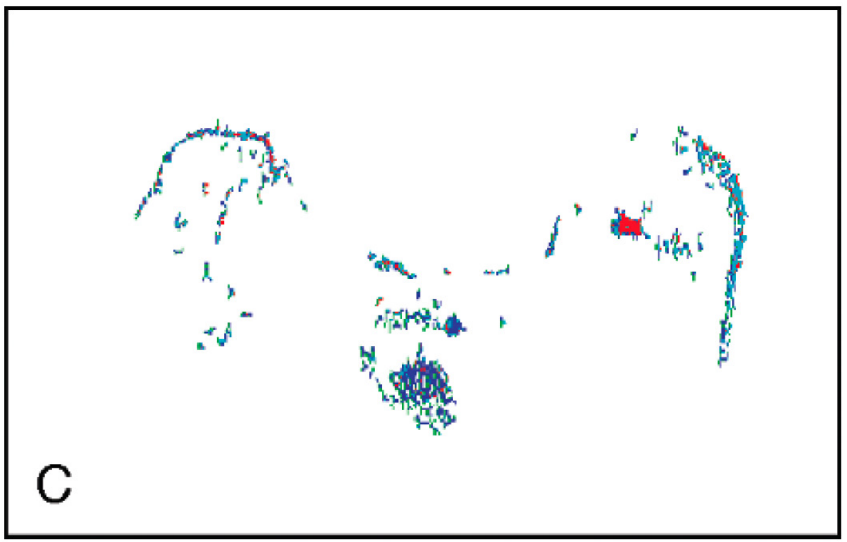
Breast MRI Example Results



Images at 10, 70, 150, 250 seconds

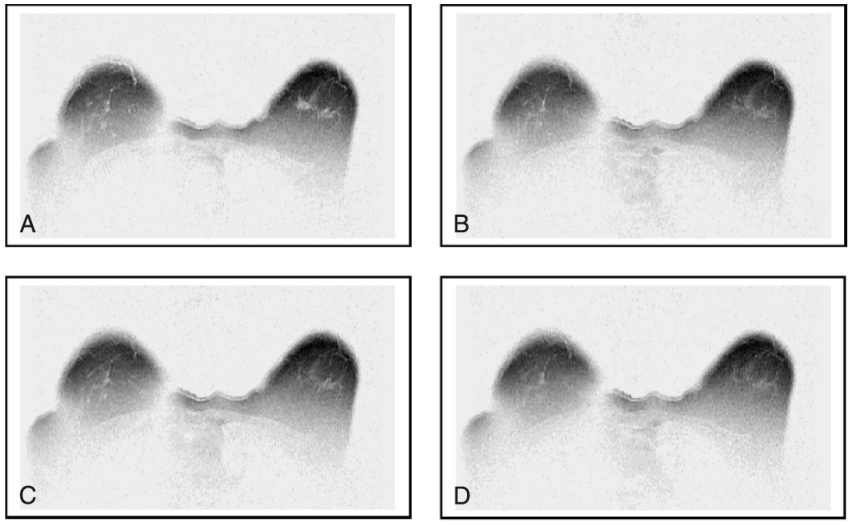


Intensity curve for one voxel

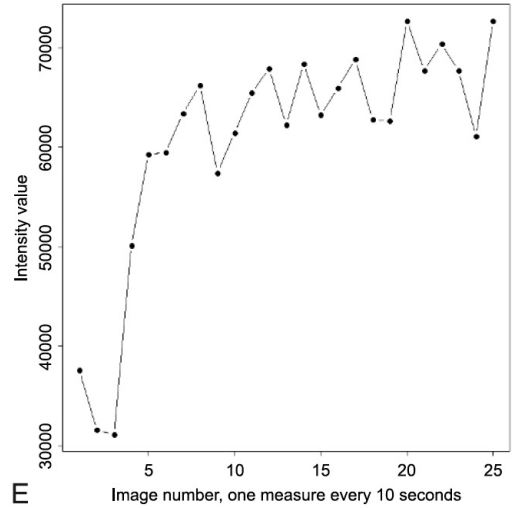


Mclust segmentation with 4 clusters

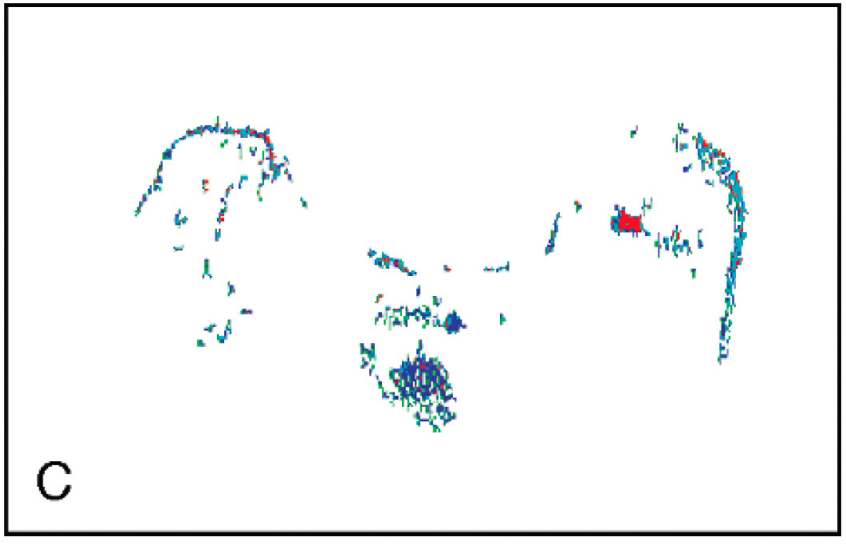
Breast MRI Example Results



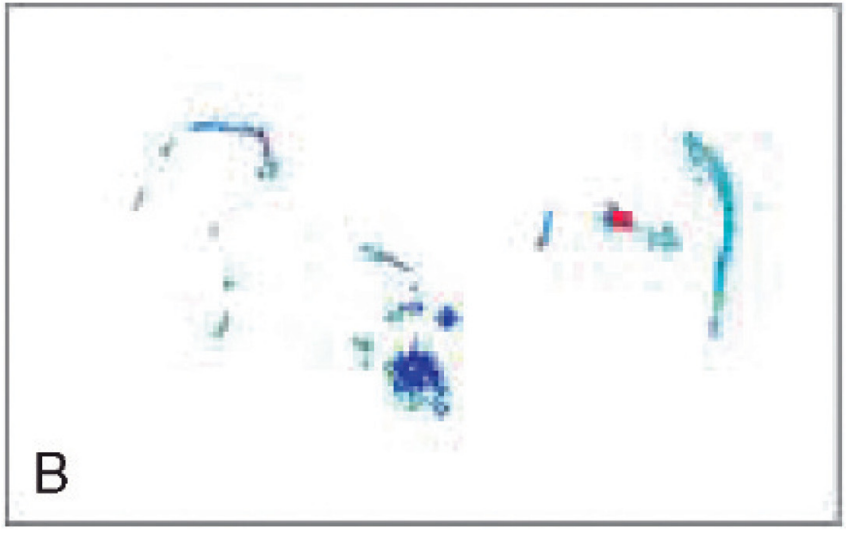
Images at 10, 70, 150, 250 seconds



Intensity curve for one voxel



Mclust segmentation with 4 clusters



Bayesian morphology restoration

Breast MRI Results for 19 patients

Breast MRI Results for 19 patients

TABLE 2. Curve Type Versus Pathology Results for the 19 Patients

Curve Type	Number of Patients	Pathology Results
1 (benign)	6	5 benign, 1 unknown
2 (uncertain)	5	1 unknown, 4 cancer
3 (malignant)	8	8 cancer

Breast MRI Results for 19 patients

TABLE 2. Curve Type Versus Pathology Results for the 19 Patients

Curve Type	Number of Patients	Pathology Results
1 (benign)	6	5 benign, 1 unknown
2 (uncertain)	5	1 unknown, 4 cancer
3 (malignant)	8	8 cancer

Reference: Forbes et al, 2006, *J. Computer Assisted Tomography*, "Finding regions of interest in dynamic breast MRI."

Other Image Processing and Pattern Recogntion Applications

Other Image Processing and Pattern Recogntion Applications

- Papers at:

Other Image Processing and Pattern Recogntion Applications

- Papers at:
  - www.stat.washington.edu/raftery/Research/publications.html

Other Image Processing and Pattern Recogntion Applications

- Papers at:
  - www.stat.washington.edu/raftery/Research/publications.html
  - OR from my home page: $\longrightarrow$ Research $\longrightarrow$ Publications

Other Image Processing and Pattern Recogntion Applications

- Papers at:
  - www.stat.washington.edu/raftery/Research/publications.html
  - OR from my home page: $\longrightarrow$ Research $\longrightarrow$ Publications
- Image segmentation with small features using incremental model-based clustering (Fraley et al, 2005, *J. Comput. Graph. Stat*)

Other Image Processing and Pattern Recogntion Applications

- Papers at:
  - www.stat.washington.edu/raftery/Research/publications.html
  - OR from my home page: $\longrightarrow$ Research $\longrightarrow$ Publications
- Image segmentation with small features using incremental model-based clustering (Fraley et al, 2005, *J. Comput. Graph. Stat*)
- Multi-band image segmentation via model-based cluster trees (Murtagh et al, 2005, *Image & Vision Computing*)

Other Image Processing and Pattern Recogntion Applications

- Papers at:
  - www.stat.washington.edu/raftery/Research/publications.html
  - OR from my home page: $\longrightarrow$ Research $\longrightarrow$ Publications
- Image segmentation with small features using incremental model-based clustering (Fraley et al, 2005, *J. Comput. Graph. Stat*)
- Multi-band image segmentation via model-based cluster trees (Murtagh et al, 2005, *Image & Vision Computing*)
- Segmentation of microarray images with inner holes, artifacts and blank spots (Li et al, 2005, *Bioinformatics*)

Other Image Processing and Pattern Recogntion Applications

- Papers at:
  - www.stat.washington.edu/raftery/Research/publications.html
  - OR from my home page: $\longrightarrow$ Research $\longrightarrow$ Publications
- Image segmentation with small features using incremental model-based clustering (Fraley et al, 2005, *J. Comput. Graph. Stat*)
- Multi-band image segmentation via model-based cluster trees (Murtagh et al, 2005, *Image & Vision Computing*)
- Segmentation of microarray images with inner holes, artifacts and blank spots (Li et al, 2005, *Bioinformatics*)
- Image segmentation with model-based clustering via sampling (Wehrens et al, 2004, *J. Classification*)

Other Image Processing and Pattern Recogntion Applications

- Papers at:
  - www.stat.washington.edu/raftery/Research/publications.html
  - OR from my home page: $\longrightarrow$ Research $\longrightarrow$ Publications
- Image segmentation with small features using incremental model-based clustering (Fraley et al, 2005, *J. Comput. Graph. Stat*)
- Multi-band image segmentation via model-based cluster trees (Murtagh et al, 2005, *Image & Vision Computing*)
- Segmentation of microarray images with inner holes, artifacts and blank spots (Li et al, 2005, *Bioinformatics*)
- Image segmentation with model-based clustering via sampling (Wehrens et al, 2004, *J. Classification*)
- Detecting features in spatial point patterns: minefields, earthquake faults (papers with Byers, Dasgupta, Walsh 1998–)

Variable/Feature Selection for Model-Based Clustering

Variable/Feature Selection for Model-Based Clustering

- Which variables to include in clustering?

Variable/Feature Selection for Model-Based Clustering

- Which variables to include in clustering?
- General approach: Treat it as a model choice problem by vewing each combination of variables as a statistical model

Variable/Feature Selection for Model-Based Clustering

- Which variables to include in clustering?
- General approach: Treat it as a model choice problem by vewing each combination of variables as a statistical model
- Formulate both choices of variables as models for (Y_1, Y_2, Y_3):

Variable/Feature Selection for Model-Based Clustering

- Which variables to include in clustering?
- General approach: Treat it as a model choice problem by vewing each combination of variables as a statistical model
- Formulate both choices of variables as models for (Y_1, Y_2, Y_3):
  - Model for (Y_1, Y_2) choice says that Y_3 is conditionally independent of the cluster assignment variable Z given (Y_1, Y_2)

Variable/Feature Selection for Model-Based Clustering

- Which variables to include in clustering?
- General approach: Treat it as a model choice problem by vewing each combination of variables as a statistical model
- Formulate both choices of variables as models for (Y_1, Y_2, Y_3):
  - Model for (Y_1, Y_2) choice says that Y_3 is conditionally independent of the cluster assignment variable Z given (Y_1, Y_2)
  - Model for (Y_1, Y_2, Y_3) choice says that all 3 variables depend on which cluster the object is in

Variable Selection Method

Variable Selection Method

- We consider whether to include one extra variable as a clustering variable

Variable Selection Method

- We consider whether to include one extra variable as a clustering variable
- Two models: One says that the new variable is useful for clustering given the current clustering variables

Variable Selection Method

- We consider whether to include one extra variable as a clustering variable
- Two models: One says that the new variable is useful for clustering given the current clustering variables
- The other model says that the variable is *not* useful for clustering given the current clustering variables

Variable Selection Method

- We consider whether to include one extra variable as a clustering variable
- Two models: One says that the new variable is useful for clustering given the current clustering variables
- The other model says that the variable is *not* useful for clustering given the current clustering variables
- We partition the data Y into 3 disjoint subsets $Y^{(clust)}$, $Y^{(?)}$ and $Y^{(other)}$ where

Variable Selection Method

- We consider whether to include one extra variable as a clustering variable
- Two models: One says that the new variable is useful for clustering given the current clustering variables
- The other model says that the variable is *not* useful for clustering given the current clustering variables
- We partition the data Y into 3 disjoint subsets $Y^{(clust)}$, $Y^{(?)}$ and $Y^{(other)}$ where
  - $Y^{(clust)}$ is the set of currently selected clustering variables

Variable Selection Method

- We consider whether to include one extra variable as a clustering variable
- Two models: One says that the new variable is useful for clustering given the current clustering variables
- The other model says that the variable is *not* useful for clustering given the current clustering variables
- We partition the data Y into 3 disjoint subsets $Y^{(clust)}$, $Y^{(?)}$ and $Y^{(other)}$ where
  - $Y^{(clust)}$ is the set of currently selected clustering variables
  - $Y^{(?)}$ is the new variable considered for inclusion into $Y^{(clust)}$

Variable Selection Method

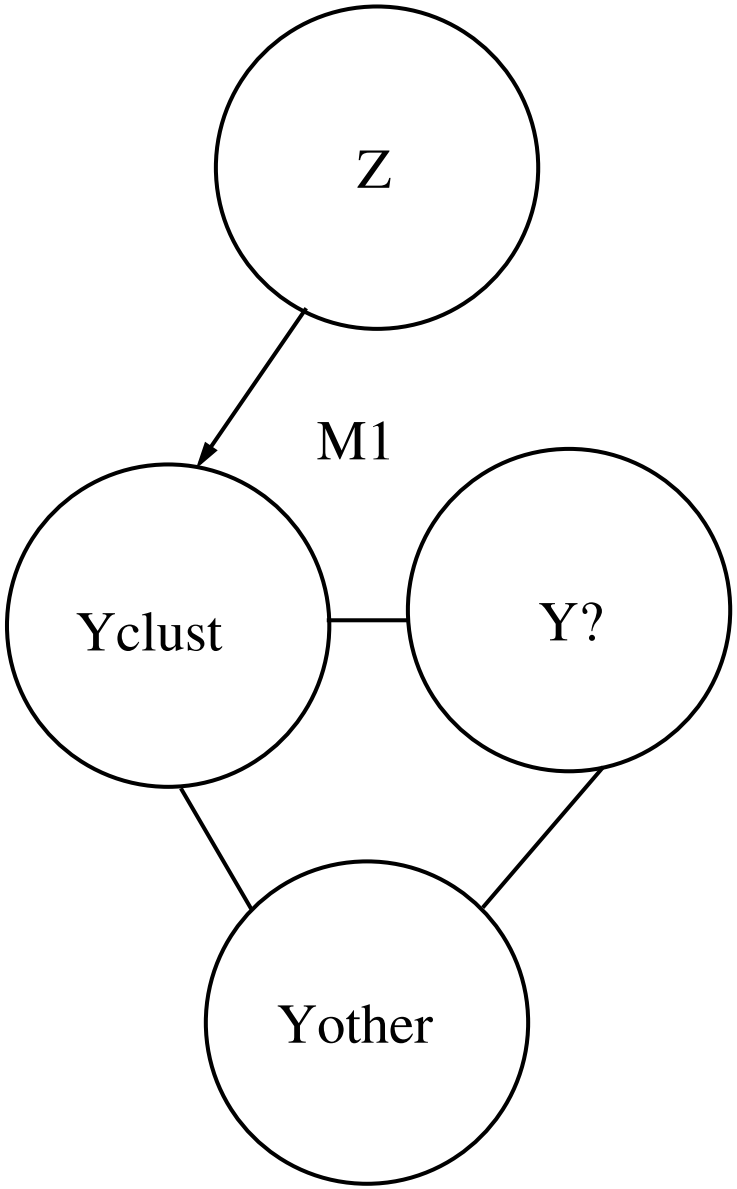
- We consider whether to include one extra variable as a clustering variable
- Two models: One says that the new variable is useful for clustering given the current clustering variables
- The other model says that the variable is *not* useful for clustering given the current clustering variables
- We partition the data Y into 3 disjoint subsets $Y^{(clust)}$, $Y^{(?)}$ and $Y^{(other)}$ where
  - $Y^{(clust)}$ is the set of currently selected clustering variables
  - $Y^{(?)}$ is the new variable considered for inclusion into $Y^{(clust)}$
  - $Y^{(other)}$ is the set of all other variables

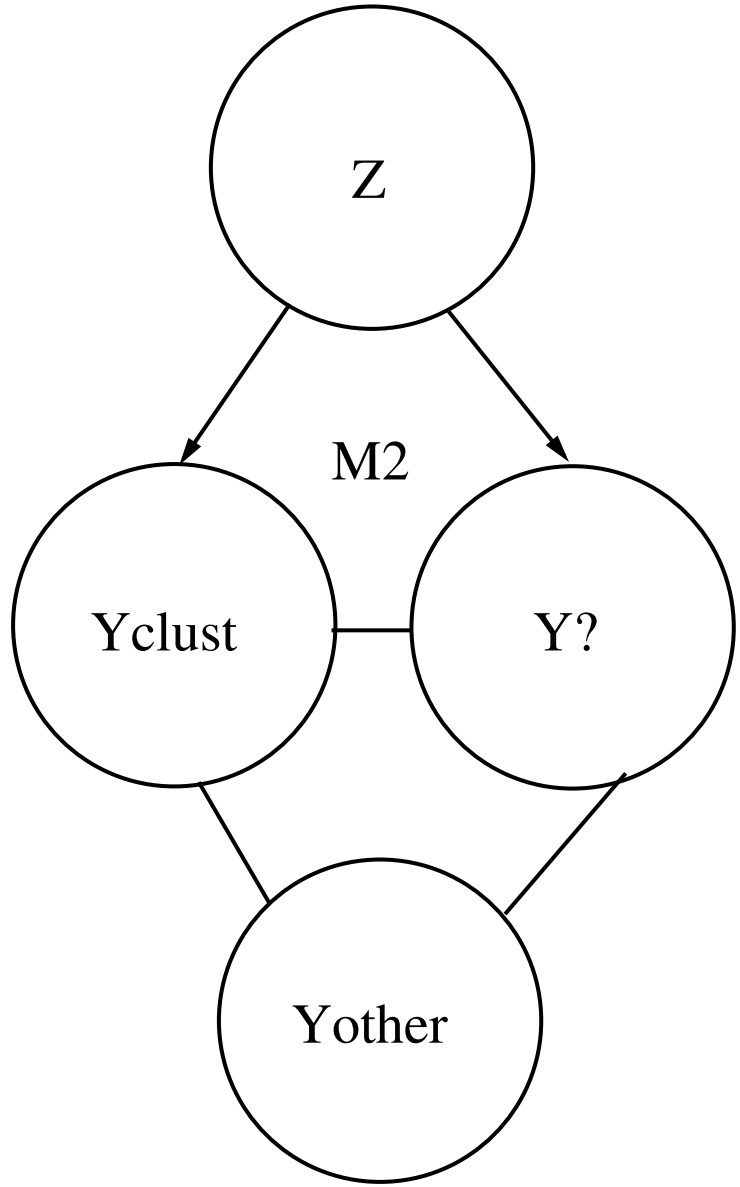
Variable Selection Method

- We consider whether to include one extra variable as a clustering variable
- Two models: One says that the new variable is useful for clustering given the current clustering variables
- The other model says that the variable is *not* useful for clustering given the current clustering variables
- We partition the data Y into 3 disjoint subsets $Y^{(clust)}$, $Y^{(?)}$ and $Y^{(other)}$ where
  - $Y^{(clust)}$ is the set of currently selected clustering variables
  - $Y^{(?)}$ is the new variable considered for inclusion into $Y^{(clust)}$
  - $Y^{(other)}$ is the set of all other variables
- Let Z be the matrix of (unobserved) variables that say which group each observation belongs to (as in EM and MCMC for mixtures)

To Include or Not To Include $Y^{(?)}$? Here are the Models

To Include or Not To Include $Y^{(?)}$? Here are the Models





The Two Models

The Two Models

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

The Two Models

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

The Two Models

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

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

The Two Models

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

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

Implementation of Variable Selection

Implementation of Variable Selection

- If $Y^{(?)}$ is a single variable, then

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

Implementation of Variable Selection

- If $Y^{(?)}$ is a single variable, then

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

- Given the partition and the two models we would like to make a decision based on the Bayes factor B_{21}.

Implementation of Variable Selection

- If $Y^{(?)}$ is a single variable, then

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

- Given the partition and the two models we would like to make a decision based on the Bayes factor B_{21}.
- We use the BIC approximation

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

Implementation of Variable Selection

- If $Y^{(?)}$ is a single variable, then

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

- Given the partition and the two models we would like to make a decision based on the Bayes factor B_{21}.
- We use the BIC approximation

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

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

Implementation of Variable Selection

- If $Y^{(?)}$ is a single variable, then

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

- Given the partition and the two models we would like to make a decision based on the Bayes factor B_{21}.
- We use the BIC approximation

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

- With mild 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.

Search Algorithm

Search Algorithm

- In order to explore all of the model space (create different partitions of the variables to check) we need a search algorithm.

Search Algorithm

- In order to explore all of the model space (create different partitions of the variables to check) we need a search algorithm.
- We iterate between inclusion and exclusion steps:

Search Algorithm

- In order to explore all of the model space (create different partitions of the variables to check) we need a search algorithm.
- We iterate between inclusion and exclusion steps:
  - Inclusion steps test new variables for inclusion into the set of clustering variables

Search Algorithm

- In order to explore all of the model space (create different partitions of the variables to check) we need a search algorithm.
- We iterate between 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

Search Algorithm

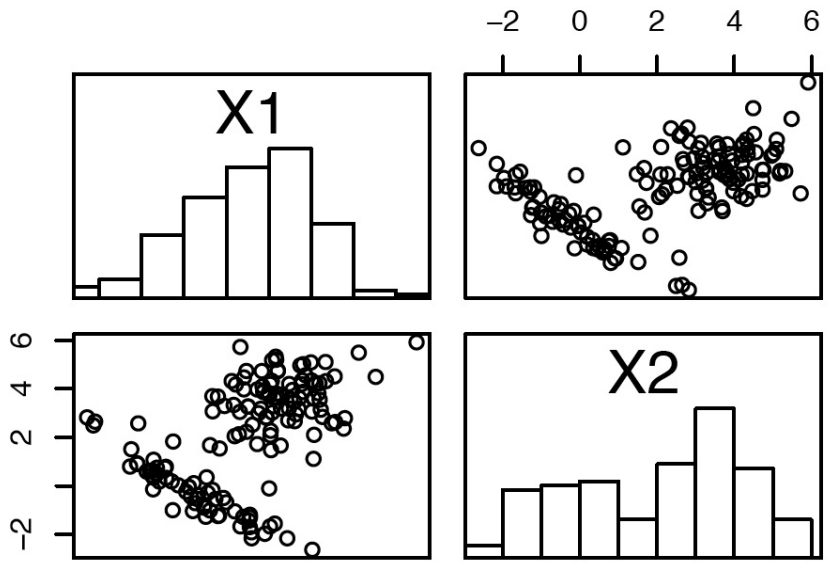
- In order to explore all of the model space (create different partitions of the variables to check) we need a search algorithm.
- We iterate between inclusion and exclusion steps:
  - Inclusion steps test new variables for inclusion into the set of clustering variables
  - Exclusion steps test variables currently in the set of clustering variables for exclusion from that set
  - Inclusion and exclusion decisions are based on the approximate Bayes factors

Search Algorithm

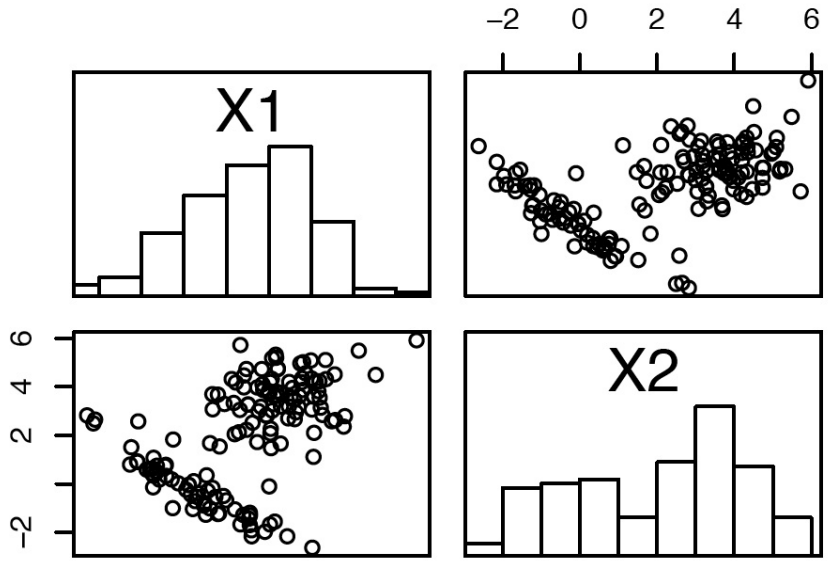
- In order to explore all of the model space (create different partitions of the variables to check) we need a search algorithm.
- We iterate between inclusion and exclusion steps:
  - Inclusion steps test new variables for inclusion into the set of clustering variables
  - Exclusion steps test variables currently in the set of clustering variables for exclusion from that set
  - Inclusion and exclusion decisions are based on the approximate Bayes factors
  - We stop when two proposed changes in a row are rejected

Simulated Data with No Noise Variables

Simulated Data with No Noise Variables

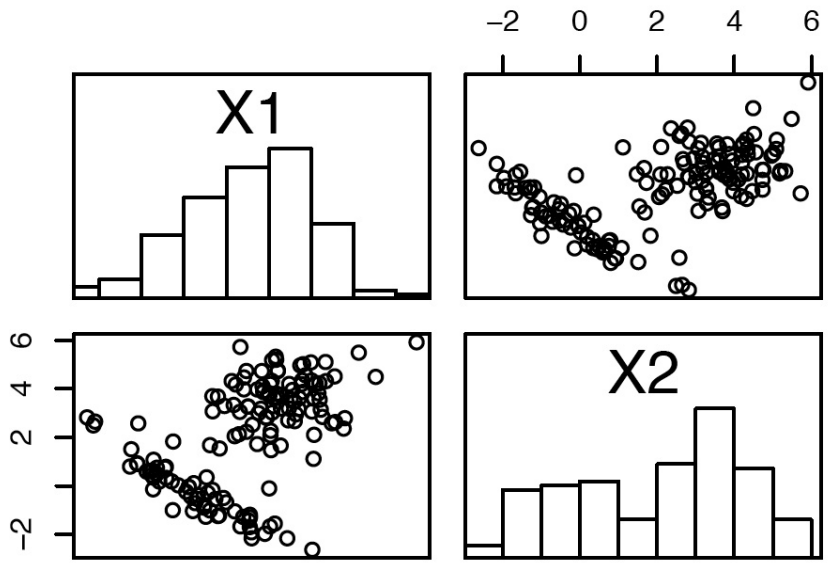


Simulated Data with No Noise Variables



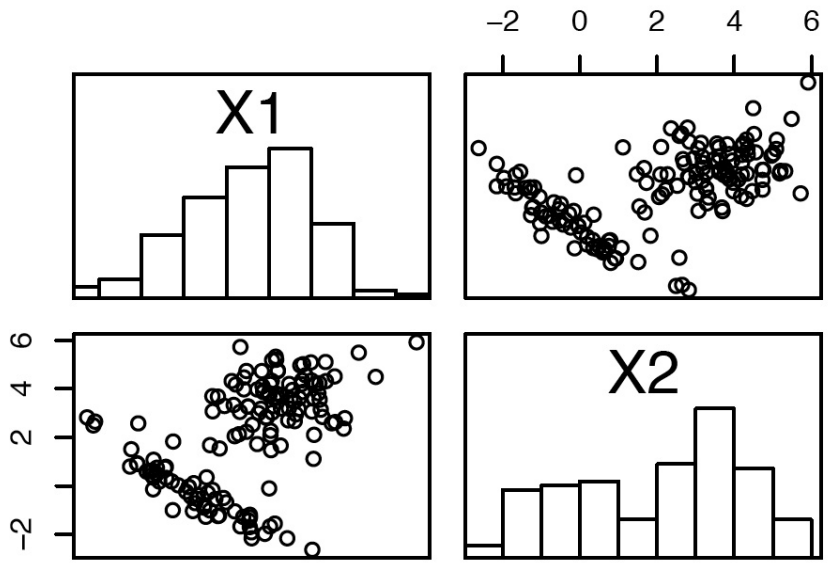
- First we look at an example where there are no noise variables present

Simulated Data with No Noise Variables



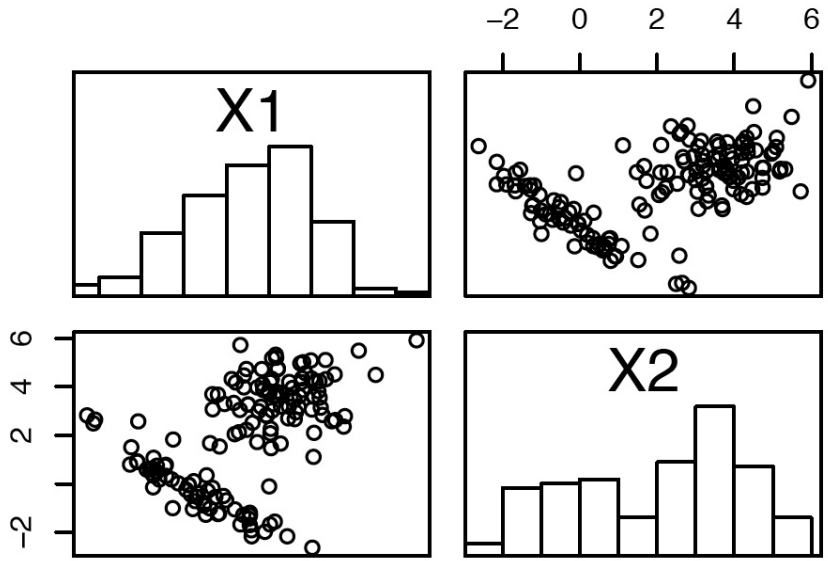
- First we look at an example where there are no noise variables present
- Have two variables with clustering information

Simulated Data with No Noise Variables



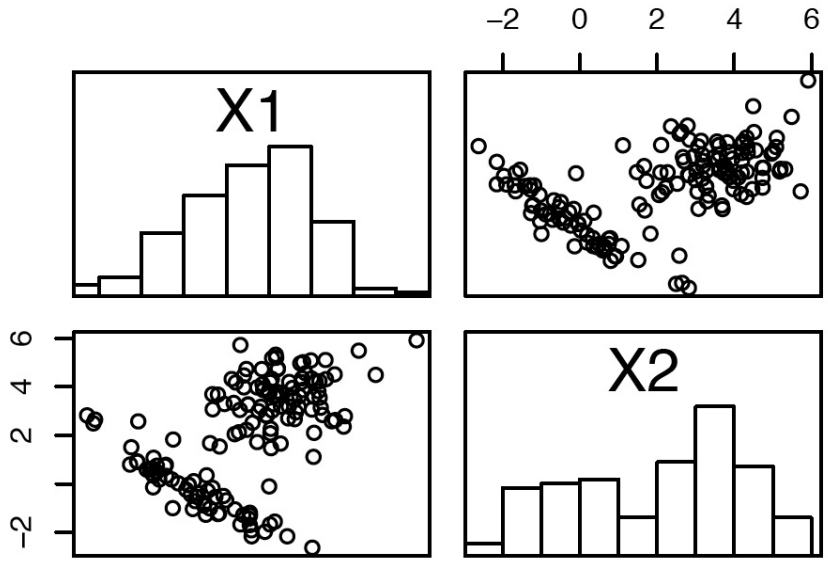
- First we look at an example where there are no noise variables present
- Have two variables with clustering information
- 150 observations

Simulated Data with 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

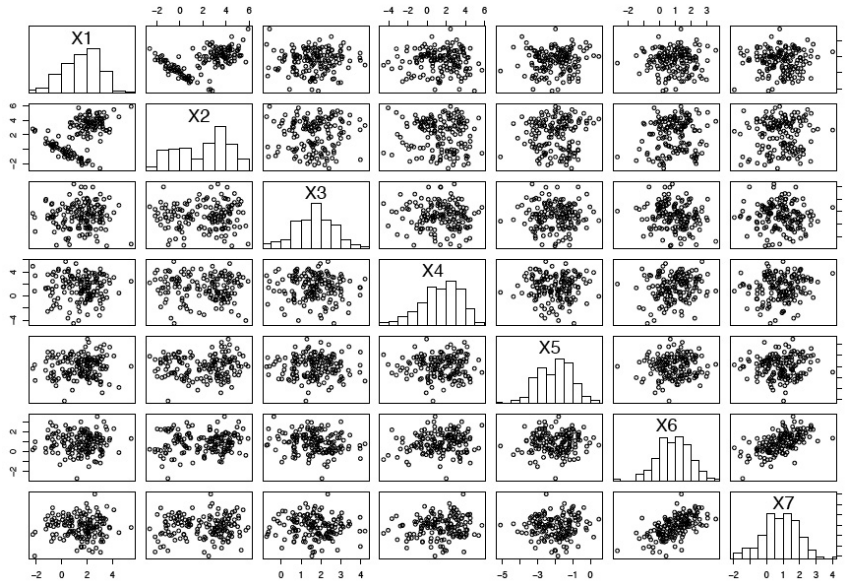
Simulated Data with No Noise Variables



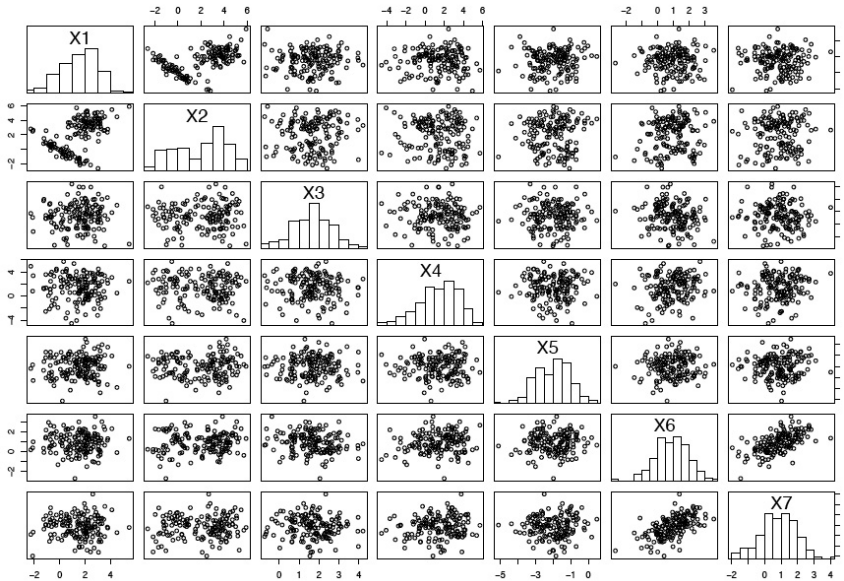
- First we look at an example where there are no noise variables present
- Have two variables with clustering information
- 150 observations
- The clusters are well separated with different variances
- The method correctly selects both variables

Simulated Data with Noise Variables

Simulated Data with Noise Variables

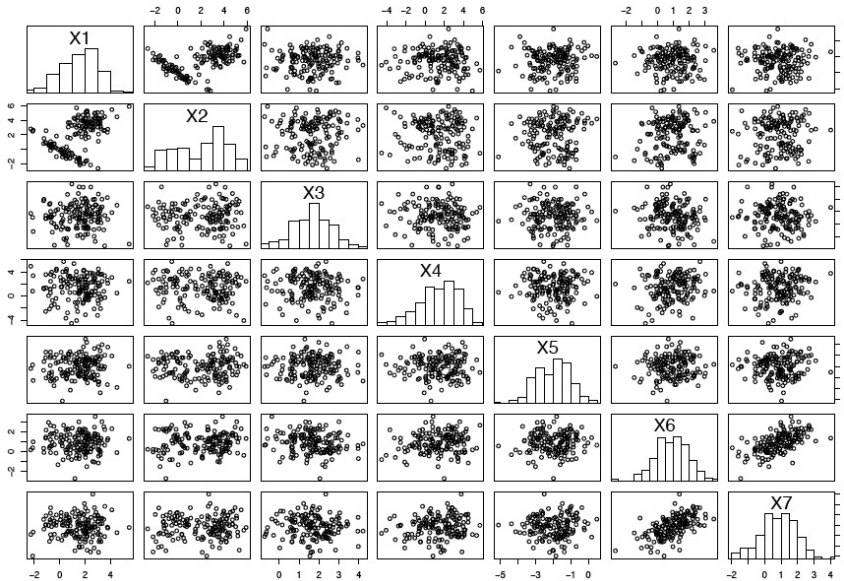


Simulated Data with Noise Variables



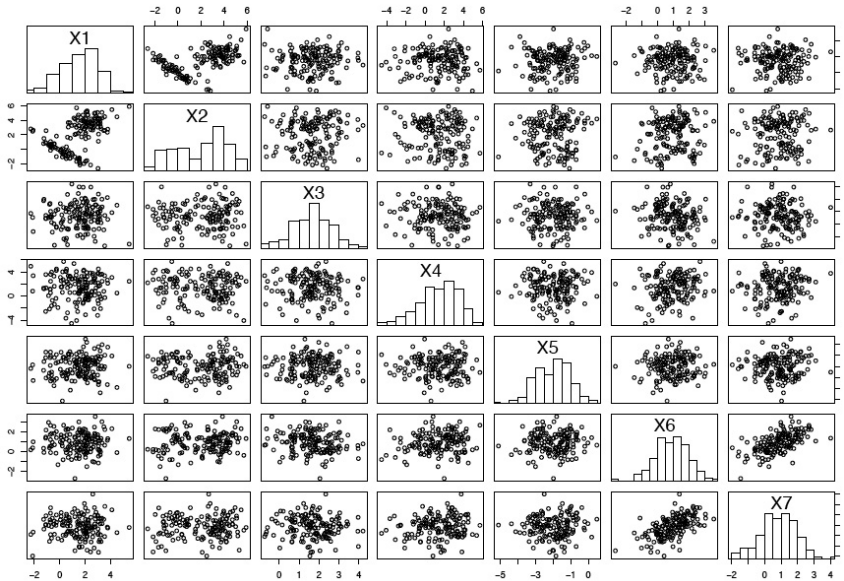
- Same 2 clustering variables as before

Simulated Data with Noise Variables



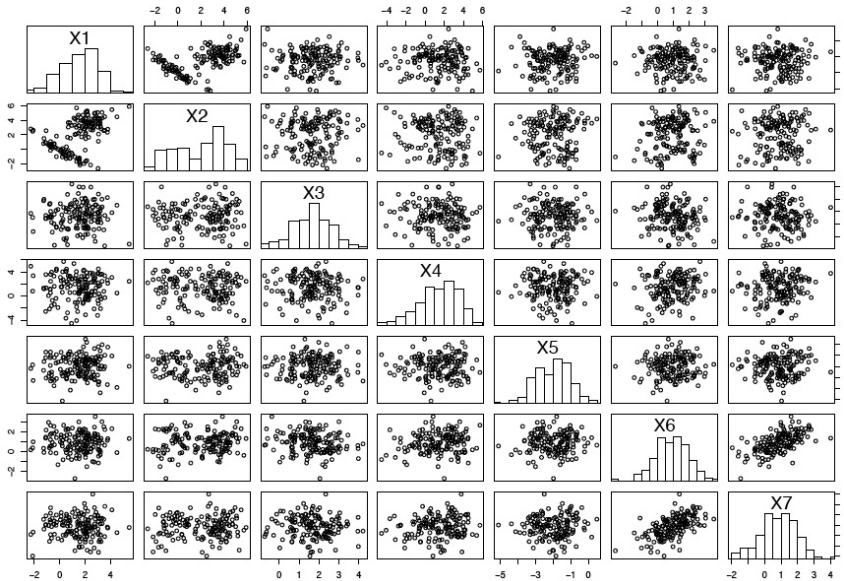
- Same 2 clustering variables as before
- 5 noise variables added:

Simulated Data with Noise Variables



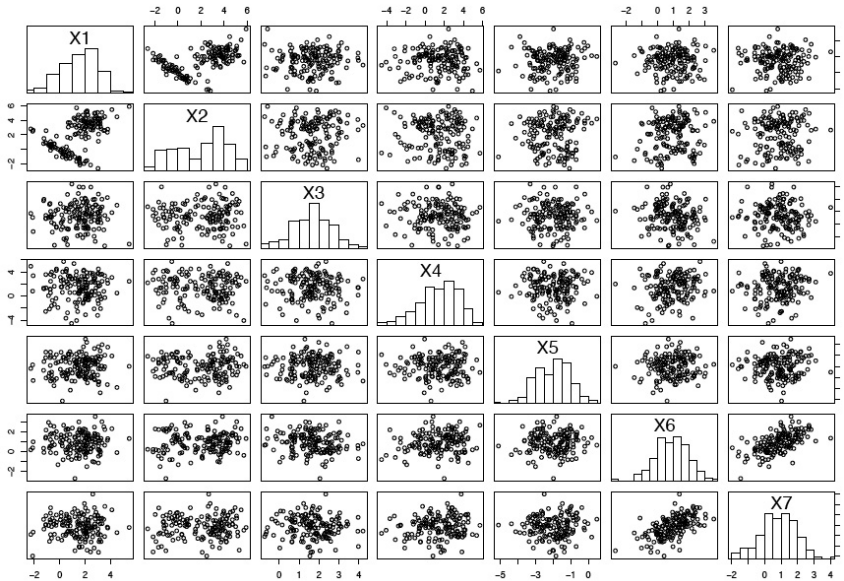
- Same 2 clustering variables as before
- 5 noise variables added:
  - X3, X4 and X5 are independent

Simulated Data with Noise Variables



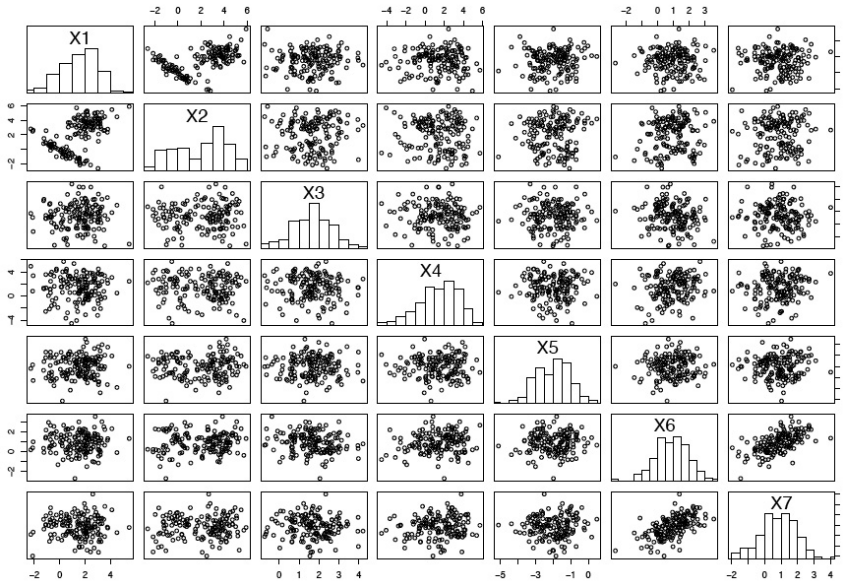
- Same 2 clustering variables as before
- 5 noise variables added:
  - X3, X4 and X5 are independent
  - X6 and X7 are dependent

Simulated Data with Noise Variables



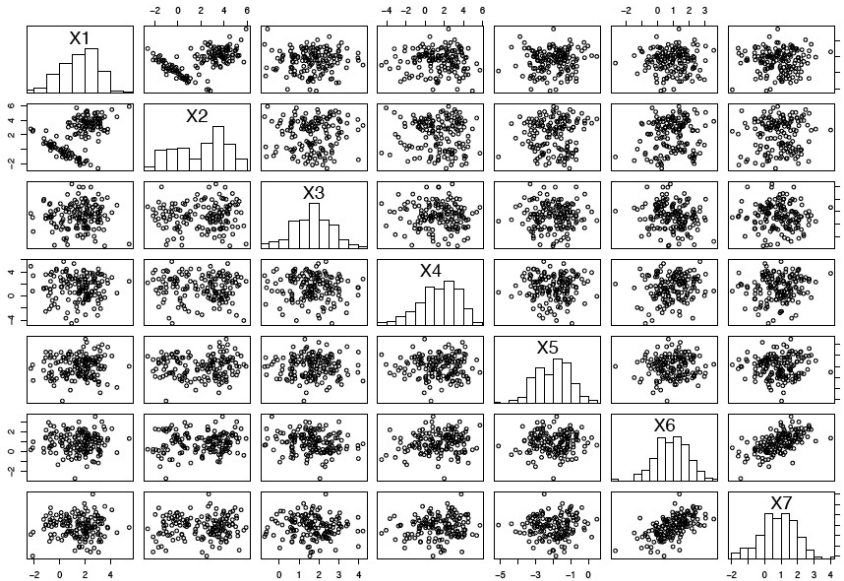
- Same 2 clustering variables as before
- 5 noise variables added:
  - X3, X4 and X5 are independent
  - X6 and X7 are dependent
- Compare clustering results:

Simulated Data with Noise Variables



- Same 2 clustering variables as before
- 5 noise variables added:
  - X3, X4 and X5 are independent
  - X6 and X7 are dependent
- Compare clustering results:

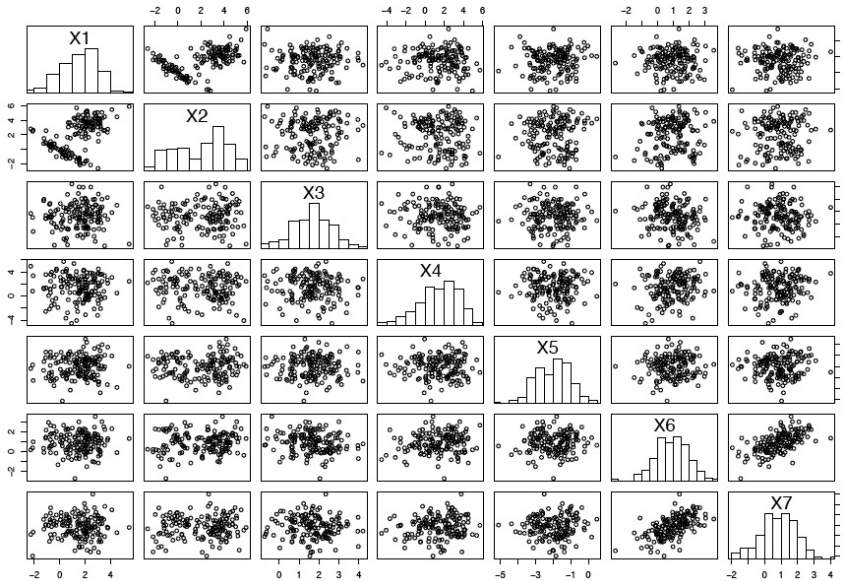
Simulated Data with Noise Variables



- Same 2 clustering variables as before
- 5 noise variables added:
  - X3, X4 and X5 are independent
  - X6 and X7 are dependent
- Compare clustering results:

Variables	# of Groups	Error rate

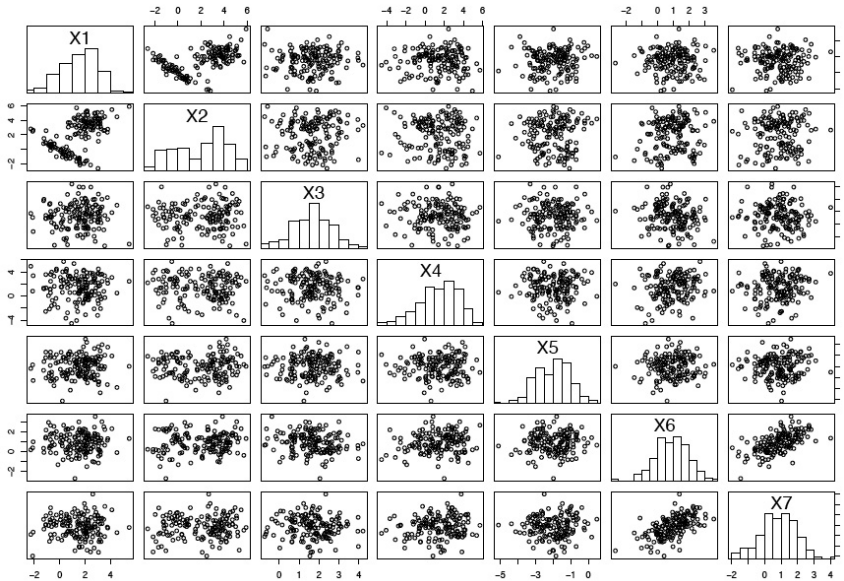
Simulated Data with Noise Variables



- Same 2 clustering variables as before
- 5 noise variables added:
  - X3, X4 and X5 are independent
  - X6 and X7 are dependent
- Compare clustering results:

Variables	# of Groups	Error rate
All 7	5	44.7%

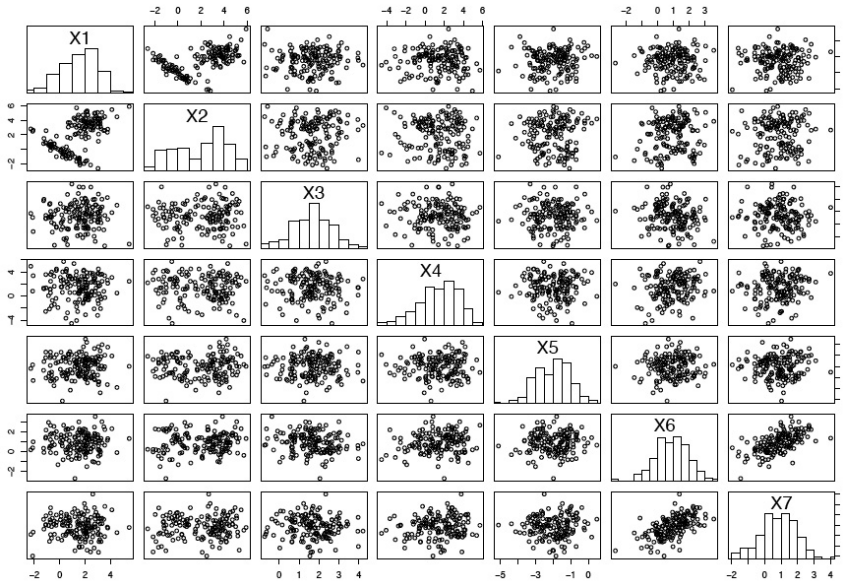
Simulated Data with Noise Variables



- Same 2 clustering variables as before
- 5 noise variables added:
  - X3, X4 and X5 are independent
  - X6 and X7 are dependent
- Compare clustering results:

Variables	# of Groups	Error rate
All 7	5	44.7%
All 7	2 (constrained)	3.3%

Simulated Data with Noise Variables



- Same 2 clustering variables as before
- 5 noise variables added:
  - X3, X4 and X5 are independent
  - X6 and X7 are dependent
- Compare clustering results:

Variables	# of Groups	Error rate
All 7	5	44.7%
All 7	2 (constrained)	3.3%
Selected 2	2	0%

Crabs Data

Crabs Data

- 4 groups: male orange, female orange, male blue and female blue

Crabs Data

- 4 groups: male orange, female orange, male blue and female blue
- 200 observations (50 per group)

Crabs Data

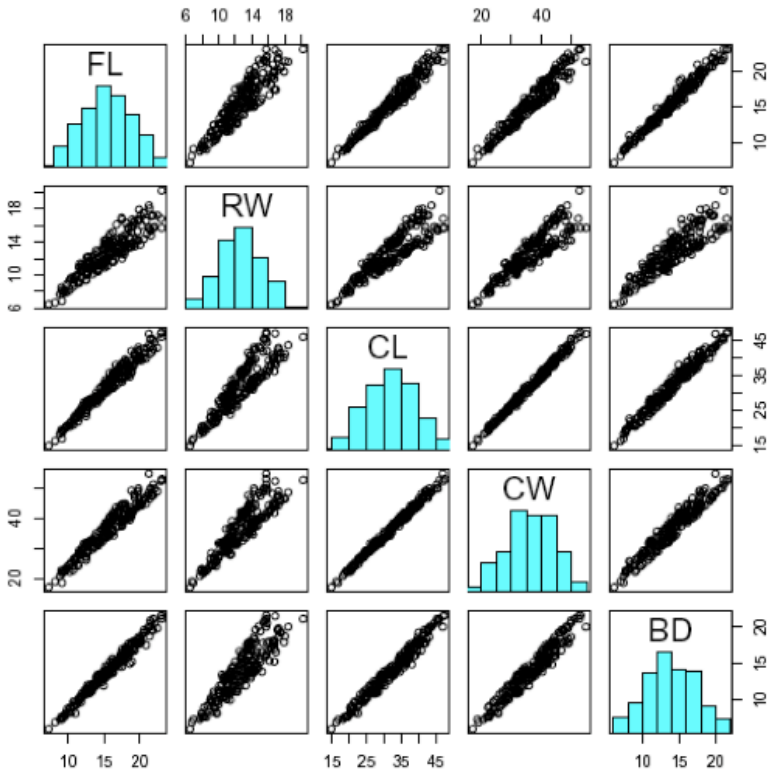
- 4 groups: male orange, female orange, male blue and female blue
- 200 observations (50 per group)
- 5 variables measuring size

Crabs Data

- 4 groups: male orange, female orange, male blue and female blue
- 200 observations (50 per group)
- 5 variables measuring size

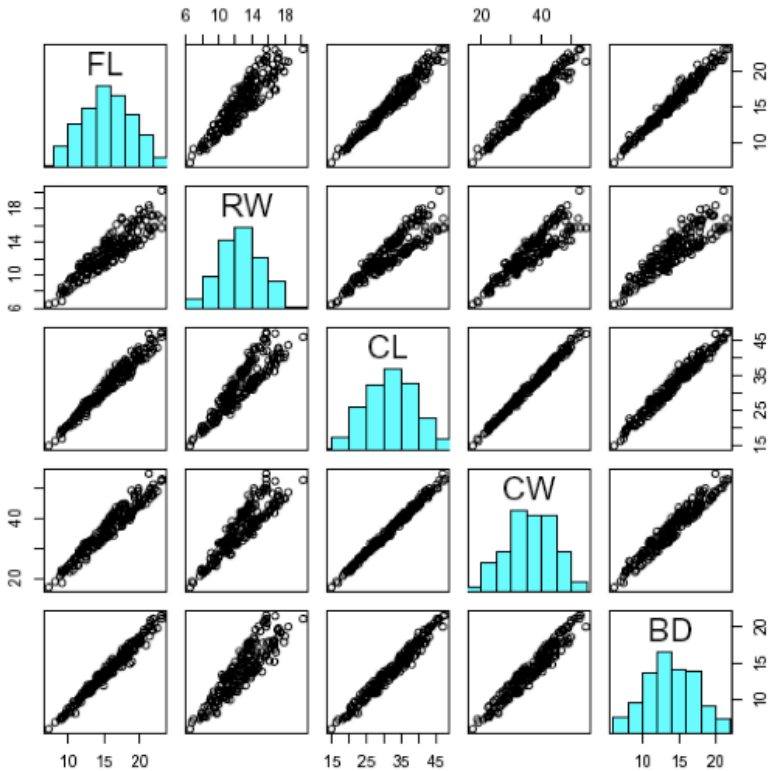
Crabs Data

- 4 groups: male orange, female orange, male blue and female blue
- 200 observations (50 per group)
- 5 variables measuring size



Crabs Data

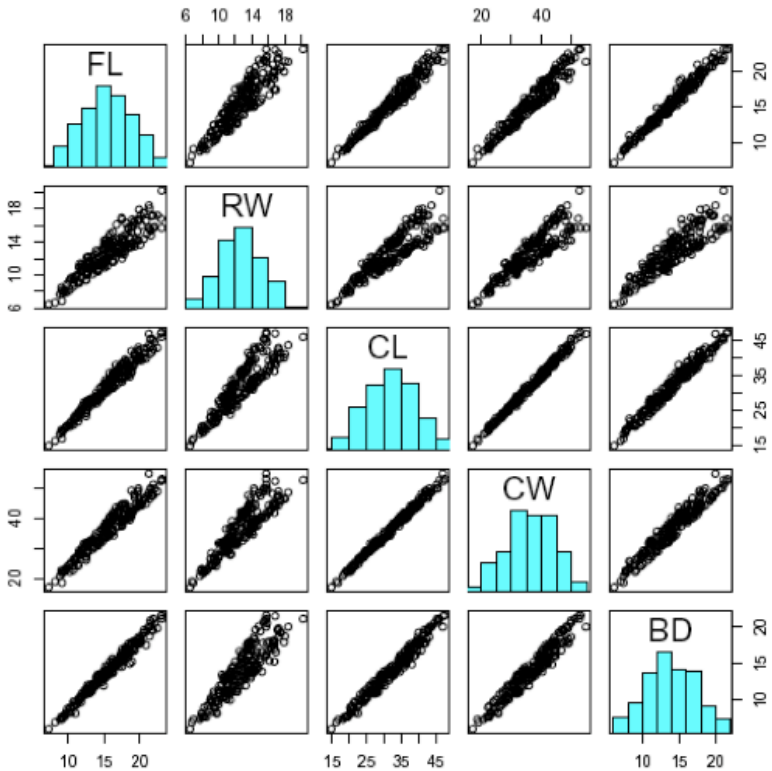
- 4 groups: male orange, female orange, male blue and female blue
- 200 observations (50 per group)
- 5 variables measuring size



- Variables selected: 4 of the 5 variables were selected, all except length along mid-line of carapace (CL)

Crabs Data

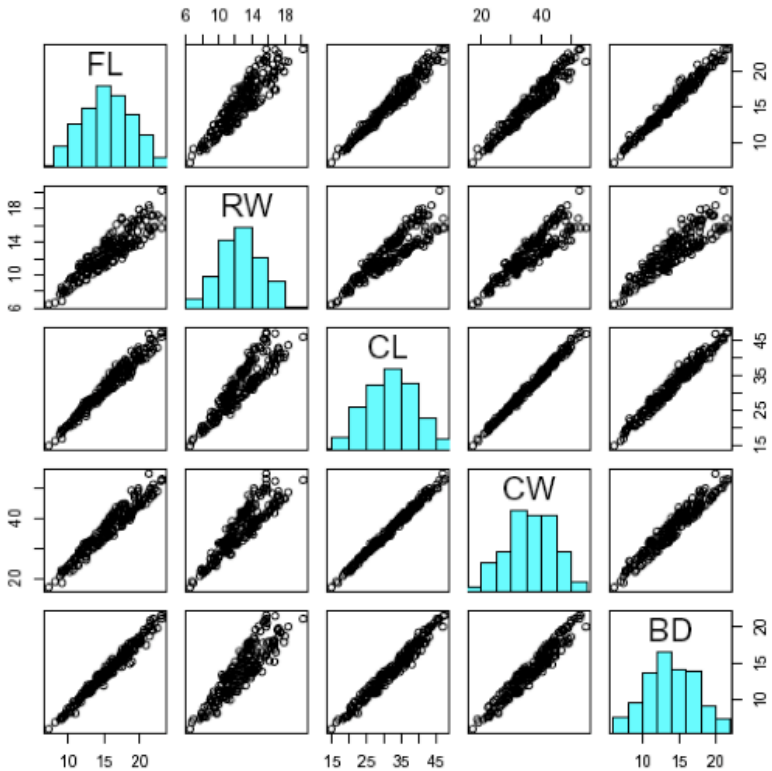
- 4 groups: male orange, female orange, male blue and female blue
- 200 observations (50 per group)
- 5 variables measuring size



- Variables selected: 4 of the 5 variables were selected, all except length along mid-line of carapace (CL)
- Compare clustering results:

Crabs Data

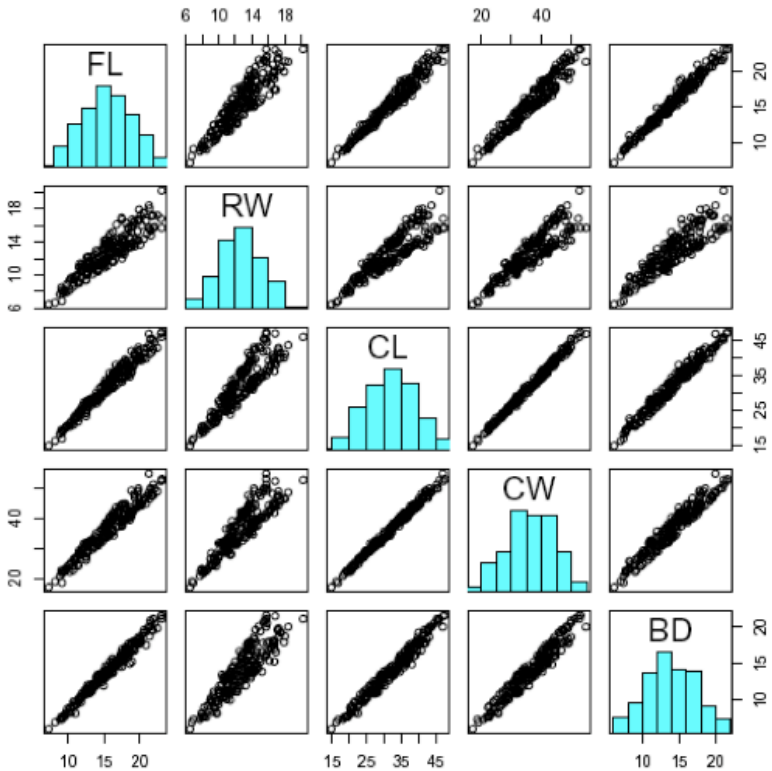
- 4 groups: male orange, female orange, male blue and female blue
- 200 observations (50 per group)
- 5 variables measuring size



- Variables selected: 4 of the 5 variables were selected, all except length along mid-line of carapace (CL)
- Compare clustering results:

Crabs Data

- 4 groups: male orange, female orange, male blue and female blue
- 200 observations (50 per group)
- 5 variables measuring size

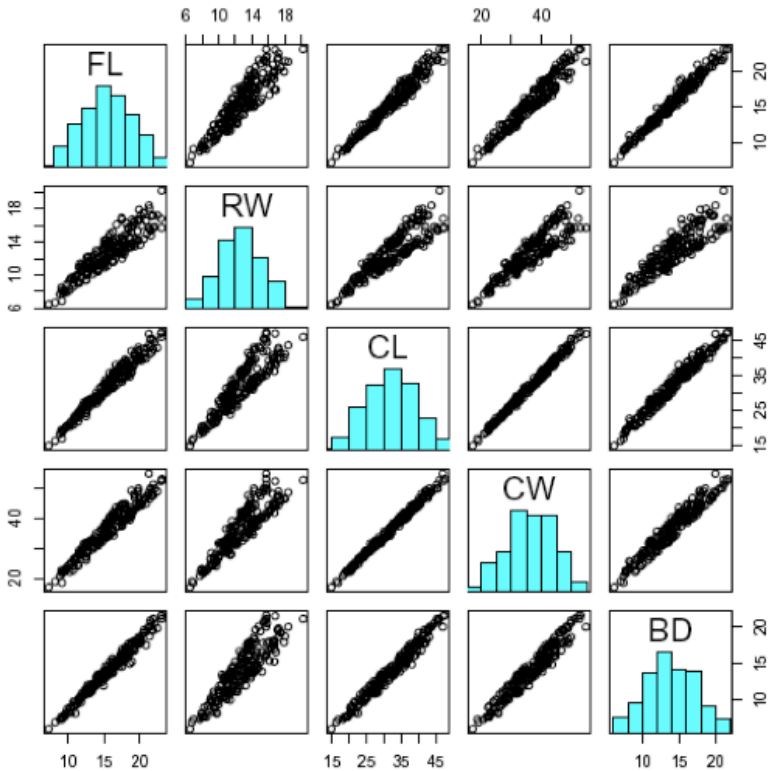


- Variables selected: 4 of the 5 variables were selected, all except length along mid-line of carapace (CL)
- Compare clustering results:

Variables	# of Groups	Error rate

Crabs Data

- 4 groups: male orange, female orange, male blue and female blue
- 200 observations (50 per group)
- 5 variables measuring size

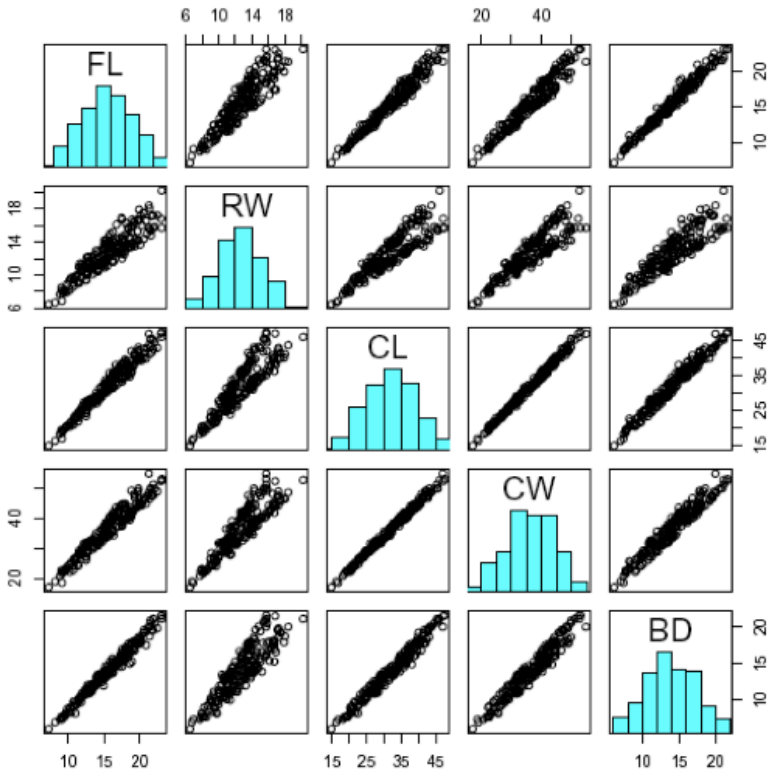


- Variables selected: 4 of the 5 variables were selected, all except length along mid-line of carapace (CL)
- Compare clustering results:

Variables	# of Groups	Error rate
All 5	7	42.5%

Crabs Data

- 4 groups: male orange, female orange, male blue and female blue
- 200 observations (50 per group)
- 5 variables measuring size

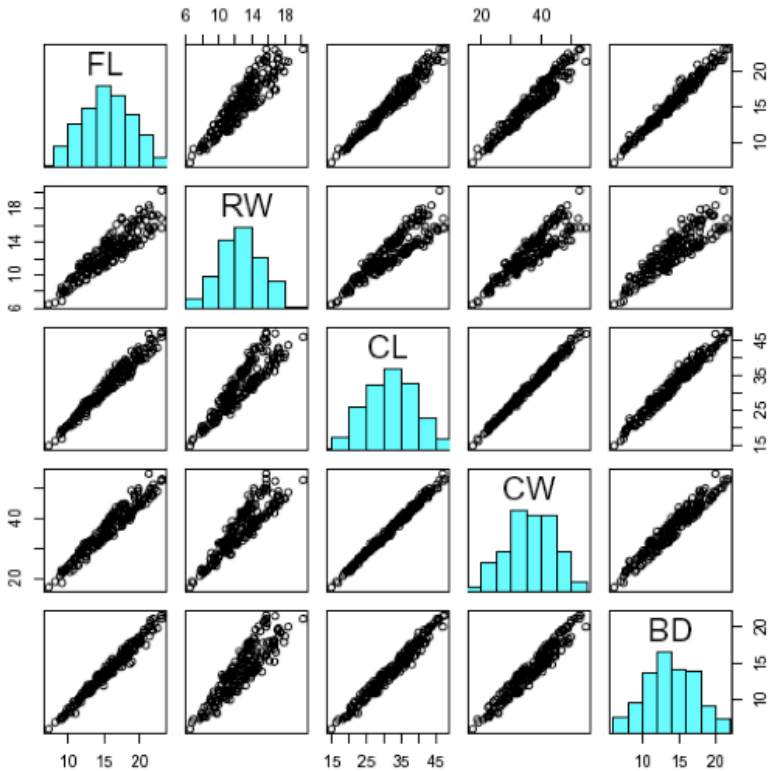


- Variables selected: 4 of the 5 variables were selected, all except length along mid-line of carapace (CL)
- Compare clustering results:

Variables	# of Groups	Error rate
All 5	7	42.5%
All 5	4 (constrained)	7.5%

Crabs Data

- 4 groups: male orange, female orange, male blue and female blue
- 200 observations (50 per group)
- 5 variables measuring size



- Variables selected: 4 of the 5 variables were selected, all except length along mid-line of carapace (CL)
- Compare clustering results:

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

Summary

Summary

- Model-based clustering puts cluster analysis on a solid statistical basis and answers questions such as:

Summary

- Model-based clustering puts cluster analysis on a solid statistical basis and answers questions such as:
  - How many groups?

Summary

- Model-based clustering puts cluster analysis on a solid statistical basis and answers questions such as:
  - How many groups?
  - Which clustering method to use?

Summary

- Model-based clustering puts cluster analysis on a solid statistical basis and answers questions such as:
  - How many groups?
  - Which clustering method to use?
  - How certain can we be about the clustering?

Summary

- Model-based clustering puts cluster analysis on a solid statistical basis and answers questions such as:
  - How many groups?
  - Which clustering method to use?
  - How certain can we be about the clustering?
  - How to deal with outliers?

Summary

- Model-based clustering puts cluster analysis on a solid statistical basis and answers questions such as:
  - How many groups?
  - Which clustering method to use?
  - How certain can we be about the clustering?
  - How to deal with outliers?
- Successfully applied to several image segmentation problems

Summary

- Model-based clustering puts cluster analysis on a solid statistical basis and answers questions such as:
  - How many groups?
  - Which clustering method to use?
  - How certain can we be about the clustering?
  - How to deal with outliers?
- Successfully applied to several image segmentation problems
- A statistically based method proposed for variable selection/feature selection in model-based clustering

Summary

- Model-based clustering puts cluster analysis on a solid statistical basis and answers questions such as:
  - How many groups?
  - Which clustering method to use?
  - How certain can we be about the clustering?
  - How to deal with outliers?
- Successfully applied to several image segmentation problems
- A statistically based method proposed for variable selection/feature selection in model-based clustering
- Software: R packages available at `http://cran.r-project.org`:

Summary

- Model-based clustering puts cluster analysis on a solid statistical basis and answers questions such as:
  - How many groups?
  - Which clustering method to use?
  - How certain can we be about the clustering?
  - How to deal with outliers?
- Successfully applied to several image segmentation problems
- A statistically based method proposed for variable selection/feature selection in model-based clustering
- Software: R packages available at `http://cran.r-project.org`:
  - Model-based clustering: `mclust`

Summary

- Model-based clustering puts cluster analysis on a solid statistical basis and answers questions such as:
  - How many groups?
  - Which clustering method to use?
  - How certain can we be about the clustering?
  - How to deal with outliers?
- Successfully applied to several image segmentation problems
- A statistically based method proposed for variable selection/feature selection in model-based clustering
- Software: R packages available at `http://cran.r-project.org`:
  - Model-based clustering: `mclust`
  - Variable selection: `clustvarsel`

Summary

- Model-based clustering puts cluster analysis on a solid statistical basis and answers questions such as:
  - How many groups?
  - Which clustering method to use?
  - How certain can we be about the clustering?
  - How to deal with outliers?
- Successfully applied to several image segmentation problems
- A statistically based method proposed for variable selection/feature selection in model-based clustering
- Software: R packages available at `http://cran.r-project.org`:
  - Model-based clustering: `mclust`
  - Variable selection: `clustvarsel`
- References:

Summary

- Model-based clustering puts cluster analysis on a solid statistical basis and answers questions such as:
  - How many groups?
  - Which clustering method to use?
  - How certain can we be about the clustering?
  - How to deal with outliers?
- Successfully applied to several image segmentation problems
- A statistically based method proposed for variable selection/feature selection in model-based clustering
- Software: R packages available at `http://cran.r-project.org`:
  - Model-based clustering: `mclust`
  - Variable selection: `clustvarsel`
- References:
  - Model-based clustering:

Summary

- Model-based clustering puts cluster analysis on a solid statistical basis and answers questions such as:
  - How many groups?
  - Which clustering method to use?
  - How certain can we be about the clustering?
  - How to deal with outliers?
- Successfully applied to several image segmentation problems
- A statistically based method proposed for variable selection/feature selection in model-based clustering
- Software: R packages available at `http://cran.r-project.org`:
  - Model-based clustering: `mclust`
  - Variable selection: `clustvarsel`
- References:
  - Model-based clustering:
    - Banfield and Raftery (1993, *Biometrics*)

Summary

- Model-based clustering puts cluster analysis on a solid statistical basis and answers questions such as:
  - How many groups?
  - Which clustering method to use?
  - How certain can we be about the clustering?
  - How to deal with outliers?
- Successfully applied to several image segmentation problems
- A statistically based method proposed for variable selection/feature selection in model-based clustering
- Software: R packages available at `http://cran.r-project.org`:
  - Model-based clustering: `mclust`
  - Variable selection: `clustvarsel`
- References:
  - Model-based clustering:
    - Banfield and Raftery (1993, *Biometrics*)
    - Fraley and Raftery (2002, *J. Amer. Statist. Ass.*)

Summary

- Model-based clustering puts cluster analysis on a solid statistical basis and answers questions such as:
  - How many groups?
  - Which clustering method to use?
  - How certain can we be about the clustering?
  - How to deal with outliers?
- Successfully applied to several image segmentation problems
- A statistically based method proposed for variable selection/feature selection in model-based clustering
- Software: R packages available at `http://cran.r-project.org`:
  - Model-based clustering: `mclust`
  - Variable selection: `clustvarsel`
- References:
  - Model-based clustering:
    - Banfield and Raftery (1993, *Biometrics*)
    - Fraley and Raftery (2002, *J. Amer. Statist. Ass.*)
  - Variable selection: Raftery and Dean (2006, *J. Amer. Statist. Ass.*)

Summary

- Model-based clustering puts cluster analysis on a solid statistical basis and answers questions such as:
  - How many groups?
  - Which clustering method to use?
  - How certain can we be about the clustering?
  - How to deal with outliers?
- Successfully applied to several image segmentation problems
- A statistically based method proposed for variable selection/feature selection in model-based clustering
- Software: R packages available at `http://cran.r-project.org`:
  - Model-based clustering: `mclust`
  - Variable selection: `clustvarsel`
- References:
  - Model-based clustering:
    - Banfield and Raftery (1993, *Biometrics*)
    - Fraley and Raftery (2002, *J. Amer. Statist. Ass.*)
  - Variable selection: Raftery and Dean (2006, *J. Amer. Statist. Ass.*)
- Website: `www.stat.washington.edu/raftery`
  $\longrightarrow$ Research $\longrightarrow$ Model-based clustering