Segmentation II

Model-based clustering

Morten Berg Jensen

Department of Economics and Business Economics

April 11, 2024

Outline

Outcome

This lecture will help you to understand

- The advantages of basing the clustering process on a statistical model
- The general idea behind model-based clustering
- Estimation and model selection

Key lessons from hierarchical and non-hierarchical cluster analysis

- There is substantial ambiguity associated with
  - The number of clusters
  - The measurement of proximity of individuals
  - The measurement of proximity of groups
- There is no single index to compare different cluster solutions
- Existing validation techniques depend on data and/or the cluster algorithm

Background: Hierarchical and non-hierarchical cluster analysis

- Both of the classic methods for doing cluster analysis are based on combinatorial methods using heuristic procedures
  - No assumptions about the class structure are made regarding the population
  - Choice of clustering method and proximity measure is based on posterior criteria like interpretability of the results

Model-based clustering

- Model-based clustering assumes that the population is made up of several distinct subsets/clusters, each governed by a different multivariate probability density function
- The parameters associated with the model can be used to assign each observation a posterior probability of belonging to a cluster
- The problems of identifying the number of clusters and selecting the clustering method boil down to a model selection problem – for which we have a number of procedures

Model-based clustering (cont'd)

- Furthermore, being based on a genuine statistical model, model-based clustering readily accommodates missing data in a way similar to the state-of-the-art ML methods
- Are also known as latent-class cluster analysis or finite mixture modeling
- K-means clustering are approximate estimation methods for certain finite mixture probability models lending credibility to these
- They allow for an integral representation of the cluster model together with predictor variables such as demographics

The formal model

- Observed data, $\mathbf{x}$ are assumed to originate from a mixture of probability density functions like this

$$f(\mathbf{x}; \mathbf{p}, \boldsymbol{\theta}) = \sum_{j=1}^{c} p_j g_j(\mathbf{x}; \boldsymbol{\theta}_j)$$

  where

  - c is the number of components/clusters
  - $\mathbf{x}$ is a p-dimensional random variable
  - $\mathbf{p}' = (p_1, p_2, \ldots, p_c)$ are mixing proportions, $\sum_{j=1}^{c} p_j = 1$ and $p_j \geq 0$
  - g_j are component densities
  - $\boldsymbol{\theta}' = (\boldsymbol{\theta}_1', \boldsymbol{\theta}_2', \ldots, \boldsymbol{\theta}_c')$ are the parameters governing each component density

The formal model (cont'd)

- Once the parameters of the model have been estimated each observation can be assigned to a cluster using estimated posterior probabilities

$$\begin{aligned}\hat{\mathbf{P}}(\text{case i belongs to cluster } j|\mathbf{x}_i) &= \hat{\mathbf{P}}(j|\mathbf{x}_i) \\ &= \frac{\hat{p}_j g_j(\mathbf{x_i}, \hat{\boldsymbol{\theta}})}{f(\mathbf{x}_i; \hat{\boldsymbol{p}}, \hat{\boldsymbol{\theta}})}, j = 1, \ldots, c \qquad (1)\end{aligned}$$

- Each observation is assigned to the cluster with the maximum estimated posterior probability

The formal model (cont'd)

- One possible solution for a specification of the component densities is to assume that they are Gaussian with different mean vectors, $\boldsymbol{\mu}_j$ and potentially different covariance matrices, $\boldsymbol{\Sigma}_j$

$$g_j(\mathbf{x};\boldsymbol{\theta}_j) = \phi(\mathbf{x}_j;\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j) = \frac{\exp\{-0.5(\mathbf{x}_j-\boldsymbol{\mu}_j)'\boldsymbol{\Sigma}_{\boldsymbol{j}}^{\boldsymbol{-1}}(\mathbf{x}_j-\boldsymbol{\mu}_j)\}}{\sqrt{|2\pi\boldsymbol{\Sigma}_j|}}$$

- But depending on the context other specifications of the component densities may be used to accommodate the specifics of each situation – skewed data, count data, multinomial data ...
- In fact these model may comprise sums of different component densities

Estimation

- Conceptually maximum likelihood estimation is straightforward
- However, from a practical point of view maximization is less simple
- In the situation of Gaussian component densities, maximization is often carried out using the iterated expectation-maximization algorithm

Estimation (cont'd)

- Given a sample of observations $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ and initial parameter estimates, $\boldsymbol{\theta}^0$ and $\mathbf{p}^0$ iterate the following
    1. Use (1) in order to calculate $\hat{\mathbf{P}}^1(j|\mathbf{x}_i), j = 1, \ldots, c$
    2. Update the parameters via

$$\hat{p}_j^1 = \frac{1}{n}\sum_{i=1}^{n}\hat{\mathbf{P}}(j|\mathbf{x}_i)$$

$$\hat{\boldsymbol{\mu}}_j^1 = \frac{1}{np_j}\sum_{i=1}^{c}\mathbf{x}_i\hat{\mathbf{P}}(j|\mathbf{x}_i)$$

$$\hat{\boldsymbol{\Sigma}}_j^1 = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_j)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_j)'\hat{\mathbf{P}}(j|\mathbf{x}_i)$$

until convergence

Estimation (cont'd)

- Thus, based on initial parameter estimates the posterior probabilities can be calculated and parameter estimates can be updated
- Initial parameter estimates may be obtained from solutions to standard heuristic hierarchical methods
- There are a number of common problems known to exist with the use of maximum likelihood estimation
  - Multiple maxima generally exist – ideally the EM algorithm must be run several times based on different initial values
  - Singularities – these are points where the likelihood function becomes infinite
  - The occurrence of singularities is associated with large parameter to observations ratio – thus to mitigate this problem restrictions are often added to the covariance matrices
- Alternatively, Bayesian analysis can be used

Estimation (cont'd)

- Calculation of standard errors of the parameter estimates may proceed in the usual way – via calculation of the Hessian
- Alternatively, bootstrapping may be used – but as for comparisons between solutions to standard heuristic hierarchical methods the label switching problem may be an issue

Estimation of multivariate normal mixture densities

- The MCLUST family is a class of models involving various restrictions attached to the covariance matrix
- The development builds on the following reparameterization of the covariance matrix

$$\mathbf{\Sigma}_j = \lambda_j \mathbf{D}_j \mathbf{\Lambda}_j \mathbf{D}_j'$$

  where
  - $\mathbf{D}_j$ is the matrix of eigenvectors – determining the orientation
  - λ_j is the largest eigenvalue of $\mathbf{\Sigma}_j$ – determining the volume
  - $\mathbf{\Lambda}_j$ is a matrix holding the eigenvalue ratios – determining the shape
- Restrictions regarding variation across clusters with one or several of these features are possible

Determining the number of clusters

- This is a key decision in connection with cluster analysis
- Deciding on the number of clusters can done using one of four possible model selection procedures
  - LRT
  - Information criteria
  - Bayes Factors
  - MCMC methods
- We will focus on the first two possibilities

Determining the number of clusters (cont'd)

- Since a comparison among models with c_0 against a model with c_{0+1} clusters involves nested models the usual likelihood ratio statistic could be an option
- However, since model the model with c_0 clusters is obtained by restricting one of the mixing proportions to zero the usual asymptotics for the likelihood ratio test does not apply
- In the literature two alternatives have been suggested – bootstrapped LRT and Lo-Mendell-Rubin LRT
- Both suggestions perform well in simulations with a slight advantage to the bootstrapped LRT
- All simulations are done under the maintained assumption of a correctly specified model

Determining the number of clusters (cont'd)

- Comparisons using information criteria allow for the possibility of comparing a set of models
- The criteria most often used is the BIC criteria
- The assumptions justifying the use of information criteria are various regularity conditions which in many situations are not fulfilled
- However, there is ample theoretical and empirical support for the use of these criteria in connection with model-based clustering
- The BIC criteria is made up of two opposite contributions – a large value of the maximized likelihood function and a penalty based on the number of parameters as well as the number of observations
- As such we are looking for large BIC values but since the likelihood function is negative this corresponds to numerically small BIC values

Background and analysis

- We will use model-based clustering for our HBAT example
- Letting the number of clusters be determined by BIC we end up in a 3 cluster solution (diagonal, equal volume, equal shape)
- Forcing a 4 cluster solution leads to the same kind of components