Model-Based Clustering I

Prof. Sharon McNicholas

STATS 780/CSE 780

Introduction

- We have seen hierarchical clustering, k-means clustering, and k-medoids clustering.
- Next, we look at mixture model-based clustering.
- The idea is to cluster based on a statistical (mixture) model.
- This material will take multiple "lectures" to cover.
- Some of the material in this lecture is taken from McNicholas (2016).
- Note that bibliographic references are given at the end of these slides.

Defining a Cluster: Similarity

- We know that clustering, or unsupervised learning, techniques are used to find labels when the observations are unlabelled or treated as such.

- Clustering is very often described as finding groups of observations such that observations within a group are more similar to one another than they are to observations in other groups.

- This definition, however, is problematic.

- An alternative definition is in terms of the modes in a mixture model.

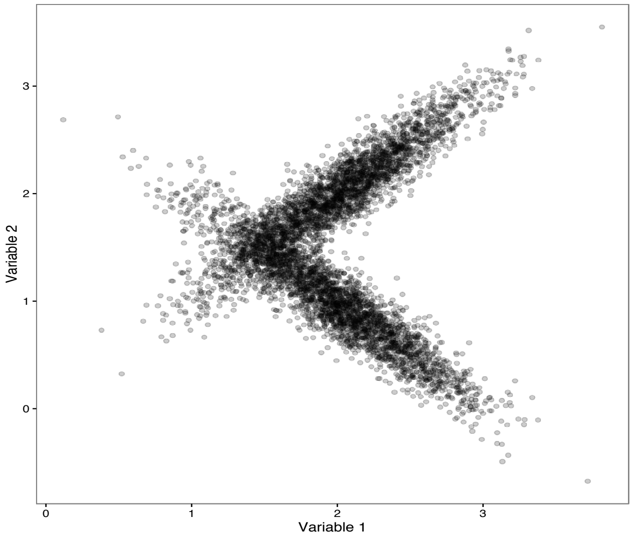
Defining a Cluster: Modes

- A cluster can be defined as a mode.

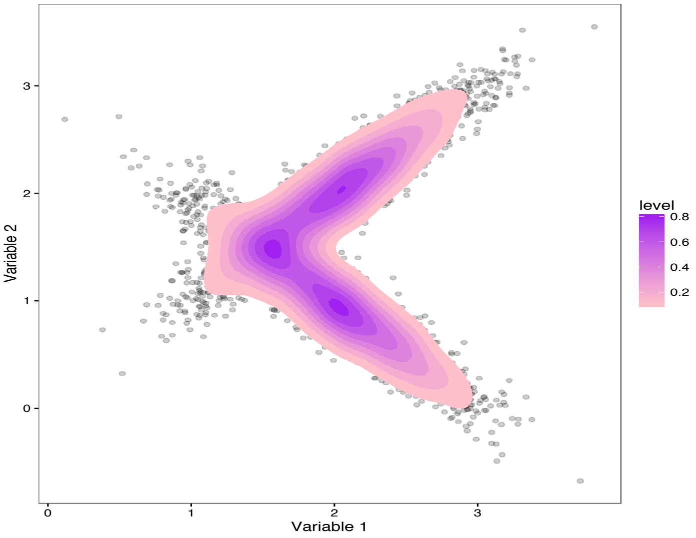
- The principal problem with this can be seen by generating two overlapping Gaussian components such that there are clearly three modes.

- Consider the following figures.

Defining a Cluster: Modes contd.



Defining a Cluster: Modes contd.



Defining a Cluster: Modes contd.

- This example illustrates three modes but two clusters.

- I think a better way of defining a cluster is in terms of a component in an appropriate mixture model.

- Before thinking further about this, we need to see the idea of a (parametric finite) mixture model.

Finite Mixture Models

- A random vector $\mathbf{X}$ arises from a parametric finite mixture distribution if, for all $\mathbf{x} \subset \mathbf{X}$, its density can be written
$$f(\mathbf{x} \mid \boldsymbol{\vartheta}) = \sum_{g=1}^{G} \pi_g f_g(\mathbf{x} \mid \boldsymbol{\theta}_g), \tag{1}$$
where $\pi_g > 0$, such that $\sum_{g=1}^{G} \pi_g = 1$, is the gth mixing proportion, $f_g(\mathbf{x} \mid \boldsymbol{\theta}_g)$ is the gth component density, and $\boldsymbol{\vartheta} = (\boldsymbol{\pi}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_G)$ is the vector of parameters, with $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_G)$.

- Note that $f(\mathbf{x} \mid \boldsymbol{\vartheta})$ in (1) is called a G-component finite mixture density. The component densities $f_1(\mathbf{x} \mid \boldsymbol{\theta}_1), f_2(\mathbf{x} \mid \boldsymbol{\theta}_2), \ldots, f_G(\mathbf{x} \mid \boldsymbol{\theta}_G)$ are often taken to be of the same type, i.e., $f_g(\mathbf{x} \mid \boldsymbol{\theta}_g) = f(\mathbf{x} \mid \boldsymbol{\theta}_g)$ for all g.

- See McLachlan and Peel (2000) for further details on finite mixtures.

Model-Based Definition

- Wolfe (1963) defines a cluster as a component in a mixture model.
- McNicholas (2016) is a little more specific:

  > A cluster is a unimodal component within an appropriate finite mixture model.

- Here, an "appropriate" mixture model is one that is appropriate in light of the data under consideration.
- For further details, see McNicholas (2016, Chapter 9).

Some Details

- Note that $\mathbf{z}_i = (z_{i1}, \ldots, z_{iG})$ is considered a realization of $\mathbf{Z}_i$, which is a random variable that follows a multinomial distribution with one draw on G categories with probabilities given by $\pi_1, \ldots, \pi_G$.
- $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$ are assumed independent and identically distributed according to a multinomial distribution with one draw on G categories with probabilities $\pi_1, \ldots, \pi_G$.
- The gth mixing proportion π_g can be interpreted as the *a priori* probability that an observation $\mathbf{x}_i$ belongs to component g.

Some Details contd.

- The corresponding *a posteriori* probability is
$$\mathbb{P}[Z_{ig} = 1 \mid \mathbf{x}_i] = \frac{\pi_g \phi(\mathbf{x}_i \mid \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)}{\sum_{h=1}^{G} \pi_h \phi(\mathbf{x}_i \mid \boldsymbol{\mu}_h, \boldsymbol{\Sigma}_h)}. \tag{2}$$

- Note that the *a posteriori* expected value $\mathbb{E}[Z_{ig} \mid \mathbf{x}_i]$ is also given by (2), i.e., $\mathbb{E}[Z_{ig} \mid \mathbf{x}_i] = \mathbb{P}[Z_{ig} = 1 \mid \mathbf{x}_i]$.

- After the parameters have been estimated, the predicted classifications are given by (2).

Predicted Classifications

- Write
$$\hat{z}_{ig} := \frac{\hat{\pi}_g \phi(\mathbf{x}_i \mid \hat{\boldsymbol{\mu}}_g, \hat{\boldsymbol{\Sigma}}_g)}{\sum_{h=1}^{G} \hat{\pi}_h \phi(\mathbf{x}_i \mid \hat{\boldsymbol{\mu}}_h, \hat{\boldsymbol{\Sigma}}_h)}, \tag{3}$$
for $i = 1, \ldots, n$ and $g = 1, \ldots, G$.

- These *a posteriori* predicted classifications are soft, i.e., each observation has a probability of belonging to each component under the fitted model.

- This is generally considered an advantage of the mixture model-based approach.

- For example, in a $G = 2$ component scenario, it is useful to know whether $\mathbf{z}_5 = (0.01, 0.99)$ or $\mathbf{z}_5 = (0.49, 0.51)$.

Predicted Classifications contd.

- Having soft classifications can be very useful in practice, e.g., when interpreting results or comparing different clustering methods.

- However, in many applications it is desirable to harden the *a posteriori* classifications and the most popular way to do this is to report maximum *a posteriori* (MAP) classifications, i.e., MAP$\{\hat{z}_{ig}\}$.

- Note that

$$\text{MAP}\{\hat{z}_{ig}\} = \begin{cases} 1 & \text{if } g = \arg\max_h \{\hat{z}_{ih}\}, \\ 0 & \text{otherwise.} \end{cases}$$

Model-Based Clustering: Likelihood

- The Gaussian model-based clustering likelihood for $\mathbf{x}_1, \ldots, \mathbf{x}_n$ is

$$\mathcal{L}(\boldsymbol{\vartheta} \mid \mathbf{x}_1, \ldots, \mathbf{x}_n) = \prod_{i=1}^{n} \sum_{g=1}^{G} \pi_g \phi(\mathbf{x}_i \mid \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g).$$

- Use z_{ig} to denote component membership, such that $z_{ig} = 1$ if observation i belongs to component g and $z_{ig} = 0$ otherwise.

- Parameter estimation is usually carried using the expectation-maximization (EM) algorithm (Dempster et al., 1977) or a variant thereof.

- See McLachlan and Krishnan (2008) for details on the EM Algorithm and various extensions.

EM Algorithm

- The EM algorithm is based on the complete-data log-likelihood, i.e.,

$$\begin{aligned} l_{\mathsf{c}}(\boldsymbol{\vartheta}) &= \log \mathcal{L}_{\mathsf{c}}(\boldsymbol{\vartheta} \mid \mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{z}_1, \ldots, \mathbf{z}_n) \\ &= \log \left\{ \prod_{i=1}^{n} \prod_{g=1}^{G} \left[\pi_g \phi(\mathbf{x}_i \mid \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)\right]^{z_{ig}} \right\}. \end{aligned}$$

- E-step: compute (update) $\mathcal{Q}$, the expected value of the complete-data log-likelihood conditional on the current parameter estimates.
- M-step: maximize $\mathcal{Q}$ wrt model parameters.
- E- and M-steps are iterated until some stopping rule is satisfied.
- Let's develop an EM algorithm for Gaussian mixture model-based clustering.

Families of Mixture Models

- For p-dimensional data, the Gaussian mixture model has $Gp(p+1)/2$ free parameters in the component covariance matrices alone.
- Parsimonious *families* of mixture models are developed by parameterizing the covariance structure and imposing constraints to give a variety of models.
- Usually, all models in a family are fitted and the "best" one is selected.
- The Gaussian parsimonious clustering models (GPCMs) make up the best known model-based clustering family in the literature.
- The GPCMs are supported by the R packages `mclust`, `mixture`, and `Rmixmod`.

GPCMs: The Covariance Structure

- Banfield and Raftery (1993) exploit an eigenvalue decomposition of the component covariance matrices for the Gaussian mixture model.
- This eigen-decomposition is given by

$$\boldsymbol{\Sigma}_g = \lambda_g \boldsymbol{\Gamma}_g \boldsymbol{\Delta}_g \boldsymbol{\Gamma}_g',$$

  where

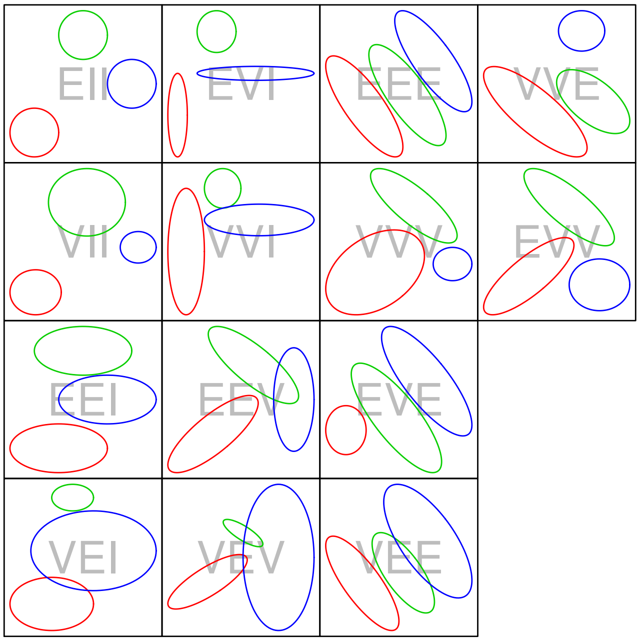
  - λ_g is a constant,
  - $\boldsymbol{\Gamma}_g$ is a matrix of eigenvectors of $\boldsymbol{\Sigma}_g$, and
  - $\boldsymbol{\Delta}_g$, with $|\boldsymbol{\Delta}_g| = 1$, is a diagonal matrix with entries proportional to the eigenvalues of $\boldsymbol{\Sigma}_g$.

- Celeux and Gavaert (1995) use this decomposition to develop a family of 14 GPCMs.

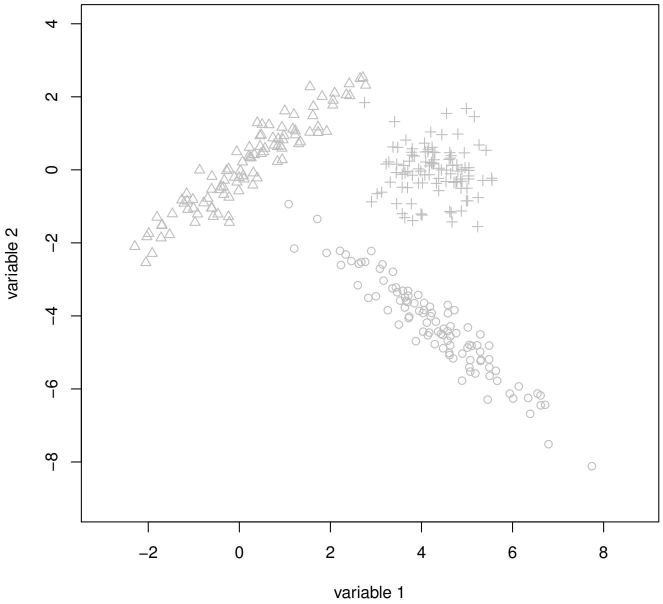
The GPCM Models

Model	Volume	Shape	Orientation	$\boldsymbol{\Sigma}_g$	No. Covariance Parameters
EII	Equal	Spherical	–	$\lambda\mathbf{I}$	1
VII	Variable	Spherical	–	$\lambda_g\mathbf{I}$	G
EEI	Equal	Equal	Axis-Aligned	$\lambda\boldsymbol{\Delta}$	p
VEI	Variable	Equal	Axis-Aligned	$\lambda_g\boldsymbol{\Delta}$	$p+G-1$
EVI	Equal	Variable	Axis-Aligned	$\lambda\boldsymbol{\Delta}_g$	$pG-G+1$
VVI	Variable	Variable	Axis-Aligned	$\lambda_g\boldsymbol{\Delta}_g$	pG
EEE	Equal	Equal	Equal	$\lambda\boldsymbol{\Gamma}\boldsymbol{\Delta}\boldsymbol{\Gamma}'$	$p(p+1)/2$
EEV	Equal	Equal	Variable	$\lambda\boldsymbol{\Gamma}_g\boldsymbol{\Delta}\boldsymbol{\Gamma}_g'$	$Gp(p+1)/2-(G-1)p$
VEV	Variable	Equal	Variable	$\lambda_g\boldsymbol{\Gamma}_g\boldsymbol{\Delta}\boldsymbol{\Gamma}_g'$	$Gp(p+1)/2-(G-1)(p-1)$
VVV	Variable	Variable	Variable	$\lambda_g\boldsymbol{\Gamma}_g\boldsymbol{\Delta}_g\boldsymbol{\Gamma}_g'$	$Gp(p+1)/2$
EVE	Equal	Variable	Equal	$\lambda\boldsymbol{\Gamma}\boldsymbol{\Delta}_g\boldsymbol{\Gamma}'$	$p(p+1)/2+(G-1)(p-1)$
VVE	Variable	Variable	Equal	$\lambda_g\boldsymbol{\Gamma}\boldsymbol{\Delta}_g\boldsymbol{\Gamma}'$	$p(p+1)/2+(G-1)p$
VEE	Variable	Equal	Equal	$\lambda_g\boldsymbol{\Gamma}\boldsymbol{\Delta}\boldsymbol{\Gamma}'$	$p(p+1)/2+(G-1)$
EVV	Equal	Variable	Variable	$\lambda\boldsymbol{\Gamma}_g\boldsymbol{\Delta}_g\boldsymbol{\Gamma}_g'$	$Gp(p+1)/2-(G-1)$
VVV	Variable	Variable	Variable	$\lambda_g\boldsymbol{\Gamma}_g\boldsymbol{\Delta}_g\boldsymbol{\Gamma}_g'$	$Gp(p+1)/2$

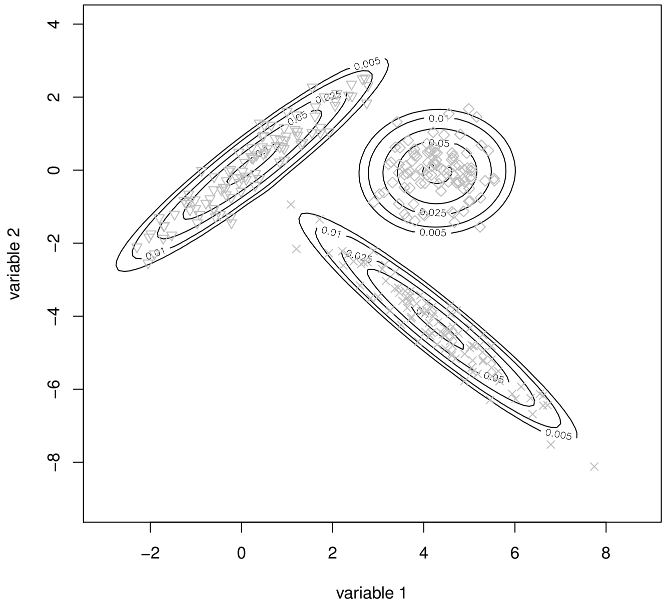
GPCM Models: A “Typical” View



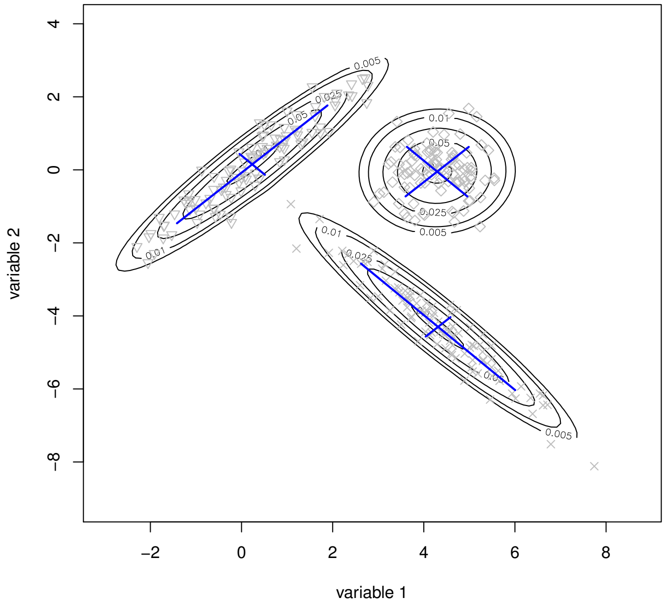
GPCM Models: `x2` Data (EVE)



GPCM Models: x2 Example (EVE)



GPCM Models: x2 Example (EVE)



Model Selection: The BIC

- After all members of a family are fitted, the BIC (Schwarz, 1978) can be used to select the best model (covariance structure and G).
- The BIC can be written

  $$\text{BIC} = 2l(\boldsymbol{\vartheta}) - \rho \log n,$$

  where $\hat{\boldsymbol{\vartheta}}$ is the MLE of $\boldsymbol{\vartheta}$, ρ is the number of free parameters, and n is the number of observations.
- Since its use in the late 1990s, the BIC has been by far the most popular approach for mixture model selection.
- There is much more to be said on all of this material.
- But now, some examples.

References

- Banfield, J.D. and Raftery, A.E. (1993), 'Model-based Gaussian and non-Gaussian clustering', *Biometrics* **49**(3), 803–821.
- Celeux, G. and Govaert, G. (1995), 'Gaussian parsimonious clustering models', *Pattern Recognition* **28**(5), 781–793.
- Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977), 'Maximum likelihood from incomplete data via the EM algorithm', *Journal of the Royal Statistical Society: Series B* **39**(1), 1–38.
- McLachlan, G.J., and Krishnan, T. (2008), *The EM Algorithm and Extensions*, 2nd ed., New York: Wiley.
- McLachlan, G.J. and Peel, D. (2000), *Finite Mixture Models*, New York: John Wiley & Sons.
- McNicholas, P.D. (2016), *Mixture Model-Based Classification*, Boca Raton: Chapman & Hall/CRC Press.
- Schwarz, G. (1978), 'Estimating the dimension of a model', *The Annals of Statistics* **6**(2), 461–464.
- Wolfe, J.H. (1963), "Object Cluster Analysis of Social Areas", Master's thesis, University of California, Berkeley.