Model-based clustering and outlier detection with missing data

Edoardo Protti Ayoube Ighissou

University of Milan - Bicocca

e.protti3@campus.unimib.it

a.ighissou@campus.unimib.it

Advanced Foundations of Statistics
January 19, 2023

Presentation Overview

Motivation

- What is the current state of the art?
- What happens when there are missing data (MCAR, MAR, MNAR)?

Goal

- **Advantages** and **drawbacks** of some multivariate mixture model such as:
  - Multivariate Student's t distribution
  - Multivariate Normal
- **Extend** the mixture of CN distributions for data sets with values missing at random (**MAR**)

Definitions

Definition

A d-variate random vector $\boldsymbol{X} = (X_1, \ldots, X_d)^\top$ is said to follow an $\boldsymbol{MCN}$ distribution with mean vector $\boldsymbol{\mu}$, scale matrix $\boldsymbol{\Sigma}$, proportion of good points $\alpha \in (0.5, 1)$, and degree of contamination $\eta > 1$ if its joint pdf is given by:

$$f_{\mathrm{MCN}}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \alpha, \eta) = \alpha f_{\mathrm{MN}}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) + (1 - \alpha) f_{\mathrm{MN}}(\boldsymbol{x}; \boldsymbol{\mu}, \eta\boldsymbol{\Sigma})$$

Definitions

Definition

A d-variate random vector $\boldsymbol{X}$ is said to follow a mixture of G MCN (MCNM) distributions if its pdf can be written as:

$$f_{\text{MCNM}}(\boldsymbol{x}; \boldsymbol{\Psi}) = \sum_{g=1}^{G} \pi_g f_{\text{MCN}}\left(\boldsymbol{x}; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g, \alpha_g, \eta_g\right)$$

Parameter Estimation

In order to estimate the parameters of the MCNM we can use the EM algorithm with the complete-data likelihood. MCNM has two sources of missing or unobserved data, which can be summarized with the following set:

$$\mathcal{D} = \{\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{V}\} = \{\boldsymbol{x}_i, \boldsymbol{z}_i, \boldsymbol{v}_i\}_{i=1}^{n}$$

But when the data are characterized by missing values at random there is a third source of missing data, indeed each observation can be decomposed in observed and missing values, so that the complete data is given by:

$$\{\boldsymbol{X}^o, \boldsymbol{X}^m, \boldsymbol{Z}, \boldsymbol{V}\} = \{\boldsymbol{x}_i^o, \boldsymbol{x}_i^m, \boldsymbol{z}_i, \boldsymbol{v}_i\}_{i=1}^{n}$$

Parameter Estimation

Furthermore, in order to understand the steps of the EM algorithm, it is useful to see how the complete-data log-likelihood is defined.

$$l(\boldsymbol{\Psi}; \mathcal{D}) = l_1(\boldsymbol{\pi}; \mathcal{D}) + l_2(\boldsymbol{\alpha}; \mathcal{D}) + l_3(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\eta}; \mathcal{D})$$

Parameter Estimation

With l_1, l_2, and l_3 defined as:

$$l_1(\pi;\mathcal{D}) = \sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig}\ln\pi_g$$

$$l_2(\boldsymbol{\alpha}; \mathcal{D}) = \sum_{i=1}^{n} \sum_{g=1}^{G} z_{ig} \left[v_{ig} \ln \alpha_{\mathrm{g}} + \left(1 - v_{ig}\right) \ln \left(1 - \alpha_{\mathrm{g}}\right) \right]$$

$$l_3(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\eta}; \mathcal{D}) = -\frac{1}{2}\sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig}\left[\ln\left|\boldsymbol{\Sigma}_g\right| + d\left(1 - v_{ig}\right)\ln\eta_g\right] + \\ -\frac{1}{2}\sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig}\left(v_{ig} + \frac{1 - v_{ig}}{\eta_g}\right)\delta\left(\begin{bmatrix} \boldsymbol{x}_i^o \\ \boldsymbol{x}_i^m \end{bmatrix}, \boldsymbol{\mu}_g; \boldsymbol{\Sigma}_\mathrm{g}\right)$$

Methodology

For parameter estimation the ECM algorithm is used (one E-step and two CM-steps, until convergence)

The E-step for iteration $r+1$ requires the calculation of the following expectations:

$$Z_{ig}, Z_{ig}V_{ig}, \text{ and } Z_{ig}\left(V_{ig}+\frac{1-V_{ig}}{\eta_g}\right)\delta\left(\begin{bmatrix} \boldsymbol{x}_i^o \\ \boldsymbol{X}_i^m \end{bmatrix}, \boldsymbol{\mu}_g; \boldsymbol{\Sigma}_g\right)$$

Given the observed data $\boldsymbol{x}_i^o$ and the current estimate of $\boldsymbol{\Psi}^{(r)}$.

Methodology

These formulas can be derived from the first expectation:

$$E_{\boldsymbol{\Psi}^{(r)}}\left(Z_{ig} \mid \boldsymbol{x}_i^o\right) = \frac{\pi_g^{(r)} f_{\text{MCN}}\left(\boldsymbol{x}_i^o; \boldsymbol{\mu}_g^{o(r)}, \boldsymbol{\Sigma}_g^{oo(r)}, \alpha_g^{(r)}, \eta_g^{(r)}\right)}{\sum_{h=1}^{G} \pi_h^{(r)} f_{\text{MCN}}\left(\boldsymbol{x}_i^o; \boldsymbol{\mu}_h^{o(r)}, \boldsymbol{\Sigma}_h^{oo(r)}, \alpha_h^{(r)}, \eta_h^{(r)}\right)} =: \tilde{z}_{ig}^{(r)}$$

Methodology

In the first CM-step the parameters π_g, α_g, $\boldsymbol{\mu}_g$, and $\boldsymbol{\Sigma}_g$ are updated by:

$$\pi_g^{(r+1)} = \frac{1}{n}\sum_{i=1}^{n} \tilde{z}_{ig}^{(r)},$$

$$\alpha_g^{(r+1)} = \frac{\sum_{i=1}^{n} \tilde{z}_{ig}^{(r)} \tilde{v}_{ig}^{(r)}}{\sum_{i=1}^{n} \tilde{z}_{ig}^{(r)}},$$

$$\boldsymbol{\mu}_g^{(r+1)} = \frac{\sum_{i=1}^{n} \tilde{w}_{ig}^{(r)} \begin{bmatrix} \boldsymbol{x}_i^o \\ \tilde{\boldsymbol{x}}_{ig}^{(r)} \end{bmatrix}}{\sum_{i=1}^{n} \tilde{w}_{ig}^{(r)}},$$

$$\boldsymbol{\Sigma}_g^{(r+1)} = \frac{\sum_{i=1}^{n} \tilde{w}_{ig}^{(r)} \tilde{\boldsymbol{\Sigma}}_{ig}^{(r)}}{\sum_{i=1}^{n} \tilde{z}_{ig}^{(r)}},$$

Methodology

In the second CM-step, η_g is updated with the maximization of the observed log-likelihood (using a variation of the ECM algorithm). Indeed, η_g can be updated in the following way:

$$\eta_g^{(r+1)} = \max\left\{\eta^*, \frac{\sum_{i=1}^n \tilde{z}_{ig}^{(r)}\left(1-\tilde{v}_{ig}^{(r)}\right)\delta\left(\boldsymbol{x}_i^o, \boldsymbol{\mu}_\mathrm{g}^{o(r+1)}; \boldsymbol{\Sigma}_\mathrm{g}^{oo(r+1)}\right)}{\sum_{i=1}^n d_i^o \tilde{z}_{ig}^{(r)}\left(1-\tilde{v}_{ig}^{(r)}\right)}\right\}$$

Algorithm implementation

Initialization is an important step in EM as the algorithm is deterministic. For this application, the k-medoids algorithm was chosen as it provides a robust clustering technique.

Convergence is determined with the Aitken acceleration criterion:

$$a^{(r+1)} = \frac{l^{(r+2)} - l^{(r+1)}}{l^{(r+1)} - l^{(r)}}$$

Algorithm implementation

In particular, convergence is reached once

$$l_{\infty}^{(r+2)} - l^{(r+1)} < \epsilon$$

For some $\epsilon > 0$, with

$$l_{\infty}^{(r+2)} = l^{(r+1)} + \frac{1}{1 - a^{(r+1)}} \left[l^{(r+2)} - l^{(r+1)} \right].$$

Algorithm Implementation

Upon convergence, cluster membership and whether the observation is good or bad are determined with the MAP probabilities.

When G is not known in advance, a range of values can be used and the best choice is determined using the BIC:

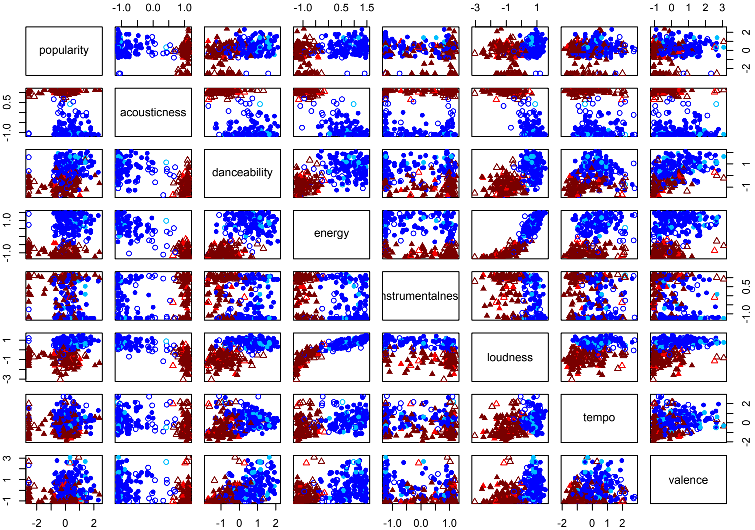
$$BIC = 2l(\hat{\Psi}) - m \ln n$$

Where $\hat{\Psi}$ and $l(\hat{\Psi})$ correspond to the estimated parameters and the associated log-likelihood, and m is the number of free-parameters in the model.

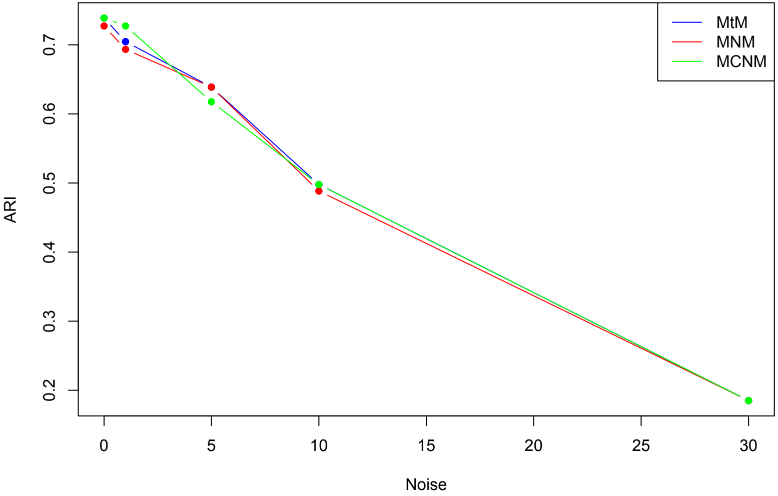
Real Application

We will apply the techniques discussed to a Music Genre dataset. In the pre-processing step we remove all categorical data so as to only deal with numerical values

MCNM Clustering Results



MCNM ARI



Estimated Parameters

Noise %	π_1	π_2	α_1	α_2	η_1	η_2
0	0.551	0.449	0.785	0.804	2.39	3.01
1	0.535	0.465	0.771	0.886	2.28	4.41
5	0.561	0.439	0.703	0.918	1.93	5.05
10	0.607	0.393	0.649	0.916	1.35	3.56
30	0.588	0.442	0.756	0.744	1.33	1.01

Questions? Comments?