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
December 19, 2022

Presentation Overview

Motivation

- Why Finite Mxiture Models?
- Which methods in the literature are useful for outlier detection?
- What happens when there are missing data?

Goal

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

Definitions

Definition

Let $\mathbf{X} = (X_1, \ldots, X_d)^T$ be a d-random vector. $\mathbf{X}$ is said to follow a **MCN** distribution with mean vector $\boldsymbol{\mu}$ and scale matrix $\boldsymbol{\Sigma}$, proportion of good points $\boldsymbol{\alpha} \in (\frac{1}{2}, 1)$, and degree of contamination $\boldsymbol{\eta} > 1$ if its joint pdf is given by:

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

Definitions

Definition

A *d*-random vector **X** is said to follow a mixture of G **MCN** distributions if its pdf can be written as:

$$f_{MCNM}(x;\boldsymbol{\Psi}) = \sum_{g=1}^{G} \pi_g f_{MCN}(\boldsymbol{x};\boldsymbol{\mu_g},\boldsymbol{\Sigma_g},\alpha_g,\eta_g)$$

Where:

- $\pi_g > 0$ and $\sum_{g=1}^{G} \pi_g = 1$
- $\boldsymbol{\Psi} = \{\boldsymbol{\pi},\boldsymbol{\theta}\}$ with $\boldsymbol{\pi} = \{\pi_g\}_{g=1}^{G}$ and $\boldsymbol{\theta} = \{\theta_g\}_{g=1}^{G}$ where $\boldsymbol{\theta}_g = \{\boldsymbol{\mu}_g,\boldsymbol{\Sigma}_g,\alpha_g,\eta_g\}$

Slide without title.

The End

Questions? Comments?