Identifying Subpopulations of Hierarchically Structured Data using a Semi-Supervised Mixture Modeling Approach

Andrew Simpson Semhar Michael Dylan Borchert Christopher Saunders

Department of Mathematics and Statistics, South Dakota State University

Acknowledgments

The research detailed in this presentation was supported by Award No. 2019-DU-BX-0011 (Dr. Larry Tang as the PI) awarded by the National Institute of Justice, Justice Programs, US Department of Justice. The opinions and conclusions or recommendations expressed in this presentation are those of the author and do not necessarily represent those of the Department of Justice.

Introduction to Forensic Statistics

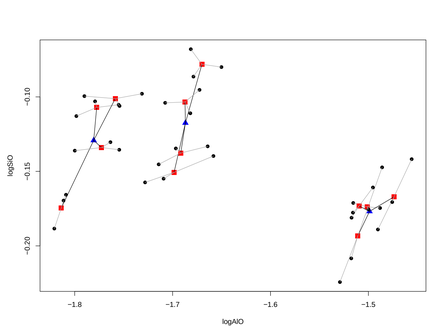
- Suppose someone breaks into a house and smashes a window in the process
- Glass fragments from the scene of the crime and taken as evidence
- A suspect is found will glass fragments stuck to his shoe
- What's the likelihood the fragments from the shoe and the fragments from the scene of the crime came from the same window?
- We need a model to calculate the likelihood ratio [1]

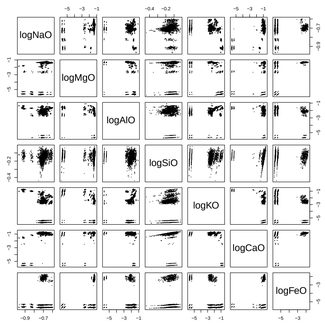
$$LR = \frac{f(e_{u1}, e_{u2}; \theta)}{f(e_{u1}; \theta) f(e_{u2}; \theta)}$$

- Estimates for θ are also needed

Hierarchical Sampling of Trace Elements as Evidence

Forensic evidence often arises from a hierarchical sampling process





Because $\boldsymbol{X}_{ij}$ is sampled from a hierarchical sampling process, $\boldsymbol{X}_{ij}$ can be written as the hierarchical random effects model [5],

$$\boldsymbol{X}_{ij} = \boldsymbol{a}_i + \boldsymbol{\epsilon}_{ij},$$

where

- $\boldsymbol{a}_i | Z_i = k \sim N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k^{'})$ is the source sampled from the k^{th} subpopulation
- $\boldsymbol{\epsilon}_{ij} \sim N(0, \boldsymbol{\Sigma}_\epsilon)$ is the piece of evidence sampled from the i^{th} within source distribution
- Σ_ϵ is the within source covariance matrix

Note that $\boldsymbol{X}_{ij} | Z_i = k \sim N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ where $\boldsymbol{\Sigma}_k = \boldsymbol{\Sigma}_k^{'} + \boldsymbol{\Sigma}_\epsilon$ given $\boldsymbol{X}_{ij}$ is in the k^{th} subpopulation. The question is how to estimate the parameters of the the GFMM

$$f(\boldsymbol{x}_{ij} | \boldsymbol{\Psi}) = \sum_{k=1}^{K} \tau_k \phi(\boldsymbol{x}_{ij} | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

Semi-Supervised Finite Mixture Model (SSFMM)

- Fragments from the same window are known to have come from the same subpopulation
- We call these observation form a block
- The SSFMM restricts the parameters space to only cases where observations from the same block are in the same subpopulation
- This modification occurs in the E-step [4] of the expectation maximization (EM) algorithm and gives the following result,

$$\ddot{\pi}_{ik} = \frac{\dot{\tau}_k^{|B_{b(i)}|} \prod_{j \in B_{b(i)}} \phi(\boldsymbol{x}_j | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k))}{\sum_{k'=1}^{K} \dot{\tau}_{k'}^{|B_{b(i)}|} \prod_{j \in B_{b(i)}} \phi(\boldsymbol{x}_j | \boldsymbol{\mu}_{k'}, \boldsymbol{\Sigma}_{k'}))}.$$

where

- $B_{b(i)}$ is the block containing the i^{th} observation

The M-step remains unchanged from the EM algorithm with normal Finite Mixture Models.

Application to Trace Evidence

There are three approaches considered in this paper for estimating the parameters of the GFMM.

GFMM Approach: Fits a GFMM using the EM algorithm using source means (using mclust 'VVV'[6]) which provides an estimate for $\hat{\boldsymbol{\Sigma}}_k^{'}$

GFMM+C Approach: Add the estimated within source covariance $\hat{\boldsymbol{\Sigma}}_\epsilon$ to the estimate $\hat{\boldsymbol{\Sigma}}_k^{'}$ from the GFMM approach to get an estimate for $\hat{\boldsymbol{\Sigma}}_k$

SSFMM Approach: The Semi-supervised approach places the fragments into blocks at which point the SSFMM is estimated

Comparison Method

We wish to compare these 3 methods and we will need a metric to compare.

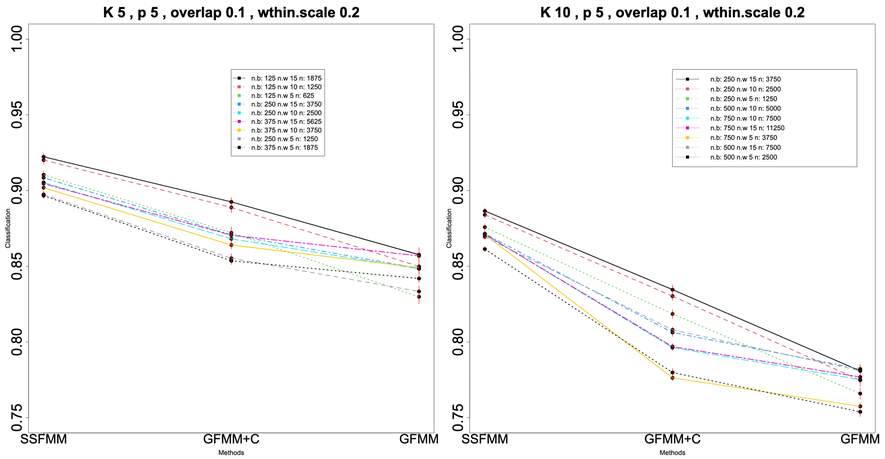
The following comparison method will be employed

1. Create train/test splits were made by randomly removing one within source observation from each source
2. Fit each model using the training dataset
3. The subpopulation membership is predicted for each observation the test dataset
4. The prediction it considered correct if the predicted membership is the same as the membership of it's source in the training dataset after fitting
5. The average classification accuracy is calculated

Comparison Simulation on Hierarchically Sampled Data

A large scale simulation was performed by generating random mixtures with MixSim [3] and varying the following parameters

- Number of components, K
- Number of dimensions, p
- Overlap
- Within covariance scale
- Number of between (n_b) and within source samples (n_w)

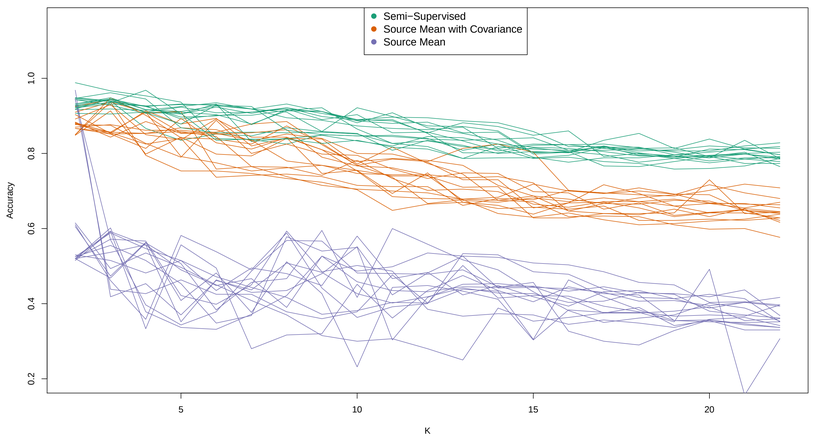


Application to Glass Data

The Zadora glass data set [1] consists of 200 window sources, each with 4 fragments, measured 3 separate times.

The comparison method is performed 15 times on the Zadora glass dataset

K us unknown, thus $K \in \{2, 3, \ldots, 22\}$ is fitted



It can be seen from the graph that the semi-supervised approach outperforms the unsupervised approach in term of prediction accuracy.

References

[1] Colin G. G. Aitken, Grzegorz Zadora, and David Lucy. A Two-Level Model for Evidence Evaluation. *Journal of Forensic Sciences*, 52(2):412–419, March 2007.

[2] Joshua R. Dettman, Alyssa A. Cassabaum, Christopher P. Saunders, Deanna L. Snyder, and JoAnn Buscaglia. Forensic Discrimination of Copper Wire Using Trace Element Concentrations. *Analytical Chemistry*, 86(16):8176–8182, 2014.

[3] Volodymyr Melnykov, Wei-Chen Chen, and Ranjan Maitra. MixSim: An R package for simulating data to study performance of clustering algorithms. *Journal of Statistical Software*, 51(12):1–25, 2012.

[4] Volodymyr Melnykov, Igor Melnykov, and Semhar Michael. Semi-supervised model-based clustering with positive and negative constraints. *Advances in Data Analysis and Classification*, 10(3):327–349, 2015.

[5] Ommen. *Approximate statistical solutions to the forensic identification of source problem*. PhD dissertation, South Dakota State University, 2017.

[6] Luca Scrucca, Michael Fop, T. Brendan Murphy, and Adrian E. Raftery. mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. *The R Journal*, 8(1):289–317, 2016.

andrew.simpson856@jacks.sdstate.edu