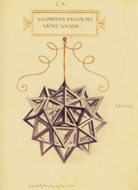


Supervised Learning in Presence of Outliers, Label Noise and Unobserved Classes

Andrea Cappozzo

University of Milano - Bicocca

joint work with **Francesca Greselin** and **Brendan Murphy**

CLADAG 2019 12-th Scientific Meeting; Cassino, September 13

Outline

Problem Statement

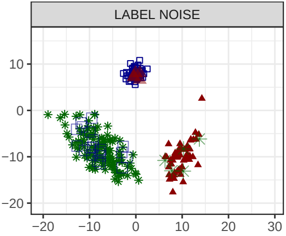
Motivating research question:

- Developing a classifier that jointly deals with 3 important issues often encountered in Supervised Learning

Problem Statement

Motivating research question:

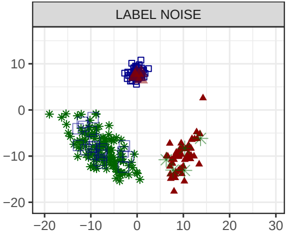
- Developing a classifier that jointly deals with 3 important issues often encountered in Supervised Learning

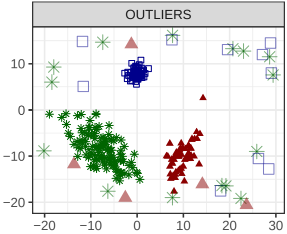


Problem Statement

Motivating research question:

- Developing a classifier that jointly deals with 3 important issues often encountered in Supervised Learning

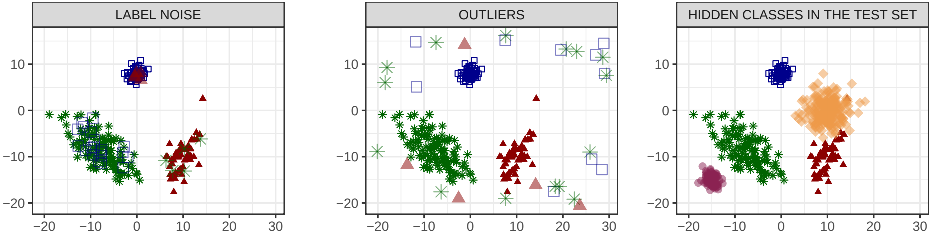




Problem Statement

Motivating research question:

- Developing a classifier that jointly deals with 3 important issues often encountered in Supervised Learning



Model-Based Discriminant Analysis

Model-based discriminant analysis (Fraley and Raftery 2002) is a probabilistic approach for supervised classification.

Model-Based Discriminant Analysis

Model-based discriminant analysis (Fraley and Raftery 2002) is a probabilistic approach for supervised classification.

- $(\mathbf{x}_1, \ldots, \mathbf{x}_N)$ independent learning observations, realizations of a continuous random vector $\mathcal{X} \in \mathbb{R}^p$
- $(\mathbf{l}_1, \ldots, \mathbf{l}_N)$ class labels, such that $l_{ng} = 1$ if observation n belongs to group g and 0 otherwise, $g = 1, \ldots, G$
- $(\mathbf{y}_1, \ldots, \mathbf{y}_M)$ test observations with unknown associated class labels

Model-Based Discriminant Analysis

- Data generating process for **genuine** observations

$$\mathcal{G} \sim Mult_G(1; \tau_1, \ldots, \tau_G)$$

$$\mathcal{X} | \mathcal{G} = g \sim \mathcal{N}_p(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$$

- Joint density of complete learning observation $(\mathbf{x}_n, \mathbf{l}_n)$:

$$f(\mathbf{x}_n, \mathbf{l}_n; \Theta) = \prod_{g=1}^{G} \left[\tau_g \phi(\mathbf{x}_n; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)\right]^{l_{ng}}$$

  - $\phi(\cdot; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$ multivariate normal density distribution
  - τ_g prior probability of the gth class

Eigenvalue Decomposition DA

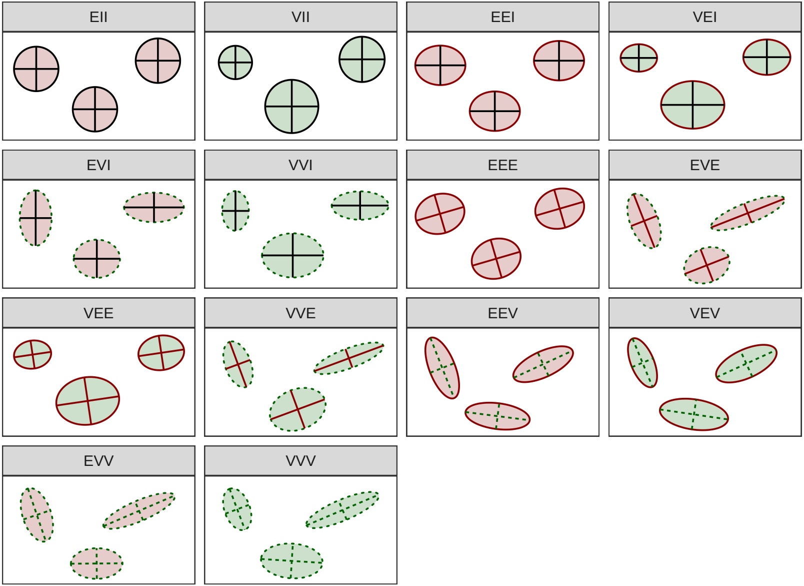
Bensmail and Celeux (Bensmail and Celeux 1996) propose an

Eigenvalue Decomposition for Σ_g

$$\Sigma_g = \lambda_g \boldsymbol{D}_g \boldsymbol{A}_g \boldsymbol{D}_g^{'}$$

- $\boldsymbol{D}_g$ orthogonal matrix of eigenvectors
- $\boldsymbol{A}_g$ diagonal matrix such that $|\boldsymbol{A}_g| = 1$
- $\lambda_g = |\Sigma_g|^{1/p}$ where p denotes the number of variables in the dataset

14 Parsimonious Models



DA Decision Phase

The **Maximum a Posteriori (MAP)** rule is employed for classifying unlabelled observations $\mathbf{y}_m$

$$\hat{z}_{mg} = \mathbb{P}(\mathcal{G} = g | \mathcal{X} = \mathbf{y}_m) = \frac{\hat{\tau}_g \phi(\mathbf{y}_m; \hat{\boldsymbol{\mu}}_g, \hat{\boldsymbol{\Sigma}}_g)}{\sum_{j=1}^{G} \hat{\tau}_j \phi(\mathbf{y}_m; \hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j)}$$

DA Decision Phase

The **Maximum a Posteriori (MAP)** rule is employed for classifying unlabelled observations $\mathbf{y}_m$

$$\hat{z}_{mg} = \mathbb{P}(\mathcal{G} = g | \mathcal{X} = \mathbf{y}_m) = \frac{\hat{\tau}_g \phi(\mathbf{y}_m; \hat{\boldsymbol{\mu}}_g, \hat{\boldsymbol{\Sigma}}_g)}{\sum_{j=1}^{G} \hat{\tau}_j \phi(\mathbf{y}_m; \hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j)}$$

MAP will automatically label such observations as belonging to one of the known classes and will not be able to detect new ones!

Back to the motivating Problem

- **Label Noise:** The class-membership is unreliable for some training observations
- **Outliers:** A proportion of observations might depart from the main structure of the data
- **Unobserved Classes:** only a subset $G \leq E$ of classes might have been encountered in the learning data, with H "hidden" classes in the test such that $E = G + H$

Back to the motivating Problem

- **Label Noise:** The class-membership is unreliable for some training observations
- **Outliers:** A proportion of observations might depart from the main structure of the data
- **Unobserved Classes:** only a subset $G \leq E$ of classes might have been encountered in the learning data, with H "hidden" classes in the test such that $E = G + H$

A **Robust** and **Adaptive** modification to EDDA is needed!

RAEDDA Model

Robust generalization of the *AMDA* methodology (Bouveyron 2014).

- We construct a procedure for maximizing the **trimmed observed data log-likelihood:**

$$\ell_{trim}(\boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathbf{X}, \mathbf{Y}, \mathbf{l}) = \sum_{n=1}^{N} \zeta(\mathbf{x}_n) \sum_{g=1}^{G} \mathrm{l}_{ng} \log\left(\tau_g \phi(\mathbf{x}_n; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)\right) + \\ + \sum_{m=1}^{M} \varphi(\mathbf{y}_m) \log\left(\sum_{g=1}^{E} \tau_g \phi(\mathbf{y}_m; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)\right)$$

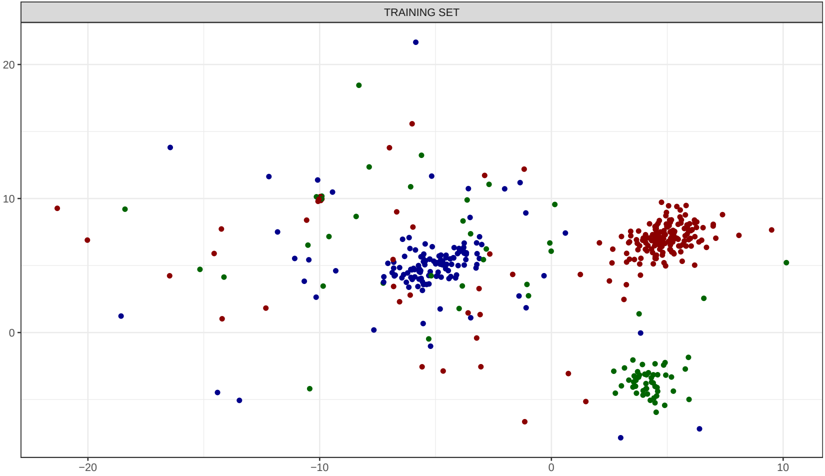
- $\zeta(\cdot)$, $\varphi(\cdot)$ 0-1 trimming indicator functions
- α_l and α_u *trimming level* for the *training* and *test* set
- $\sum_{n=1}^{N} \zeta(\mathbf{x}_n) = \lceil N(1-\alpha_l) \rceil$, $\sum_{m=1}^{M} \varphi(\mathbf{y}_m) = \lceil M(1-\alpha_u) \rceil$

RAEDDA Model

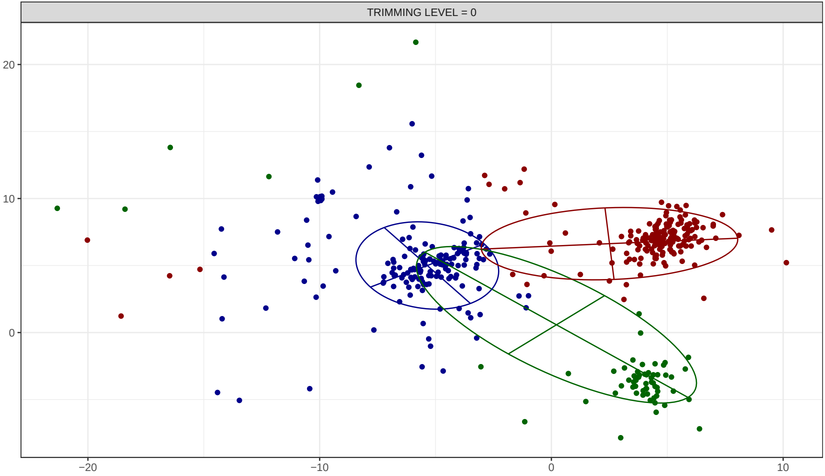
- **Robustness** is achieved employing *impartial trimming* and *constraints* on the parameter space
  - **Impartial Trimming:** observations with the lowest contributions to the overall likelihood will not be accounted for in the parameter estimation (Cuesta-Albertos, Gordaliza, and Matrán 1997)
  - **Constrained Estimation:** eigenvalues-ratio restrictions to avoid singularities and reduce spurious solutions (Ingrassia 2004)

- **Adaptive Learning** is obtained by deploying EM-based approaches to parameter estimation (Bouveyron 2014)
  - **Transductive Approach**
  - **Inductive Approach**

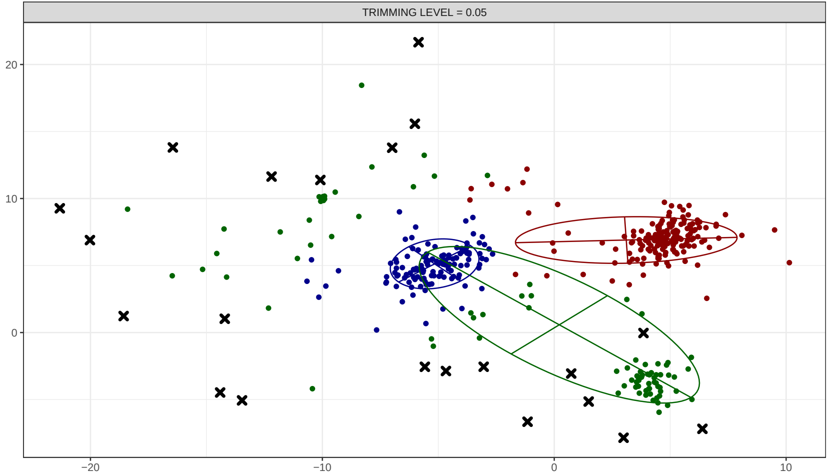
Impartial Trimming: Intuition



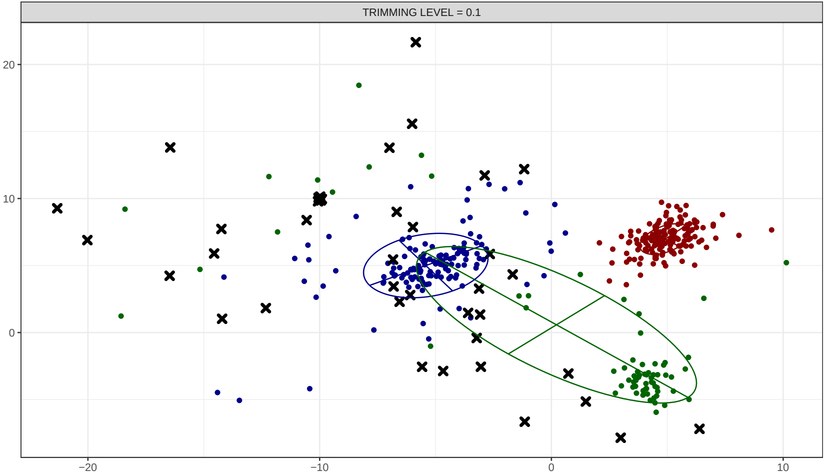
Impartial Trimming: Intuition



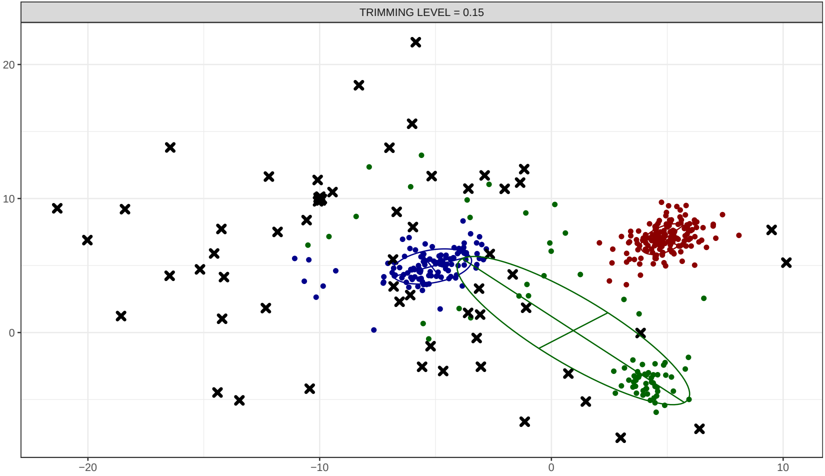
Impartial Trimming: Intuition



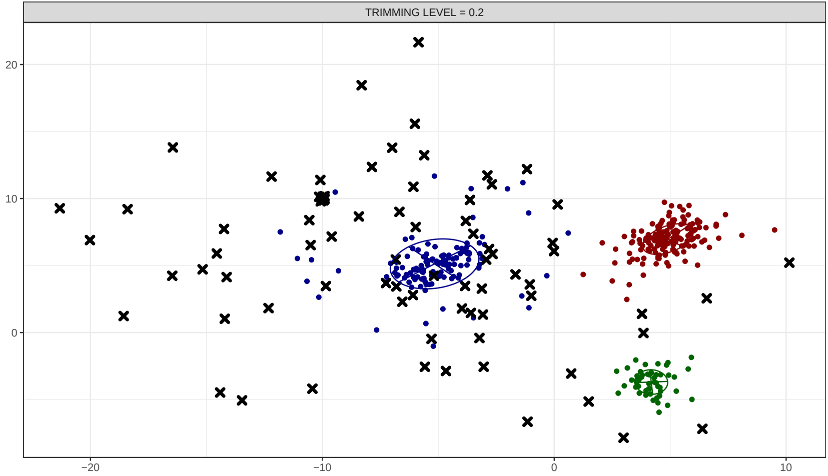
Impartial Trimming: Intuition



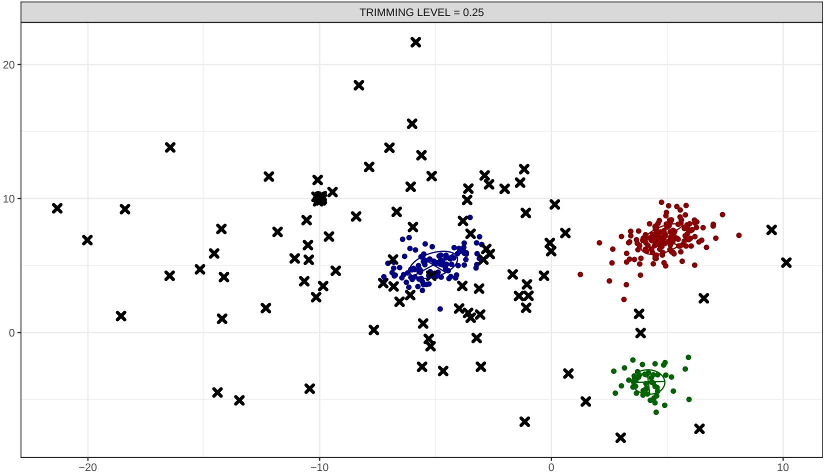
Impartial Trimming: Intuition



Impartial Trimming: Intuition



Impartial Trimming: Intuition



Constrained Estimation

Singularity issues for heteroscedastic covariance matrices $\boldsymbol{\Sigma}_g$ are avoided considering an eigenvalues-ratio restriction:

$$\Pi_n / \pi_n \leq c$$

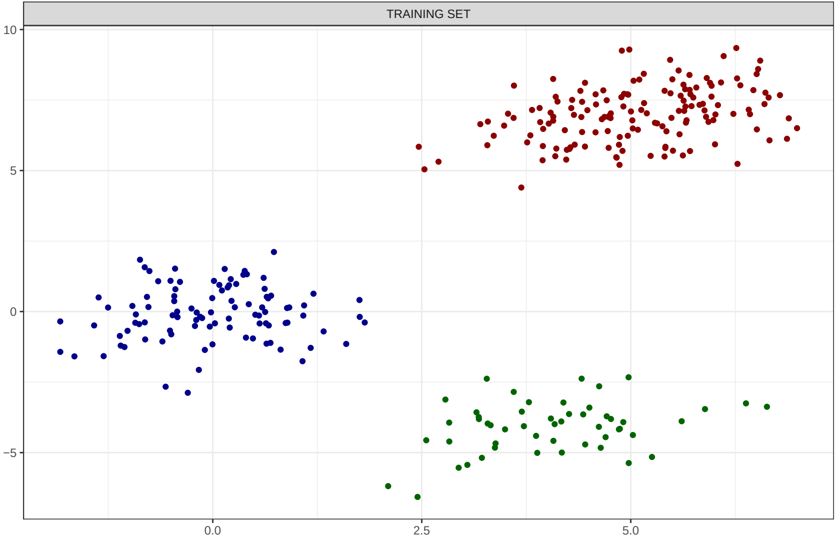
where

$$\Pi_n = \max_{g=1\ldots G} \max_{l=1\ldots p} d_{lg} \quad \text{and} \quad \pi_n = \min_{g=1\ldots G} \min_{l=1\ldots p} d_{lg},$$

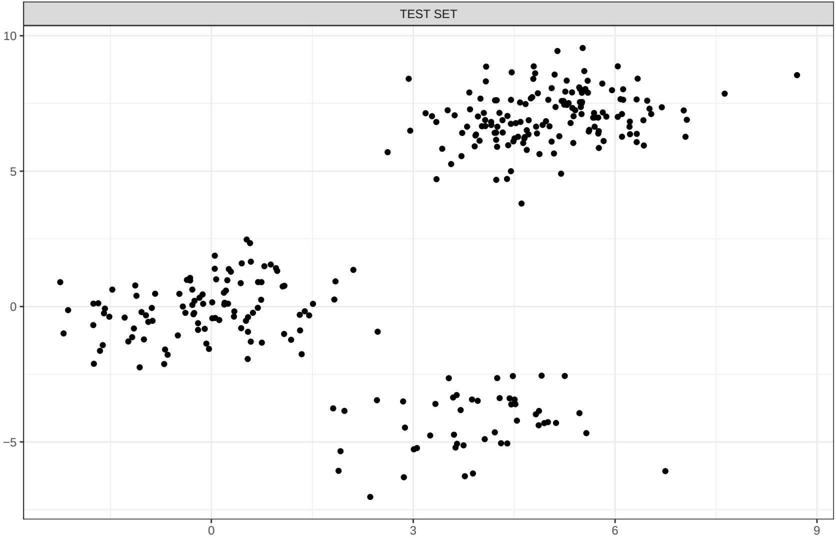
$d_{lg}, l = 1, \ldots, p$ being the eigenvalues of the matrix $\boldsymbol{\Sigma}_g$ and $c \geq 1$ being a fixed constant (García-Escudero et al. 2008)

Still needed when either the **shape** or the **volume** is free to vary across components (García-Escudero et al. 2018)!

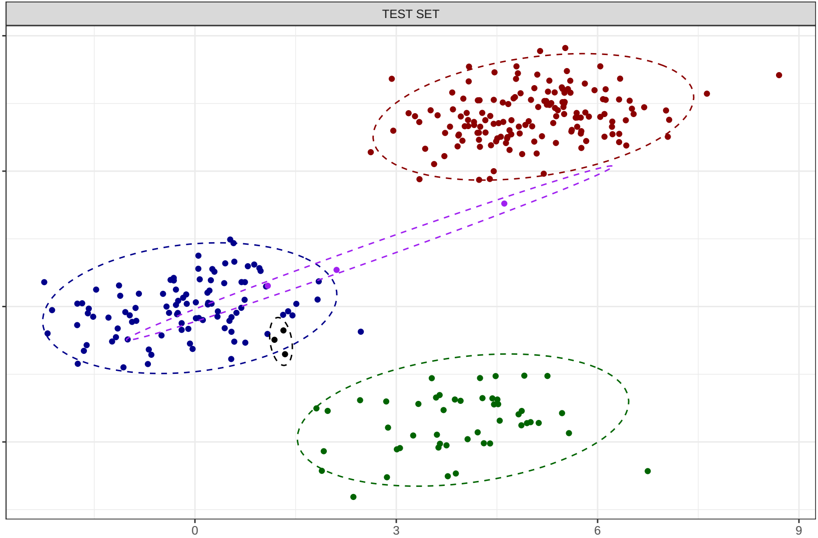
Spurious Solutions: Intuition



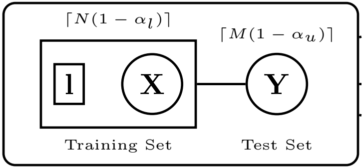
Spurious Solutions: Intuition



Unconstrained Spurious Fitting



Transductive Approach: Scheme

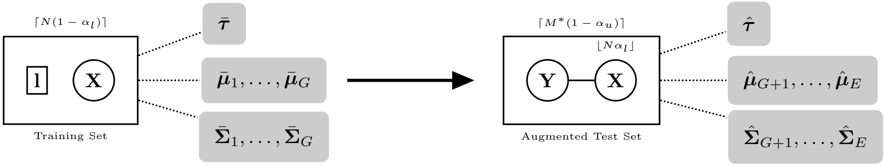


$\hat{\boldsymbol{\tau}}$

$\hat{\boldsymbol{\mu}}_1, \ldots, \hat{\boldsymbol{\mu}}_G, \hat{\boldsymbol{\mu}}_{G+1}, \ldots, \hat{\boldsymbol{\mu}}_E$

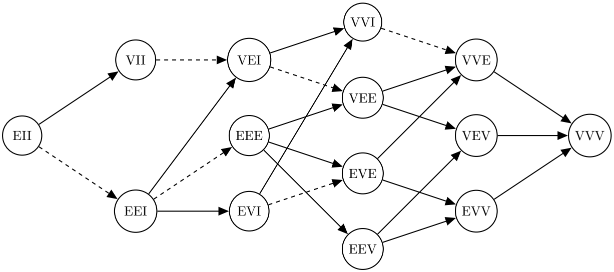
$\hat{\boldsymbol{\Sigma}}_1, \ldots, \hat{\boldsymbol{\Sigma}}_G, \hat{\boldsymbol{\Sigma}}_{G+1}, \ldots, \hat{\boldsymbol{\Sigma}}_E$

Inductive Approach: Scheme



Robust Learning Phase

Robust Discovery Phase



Model Selection using BIC

We propose to use the Robust BIC (Cerioli et al. 2018) to select:

1. Best model among the 14 covariance structures
2. H number of extra classes
3. Constant c constraining the allowed differences among group scatters

$$RBIC = 2\ell_{trim}(\hat{\boldsymbol{\tau}}, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}) - v_M^c log(n^*)$$

- v_M^c penalty term
- $n^* = \begin{cases} \lceil N(1-\alpha_l) \rceil + \lceil M(1-\alpha_u) \rceil & \text{Transductive EM} \\ \lceil M^*(1-\alpha_u) \rceil & \text{Inductive EM} \end{cases}$

More on v_M^c

$$v_M^c = \kappa + \gamma + (\delta - 1)\left(1 - \frac{1}{c}\right) + 1$$

- κ number of parameters related to mixing proportions and mean vectors
- γ number of parameters related to orthogonal rotation
- δ number of parameters related to eigenvalues
- $c \geq 1$ constant allowing differences among group scatters

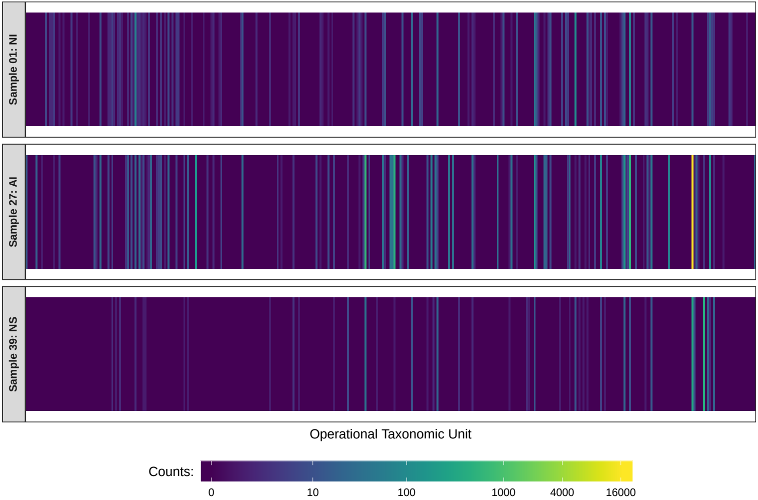
Interesting Fact:
In the Inductive approach the penalty term for the Discovery Phase depends only on the model chosen in this second phase

Grapevine microbiome Dataset

Abundance table of **836 bacterial communities** for 45 grape samples in $E = 3$ regions: NI, AI, NS (Mezzasalma et al. 2018).

Grapevine microbiome Dataset

Abundance table of **836 bacterial communities** for 45 grape samples in $E = 3$ regions: NI, AI, NS (Mezzasalma et al. 2018).



Grapevine microbiome: Results

Data Preprocessing: ROBPCA (Hubert, Rousseeuw, and Vanden Branden 2005)

Grapevine microbiome: Results

Data Preprocessing: ROBPCA (Hubert, Rousseeuw, and Vanden Branden 2005)

Learning scenario:

- 2 of **NI** samples in the training set wrongly labelled as **AI**
- **NS** region not present in the training set

Grapevine microbiome: Results

Data Preprocessing: ROBPCA (Hubert, Rousseeuw, and Vanden Branden 2005)

Learning scenario:

- 2 of **NI** samples in the training set wrongly labelled as **AI**
- **NS** region not present in the training set

Classification Results (inductive approach):

Robust learning phase

# Classes	Covariance Structure EII	VII	EEI	VEI	EVI	**VVI**
2	-719.26	-709.13	-718.97	-712.11	-688.40	**-678.29**

Robust discovery phase

# Classes	Covariance Structure **VVI**
2	-639.85
3	**-506.59**
4	-511.43

Grapevine microbiome: Results

Data Preprocessing: ROBPCA (Hubert, Rousseeuw, and Vanden Branden 2005)

Learning scenario:

- 2 of **NI** samples in the training set wrongly labelled as **AI**
- **NS** region not present in the training set

Classification Results (inductive approach):

Robust learning phase

# Classes	Covariance Structure EII	VII	EEI	VEI	EVI	**VVI**
2	-719.26	-709.13	-718.97	-712.11	-688.40	**-678.29**

Robust discovery phase

# Classes	Covariance Structure **VVI**
2	-639.85
3	**-506.59**
4	-511.43

Classification	Truth NI	NS	AI
NI	2	1	0
AI	0	0	3
HIDDEN GROUP 1	1	14	0

Conclusions

We proposed a model-based discriminant analysis method for anomaly and novelty detection

WIP:

- Selecting the trimming levels α_l and α_u
- Development of `raedda` R package
- Robust wrapper variable selection procedure for adaptive classification with high dimensional data

References

Bensmail, Halima and Gilles Celeux (1996). “Regularized Gaussian discriminant analysis through eigenvalue decomposition”. In: *Journal of the American Statistical Association* 91.436, pp. 1743–1748.

Bouveyron, Charles (2014). “Adaptive mixture discriminant analysis for supervised learning with unobserved classes”. In: *Journal of Classification* 31.1, pp. 49–84.

Cappozzo, Andrea, Francesca Greselin, and Thomas Brendan Murphy (2019). “A robust approach to model-based classification based on trimming and constraints”. In: *Advances in Data Analysis and Classification*. arXiv: 1904.06136.

References

Celeux, Gilles and Gérard Govaert (1995). "Gaussian parsimonious clustering models". In: *Pattern Recognition* 28.5, pp. 781–793.

Cerioli, Andrea et al. (2018). "Finding the number of normal groups in model-based clustering via constrained likelihoods". In: *Journal of Computational and Graphical Statistics* 27.2, pp. 404–416.

Cuesta-Albertos, J. A., A. Gordaliza, and C. Matrán (1997). "Trimmed k-means: An attempt to robustify quantizers". In: *Annals of Statistics* 25.2, pp. 553–576.

References

Fraley, Chris and Adrian E Raftery (2002). “Model-based clustering, discriminant analysis, and density estimation”. In: *Journal of the American Statistical Association* 97.458, pp. 611–631. arXiv: `arXiv:1011.1669v3`.

García-Escudero, Luis A. et al. (2008). “A general trimming approach to robust cluster Analysis”. In: *The Annals of Statistics* 36.3, pp. 1324–1345. arXiv: `0806.2976`.

García-Escudero, Luis Angel et al. (2018). “Eigenvalues and constraints in mixture modeling: geometric and computational issues”. In: *Advances in Data Analysis and Classification* 12.2, pp. 203–233.

References

Hubert, Mia, Peter J Rousseeuw, and Karlien Vanden Branden (2005). "ROBPCA: A New Approach to Robust Principal Component Analysis". In: *Technometrics* 47.1, pp. 64–79.

Ingrassia, Salvatore (2004). "A likelihood-based constrained algorithm for multivariate normal mixture models". In: *Statistical Methods and Applications* 13.2, pp. 151–166.

Mezzasalma, Valerio et al. (2018). "Geographical and Cultivar Features Differentiate Grape Microbiota in Northern Italy and Spain Vineyards". In: *Frontiers in Microbiology* 9.MAY, pp. 1–13.

Thank You!

Transductive EM: initialization

- *Robust Initialization for the G known groups:*
  1. For each known class g, draw a random $(p+1)$-subset J_g and compute its empirical mean $\bar{\boldsymbol{\mu}}_g^{(0)}$ and variance covariance matrix $\bar{\boldsymbol{\Sigma}}_g^{(0)}$ according to the considered parsimonious structure.
  2. Set $\bar{\boldsymbol{\theta}} = \{\bar{\tau}_1^{(0)}, \ldots, \bar{\tau}_G^{(0)}, \bar{\boldsymbol{\mu}}_1^{(0)}, \ldots, \bar{\boldsymbol{\mu}}_G^{(0)}, \bar{\boldsymbol{\Sigma}}_1^{(0)}, \ldots, \bar{\boldsymbol{\Sigma}}_G^{(0)}$ where $\bar{\tau}_1^{(0)} = \ldots = \bar{\tau}_G^{(0)} = 1/G$.
  3. For each $\mathbf{x}_n$, $n = 1, \ldots, N$, compute the conditional density

$$f(\mathbf{x}_n | l_{ng} = 1; \bar{\boldsymbol{\theta}}) = \phi\left(\mathbf{x}_n; \bar{\boldsymbol{\mu}}_g, \bar{\boldsymbol{\Sigma}}_g\right) \qquad g = 1, \ldots, G.$$

     $\lfloor N\alpha_l \rfloor$%of the samples with lowest value are temporarily discarded as possible outliers

Transductive EM: initialization

4. The parameter estimates are updated, based on the non-discarded observations:

$$
\bar{\tau}_g = \frac{\sum_{n=1}^{N} \zeta(\mathbf{x}_n) l_{ng}}{\lceil N(1-\alpha_l) \rceil} \quad g = 1, \dots, G
$$

$$
\bar{\boldsymbol{\mu}}_g = \frac{\sum_{n=1}^{N} \zeta(\mathbf{x}_n) l_{ng} \mathbf{x}_n}{\sum_{n=1}^{N} \zeta(\mathbf{x}_n) l_{ng}} \quad g = 1, \dots, G.
$$

Estimation of $\boldsymbol{\Sigma}_g$ depends on the patterned model

5. Iterate $3-4$ until the $\lfloor N\alpha_l \rfloor$ discarded observations are exactly the same on two consecutive iterations

- Retain the estimates that lead to the highest value of $\ell_{trim}(\bar{\boldsymbol{\tau}}, \bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\Sigma}} | \mathbf{X}, \mathbf{l}) = \sum_{n=1}^{N} \zeta(\mathbf{x}_n) \sum_{g=1}^{G} l_{ng} \log \left[\bar{\tau}_g \phi(\mathbf{x}_n; \bar{\boldsymbol{\mu}}_g, \bar{\boldsymbol{\Sigma}}_g) \right]$ out of `n_init` repetitions

Transductive EM: initialization

- *Robust Initialization for the H hidden classes:*
  1. For each hidden class h, $h = G+1, \dots, E$, draw a random $(p+1)$-subset J_h and compute its empirical mean $\hat{\boldsymbol{\mu}}_h^{(0)}$ and variance covariance matrix $\hat{\boldsymbol{\Sigma}}_h^{(0)}$
  2. Mixing proportions τ_h are drawn from $\mathcal{U}_{[0,1]}$ and initial values set equal to $\hat{\tau}_h^{(0)} = \frac{\tau_h}{\sum_{j=G+1}^{E} \tau_j} \frac{H}{E}$, $h = G+1, \dots, E$ and $\hat{\tau}_g^{(0)} = \bar{\tau}_g \frac{G}{E}$, $g = 1, \dots, G$
- Enforce the eigenvalue-ratio on $\hat{\boldsymbol{\Sigma}}_g^{(0)}$, $g = 1, \dots, E$. A possible choice for c could be:

$$\tilde{c} = \frac{\max_{g=1\dots G} \max_{l=1\dots p} \bar{d}_{lg}}{\min_{g=1\dots G} \min_{l=1\dots p} \bar{d}_{lg}}$$

with $\bar{d}_{lg}$, $l = 1, \dots, p$ being the eigenvalues of the matrix $\bar{\boldsymbol{\Sigma}}_g$, $g = 1, \dots, G$.

Transductive EM: iterations

- *Step 1 - Trimming*: discard the $\lfloor N\alpha_l \rfloor$ observations $\mathbf{x}_n$ with smaller values of

$$D\left(\mathbf{x}_n; \hat{\boldsymbol{\Theta}}^{(k)}\right) = \prod_{g=1}^{E} \left[\phi\left(\mathbf{x}_n; \hat{\boldsymbol{\mu}}_g^{(k)}, \hat{\boldsymbol{\Sigma}}_g^{(k)}\right)\right]^{l_{ng}} \quad n = 1, \ldots, N$$

discard the $\lceil M\alpha_u \rceil$ observations $\mathbf{y}_m$ with smaller values of

$$D\left(\mathbf{y}_m; \hat{\boldsymbol{\Theta}}^{(k)}\right) = \sum_{g=1}^{E} \hat{\tau}_g^{(k)} \phi\left(\mathbf{y}_m; \hat{\boldsymbol{\mu}}_g^{(k)}, \hat{\boldsymbol{\Sigma}}_g^{(k)}\right) \quad m = 1, \ldots, M.$$

Note that $l_{ng} = 0 \,\forall\, n = 1, \ldots, N, g = G + 1, \ldots, E$ is implicitly set in the training set.

Transductive EM: iterations

- *Step 2 - Expectation*: for each non-trimmed observation $\mathbf{y}_m$ compute the posterior probabilities

$$\hat{z}_{mg}^{(k+1)} = \frac{\hat{\tau}_g^{(k)} \phi\left(\mathbf{y}_m; \hat{\boldsymbol{\mu}}_g^{(k)}, \hat{\boldsymbol{\Sigma}}_g^{(k)}\right)}{D\left(\mathbf{y}_m; \hat{\boldsymbol{\theta}}^{(k)}\right)} \quad g = 1, \ldots, E; \quad m = 1, \ldots, M$$

Transductive EM: iterations

- *Step 2 - Expectation*: for each non-trimmed observation $\mathbf{y}_m$ compute the posterior probabilities

$$\hat{z}_{mg}^{(k+1)} = \frac{\hat{\tau}_g^{(k)} \phi\left(\mathbf{y}_m; \hat{\boldsymbol{\mu}}_g^{(k)}, \hat{\boldsymbol{\Sigma}}_g^{(k)}\right)}{D\left(\mathbf{y}_m; \hat{\boldsymbol{\theta}}^{(k)}\right)} \quad g = 1, \dots, E; \quad m = 1, \dots, M$$

- *Step 3 - Constrained Maximization*:

$$\hat{\tau}_g^{(k+1)} = \frac{\sum_{n=1}^{N} \zeta(\mathbf{x}_n) l_{ng} + \sum_{m=1}^{M} \varphi(\mathbf{y}_m) \hat{z}_{mg}^{(k+1)}}{\lceil N(1-\alpha_l) \rceil + \lceil M(1-\alpha_u) \rceil} \quad g = 1, \dots, E$$

$$\hat{\boldsymbol{\mu}}_g^{(k+1)} = \frac{\sum_{n=1}^{N} \zeta(\mathbf{x}_n) l_{ng} \mathbf{x}_n + \sum_{m=1}^{M} \varphi(\mathbf{y}_m) \hat{z}_{mg}^{(k+1)} \mathbf{y}_m}{\sum_{n=1}^{N} \zeta(\mathbf{x}_n) l_{ng} + \sum_{m=1}^{M} \varphi(\mathbf{y}_m) \hat{z}_{mg}^{(k+1)}} \quad g = 1, \dots, E.$$

Estimation of $\boldsymbol{\Sigma}_g$ depends on the considered patterned model and on the eigenvalues-ratio constraint

Robust Learning Phase

- Only labeled observations are considered
- This phase can be seen as a Robust EDDA model, whose parameters are obtained maximizing the associated log-likelihood (Cappozzo, Greselin, and Murphy 2019):

$$\ell_{trim}(\boldsymbol{\tau}, \boldsymbol{\mu}, \Sigma | \boldsymbol{X}, \mathrm{l}) = \sum_{n=1}^{N} \zeta(\boldsymbol{x}_n) \sum_{g=1}^{G} \mathrm{l}_{ng} \log\left(\tau_g \phi(\boldsymbol{x}_n; \boldsymbol{\mu}_g, \Sigma_g)\right)$$

- $\bar{\tau}_g = \frac{\sum_{n=1}^{N} \zeta(\mathbf{x}_n) l_{ng}}{\lceil N(1-\alpha_l) \rceil} \qquad \bar{\boldsymbol{\mu}}_g = \frac{\sum_{n=1}^{N} \zeta(\mathbf{x}_n) l_{ng} \mathbf{x}_n}{\sum_{n=1}^{N} \zeta(\mathbf{x}_n) l_{ng}}, \qquad g = 1, \dots, G$
- $\bar{\Sigma}_g$ depends on the considered patterned model and on the eigenvalues-ratio constraint
- The best model among the 14 covariance structures is selected using Robust BIC (Cerioli et al. 2018)

Robust Discovery Phase

- Considering the augmented test observations $\mathbf{Y}^* = \mathbf{Y} \cup \mathbf{X}^{(\alpha_l)}$, with elements $\mathbf{y}^*_m$, $m = 1, \dots, M^*$, $M^* = (M + \lfloor N\alpha_l \rfloor)$, we look for $E - G$ novel classes maximizing the associated log-likelihood:

$$\ell_{trim}(\boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\Sigma} | \mathbf{Y}^*, \bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\Sigma}}) = \sum_{m=1}^{M^*} \varphi(\mathbf{y}^*_m) \log \left(\sum_{g=1}^{G} \tau_g \phi(\mathbf{y}^*_m; \bar{\boldsymbol{\mu}}_g, \bar{\boldsymbol{\Sigma}}_g) + \right.$$
$$\left. + \sum_{h=G+1}^{E} \tau_h \phi(\mathbf{y}^*_m; \boldsymbol{\mu}_h, \boldsymbol{\Sigma}_h) \right)$$

- $\bar{\boldsymbol{\mu}}_g, \bar{\boldsymbol{\Sigma}}_g$ for $g = 1, \dots, G$ were already estimated in the learning phase
- Only $\hat{\boldsymbol{\mu}}_h, \hat{\boldsymbol{\Sigma}}_h$ for $h = G + 1, \dots, E$ remain to be estimated: again a constrained EM algorithm is employed

Inductive EM: initialization

- *Robust Initialization for the H hidden classes:*
  1. For each hidden class h, $h = G+1, \dots, E$, draw a random $(p+1)$-subset J_h and compute its empirical mean $\hat{\boldsymbol{\mu}}_h^{(0)}$ and variance covariance matrix $\hat{\boldsymbol{\Sigma}}_h^{(0)}$
  2. Mixing proportions τ_h are drawn from $\mathcal{U}_{[0,1]}$ and initial values set equal to $\hat{\tau}_h^{(0)} = \frac{\tau_h}{\sum_{j=G+1}^{E} \tau_j} \frac{H}{E}$, $h = G+1, \dots, E$ and $\hat{\tau}_g^{(0)} = \bar{\tau}_g \frac{G}{E}$, $g = 1, \dots, G$
- Enforce the eigenvalue-ratio on $\hat{\boldsymbol{\Sigma}}_h^{(0)}$, $h = G+1, \dots, E$. A possible choice for c could be:

$$\tilde{c} = \frac{\max_{g=1\dots G} \max_{l=1\dots p} \bar{d}_{lg}}{\min_{g=1\dots G} \min_{l=1\dots p} \bar{d}_{lg}}$$

with $\bar{d}_{lg}$, $l = 1, \dots, p$ being the eigenvalues of the matrix $\bar{\boldsymbol{\Sigma}}_g$, $g = 1, \dots, G$.

Inductive EM: iterations

- *Step 1 - Trimming*:
  Define

$$D_g\left(\mathbf{y}_m^*; \hat{\mathbf{\Theta}}^{(k)}\right) = \begin{cases} \hat{\tau}_g^{(k)}\phi\left(\mathbf{y}_m^*; \bar{\boldsymbol{\mu}}_g, \bar{\mathbf{\Sigma}}_g\right) & g = 1, \ldots, G \\ \hat{\tau}_g^{(k)}\phi\left(\mathbf{y}_m^*; \hat{\boldsymbol{\mu}}_g^{(k)}, \hat{\mathbf{\Sigma}}_g^{(k)}\right) & g = G+1, \ldots, E \end{cases}$$

  discard the $\lceil M^* \alpha_u \rceil$ observations $\mathbf{y}_m^*$ with smaller values of

$$D\left(\mathbf{y}_m^*; \hat{\mathbf{\Theta}}^{(k)}\right) = \sum_{g=1}^{E} D_g\left(\mathbf{y}_m^*; \hat{\mathbf{\Theta}}^{(k)}\right) \quad m = 1, \ldots, M^*.$$

Inductive EM: iterations

- *Step 2 - Expectation*: for each non-trimmed observation $\mathbf{y}_m^*$ compute the posterior probabilities

$$\hat{z}_{mg}^{*(k+1)} = \frac{D_g\left(\mathbf{y}_m^*; \hat{\boldsymbol{\Theta}}^{(k)}\right)}{D\left(\mathbf{y}_m^*; \hat{\boldsymbol{\Theta}}^{(k)}\right)} \quad g = 1, \ldots, E; \quad m = 1, \ldots, M^*.$$

Inductive EM: iterations

- *Step 2 - Expectation*: for each non-trimmed observation $\mathbf{y}_m^*$ compute the posterior probabilities

$$\hat{z}_{mg}^{*(k+1)} = \frac{D_g\left(\mathbf{y}_m^*; \hat{\boldsymbol{\Theta}}^{(k)}\right)}{D\left(\mathbf{y}_m^*; \hat{\boldsymbol{\Theta}}^{(k)}\right)} \quad g = 1, \dots, E; \quad m = 1, \dots, M^*.$$

- *Step 3 - Constrained Maximization*:

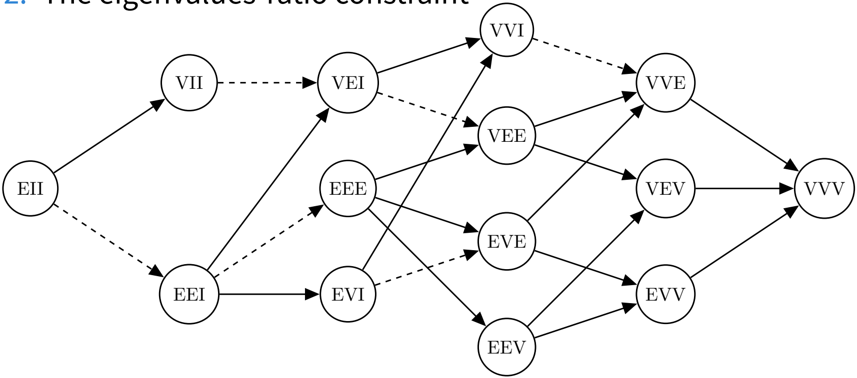
$$\hat{\tau}_g^{(k+1)} = \begin{cases} \bar{\tau}_g \left(1 - \sum_{h=G+1}^{E} \frac{\sum_{m=1}^{M} \varphi(\mathbf{y}_m^*)\hat{z}_{mh}^{*(k+1)}}{\lceil M^*(1-\alpha_u)\rceil}\right) & g = 1, \dots, G \\ \frac{\sum_{m=1}^{M} \varphi(\mathbf{y}_m^*)\hat{z}_{mg}^{*(k+1)}}{\lceil M^*(1-\alpha_u)\rceil} & g = G+1, \dots, E \end{cases}$$

$$\hat{\boldsymbol{\mu}}_h^{(k+1)} = \frac{\sum_{m=1}^{M^*} \varphi(\mathbf{y}_m^*)\hat{z}_{mh}^{*(k+1)}\mathbf{y}_m^*}{\sum_{m=1}^{M^*} \varphi(\mathbf{y}_m^*)\hat{z}_{mh}^{*(k+1)}} \quad h = G+1, \dots, E.$$

Inductive EM algorithm

Estimation of $\Sigma_h, h = G+1, \ldots, E$ depends on:

1. The available patterned models given the model identified in the Learning Phase
2. The eigenvalues-ratio constraint



Partial-order structure in the eigen-decomposition for the covariance matrices. Model complexity increases from left to right. Dashed arrows denote equivalent models in terms of parameters to be estimated in the Discovery Phase.

Explanatory example Σ_h estimation

- Imagine to have selected a *VEE* model in the Learning Phase:

$$\bar{\Sigma}_g = \bar{\lambda}_g \bar{\boldsymbol{D}}\bar{\boldsymbol{A}}\bar{\boldsymbol{D}}', \qquad g = 1, \dots, G$$

- Due to the Inductive approach, only *VEE*, *VVE*, *VEV* and *VVV* models can be selected in the Discovery Phase
- If for example we select to employ a *VEV* model the estimate for Σ_h, $h = G+1, \dots, E$ at the $(k+1)$-th iteration of the EM algorithm will be:

$$\hat{\Sigma}_h^{(k+1)} = \hat{\lambda}_h^{(k+1)} \hat{\boldsymbol{D}}_h^{(k+1)} \bar{\boldsymbol{A}} \hat{\boldsymbol{D}}_h^{(k+1)'} \qquad h = G+1, \dots, E$$

- Closed form solutions are obtained for all models in (Celeux and Govaert 1995), no matter the model selected in the Learning Phase

More on V_M^C: Transductive Approach

Model ID	γ	δ	Constraint needed
EII	0	1	N
VII	0	G	Y
EEI	0	p	N
VEI	0	$G+p-1$	Y
EVI	0	$Gp-(G-1)$	Y
VVI	0	Gp	Y
EEE	$p(p-1)/2$	p	N
VEE	$p(p-1)/2$	$G+p-1$	Y
EVE	$p(p-1)/2$	$Gp-(G-1)$	Y
VVE	$p(p-1)/2$	Gp	Y
EEV	$Gp(p-1)/2$	p	N
VEV	$Gp(p-1)/2$	$G+p-1$	Y
EVV	$Gp(p-1)/2$	$Gp-(G-1)$	Y
VVV	$Gp(p-1)/2$	Gp	Y

$\boldsymbol{\kappa} = Gp + G - 1$ for every model, that is the p entries for $\boldsymbol{\mu}_g, g = 1, \ldots, G$ and $G - 1$ mixing proportions

More on V_M^C: Inductive Approach

Model ID	γ	δ	Constraint needed
EII	0	0	N
VII	0	H	Y
EEI	0	0	N
VEI	0	H	Y
EVI	0	$Hp - H$	Y
VVI	0	Hp	Y
EEE	0	0	N
VEE	0	H	Y
EVE	0	$Hp - H$	Y
VVE	0	Hp	Y
EEV	$Hp(p-1)/2$	0	N
VEV	$Hp(p-1)/2$	H	Y
EVV	$Hp(p-1)/2$	$Hp - H$	Y
VVV	$Hp(p-1)/2$	Hp	Y

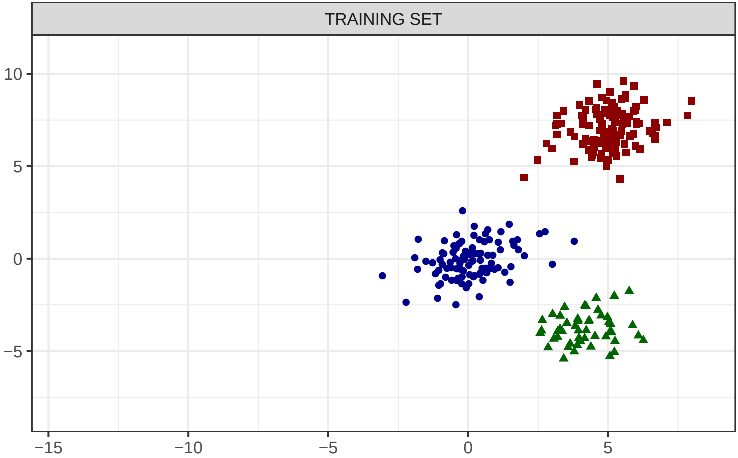
$\boldsymbol{\kappa} = Hp + H$ for every model, that is the p entries for $\boldsymbol{\mu}_h, h = G+1, \ldots, C$ and H mixing proportions

Simulated Experiment

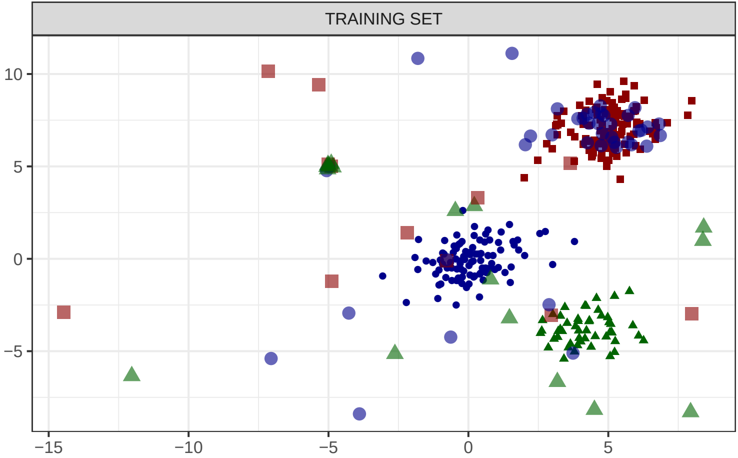
RAEDDA Model is employed for performing Supervised Learning in a Scenario that involves:

- Extra Classes in the Test Set
- Label Noise
- Possion Noise (both in Training and Test)
- Pointwise Contamination (both in Training and Test)

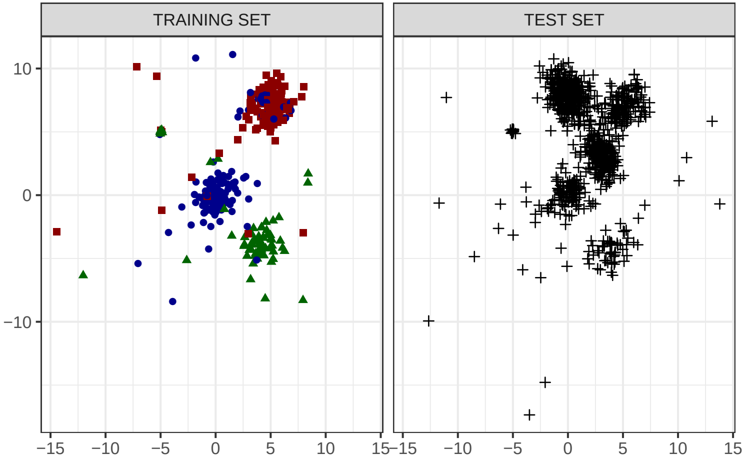
Uncontaminated Learning Set



Actual Learning Set



Classification Problem



Classification Result