Robust variable selection for model-based learning from adulterated samples

Andrea Cappozzo

University of Milano - Bicocca

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

12th International Conference of the ERCIM WG on Computational and Methodological Statistics; London, December 15

Outline

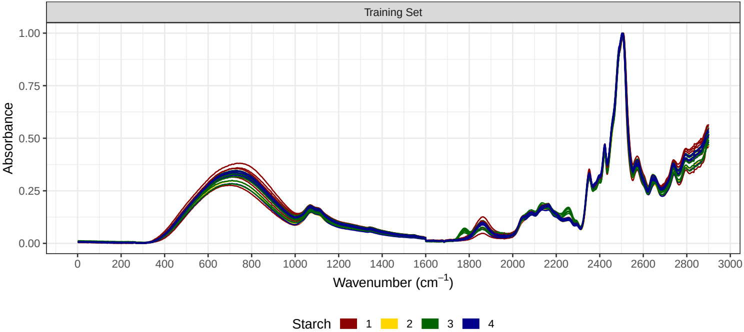
1. Chemometric contest
2. Feature selection in classification
3. Robust model-based Discriminant Analysis
4. Robust variable selection
    - Stepwise greedy-forward approach via TBIC
    - ML subset selector approach
5. Starches discrimination
6. Open Problems and Future Research

Chemometric contest

- MIR spectra of starches of four different classes (Fernández Pierna and Dardenne 2007)
- $P = 2901$ absorbance measurements for each sample
- Training and test sets of $N = 215$ and $M = 43$ units, respectively
- Adulterated samples (more details later!)

Chemometric contest

- MIR spectra of starches of four different classes (Fernández Pierna and Dardenne 2007)
- $P = 2901$ absorbance measurements for each sample
- Training and test sets of $N = 215$ and $M = 43$ units, respectively
- Adulterated samples (more details later!)



Motivating problem

Classification framework:

- High dimensional ($P = 2901$)
- Contaminated units (label noise and modifications)

Motivating problem

Classification framework:

- High dimensional ($P = 2901$)
- Contaminated units (label noise and modifications)

Expected output:

- High accuracy
- Anomaly detection
- Interpretable solution

Motivating problem

Classification framework:

- High dimensional ($P = 2901$)
- Contaminated units (label noise and modifications)

Expected output:

- High accuracy
- Anomaly detection
- Interpretable solution

Model-based method with variable selection would be optimal, but attribute and class noise can heavily damage the performance of standard methods (Zhu and Wu 2004)!

Variable selection in classification

The detection of *p* relevant features (out of $P >> p$) is particularly desirable as (McLachlan 1992):

Variable selection in classification

The detection of p relevant features (out of $P >> p$) is particularly desirable as (McLachlan 1992):

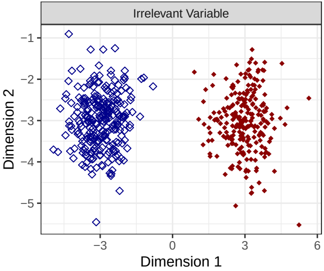
- it simplifies parameters estimation and interpretation
- it avoids loss on predictive power
- it leads to cost reduction on future data collection
- it mitigates the *curse of dimensionality* (Bellman 1957) in model-based methods
- for MIR spectra, adjacent wavelengths are often correlated and virtually contain the same information (Indahl and Næs 2004)

Variables role in DA

- *Relevant variables*: their distribution directly depends on the class membership
- *Irrelevant or noisy variables*: their distribution is completely independent from the group structure
- *Redundant variables*: their distribution is conditionally independent on the class membership, given the relevant features

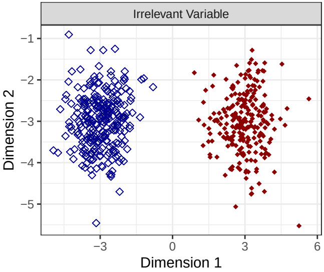
Variables role in DA

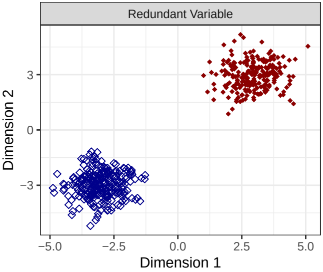
- *Relevant variables*: their distribution directly depends on the class membership
- *Irrelevant or noisy variables*: their distribution is completely independent from the group structure
- *Redundant variables*: their distribution is conditionally independent on the class membership, given the relevant features



Variables role in DA

- *Relevant variables*: their distribution directly depends on the class membership
- *Irrelevant or noisy variables*: their distribution is completely independent from the group structure
- *Redundant variables*: their distribution is conditionally independent on the class membership, given the relevant features





Robust Model-Based Classification

- A complete set of N learning observations:

$$(\mathbf{x}, \mathbf{l}) = \left\{ (\mathbf{x}_1, \mathbf{l}_1), \dots, (\mathbf{x}_N, \mathbf{l}_N); \mathbf{x}_n \in \mathbb{R}^P, l_n \in \{1, \dots, G\} \right\}$$

$\mathbf{x}_n$ is a P-dimensional predictor and $\mathbf{l}_n$ its associated label

- Data generating process for **genuine** observations

$$\mathcal{G} \sim Mult_G(1; \tau_1, \dots, \tau_G) \quad \mathcal{X} | \mathcal{G} = g \sim \mathcal{N}_P(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$$

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

  - $\phi(\cdot; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$ multivariate normal density distribution
  - τ_g prior probability of the gth class
  - $\boldsymbol{\Sigma}_g = \lambda_g \boldsymbol{D}_g \boldsymbol{A}_g \boldsymbol{D}_g'$ (Bensmail and Celeux 1996)

Robust Model-Based Classification

REDDA protects the estimates against label noise and outliers defining a suitable trimmed mixture log-likelihood (Cappozzo, Greselin, and Murphy 2019)

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

- $\zeta(\cdot)$ 0-1 trimming indicator function
- α_l *labelled trimming level*: $\sum_{n=1}^{N} \zeta(\mathbf{x}_n) = \lceil N(1-\alpha_l) \rceil$
- (1) maximized via a generalization of the FastMCD algorithm (Rousseeuw and Driessen 1999)
- *Concentration step* discards $\lfloor N\alpha_l \rfloor$% units with lowest:

$$f(\mathbf{x}_n | l_{ng} = 1; \hat{\boldsymbol{\mu}}_g, \hat{\boldsymbol{\Sigma}}_g) = \phi\left(\mathbf{x}_n; \hat{\boldsymbol{\mu}}_g, \hat{\boldsymbol{\Sigma}}_g\right) \qquad n = 1, \ldots, N.$$

Robust variable selection

Two proposals for robust variable selection in model-based classification

Robust variable selection

Two proposals for robust variable selection in model-based classification

- Robust stepwise greedy-forward approach via TBIC
  - Robust classification rule built in a step-wise manner
  - TBIC used for model comparison
  - Automatic selection of the relevant subset size

Robust variable selection

Two proposals for robust variable selection in model-based classification

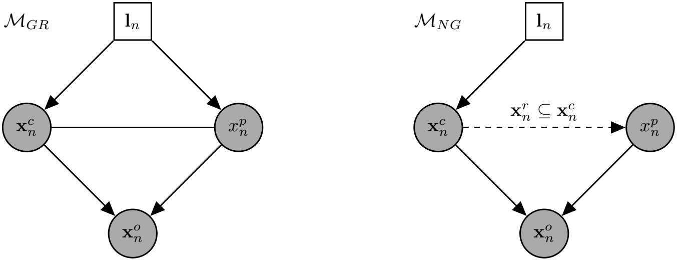
- Robust stepwise greedy-forward approach via TBIC
  - Robust classification rule built in a step-wise manner
  - TBIC used for model comparison
  - Automatic selection of the relevant subset size

- ML subset selector approach
  - Based on MLE theory and irrelevance in Gaussian mixtures
  - Relevant subset as a parameter to be estimated via ML
  - Relevant subset size is a-priori specified

Robust stepwise via TBIC

At each step of the algorithm, the learning observations are partitioned as $\mathbf{x}_n = (\mathbf{x}_n^c, x_n^p, \mathbf{x}_n^o)$ (Raftery and Dean 2006):

- $\mathbf{x}_n^c$ the variables currently included in the model
- x_n^p the variable proposed for inclusion
- $\mathbf{x}_n^o$ the remaining variables



Grouping

$p(\mathbf{x}_n^c, x_n^p|\mathbf{l}_n)p(\mathbf{x}_n^o|x_n^p, \mathbf{x}_n^c)$

No Grouping

$p(\mathbf{x}_n^c|\mathbf{l}_n)p(x_n^p|\mathbf{x}_n^r \subseteq \mathbf{x}_n^c)p(\mathbf{x}_n^o|x_n^p, \mathbf{x}_n^c)$

Robust stepwise via TBIC

Model comparison is carried out employing a robust approximation to the Bayes Factor (Kass and Raftery 1995):

$$\mathcal{B}_{GR,NG} = \frac{p(\mathbf{x}_n|\mathcal{M}_{GR})}{p(\mathbf{x}_n|\mathcal{M}_{NG})} = \frac{\int p(\mathbf{x}_n|\boldsymbol{\theta}_{GR}, \mathcal{M}_{GR}) p(\boldsymbol{\theta}_{GR}|\mathcal{M}_{GR}) d\boldsymbol{\theta}_{GR}}{\int p(\mathbf{x}_n|\boldsymbol{\theta}_{NG}, \mathcal{M}_{NG}) p(\boldsymbol{\theta}_{NG}|\mathcal{M}_{NG}) d\boldsymbol{\theta}_{NG}}$$

Robust stepwise via TBIC

Model comparison is carried out employing a robust approximation to the Bayes Factor (Kass and Raftery 1995):

$$\mathcal{B}_{GR,NG} = \frac{p(\mathbf{x}_n|\mathcal{M}_{GR})}{p(\mathbf{x}_n|\mathcal{M}_{NG})} = \frac{\int p(\mathbf{x}_n|\boldsymbol{\theta}_{GR}, \mathcal{M}_{GR}) p(\boldsymbol{\theta}_{GR}|\mathcal{M}_{GR}) d\boldsymbol{\theta}_{GR}}{\int p(\mathbf{x}_n|\boldsymbol{\theta}_{NG}, \mathcal{M}_{NG}) p(\boldsymbol{\theta}_{NG}|\mathcal{M}_{NG}) d\boldsymbol{\theta}_{NG}}$$

Trimmed BIC (Neykov et al. 2007), is employed as a robust proxy for the integrated likelihoods

$$2\log\left(\mathcal{B}_{GR,NG}\right) \approx TBIC(Grouping) - TBIC(No\ Grouping) \qquad (2)$$

Variable x_n^p with a positive difference in (2) is a candidate for being added (removed) to (from) the model

Robust stepwise via TBIC

$$TBIC(GR) = \underbrace{2\sum_{n=1}^{N}\zeta(\mathbf{x}_n^c, x_n^p)\sum_{g=1}^{G}\mathfrak{l}_{ng}\log\left(\hat{\tau}_g^{cp}\phi(\mathbf{x}_n^c, x_n^p; \hat{\boldsymbol{\mu}}_g^{cp}, \hat{\boldsymbol{\Sigma}}_g^{cp})\right)}_{2\times\text{trimmed log maximized likelihood of } p(\mathbf{x}_n^c, x_n^p, \mathbf{l}_n)} +$$

$$- v^{cp} log(N^*)$$

$$TBIC(NG) = \underbrace{2\sum_{n=1}^{N}\iota(\mathbf{x}_n^c, x_n^p)\sum_{g=1}^{G}\mathfrak{l}_{ng}\log\left(\hat{\tau}_g^{c}\phi(\mathbf{x}_n^c; \hat{\boldsymbol{\mu}}_g^{c}, \hat{\boldsymbol{\Sigma}}_g^{c})\right)}_{2\times\text{trimmed log maximized likelihood of } p(\mathbf{x}_n^c, \mathbf{l}_n)} - v^c log(N^*) +$$

$$\underbrace{+2\sum_{n=1}^{N}\iota(\mathbf{x}_n^c, x_n^p)\log\left[\phi\left(x_n^p; \hat{\alpha} + \hat{\boldsymbol{\beta}}'\mathbf{x}_n^r, \hat{\sigma}^2\right)\right]}_{2\times\text{trimmed log maximized likelihood of } p(x_n^p|\mathbf{x}_n^r\subseteq\mathbf{x}_n^c)} - v^p log(N^*).$$

ML subset selector

A model for the entire P-dimensional space is built:

- $F \subseteq 1, \ldots, P$ set of relevant variables, $|F| = p$
- $E = \bar{F}$ set of irrelevant variables, $|E| = P - p$

Exploiting the theory for the multivariate Gaussian under irrelevance (Ritter 2014)

$$\ell_{trim}(\boldsymbol{\tau}, \boldsymbol{\mu}_F, \boldsymbol{\Sigma}_F, \boldsymbol{G}_{E|F}, \boldsymbol{\mu}_{E|F}, \boldsymbol{\Sigma}_{E|F} | \mathbf{X}, \mathbf{l}) =$$

$$= \sum_{n=1}^{N} \zeta(\mathbf{x}_n) \left(\sum_{g=1}^{G} l_{ng} \log \left[\tau_g \phi(\mathbf{x}_{n,F}; \boldsymbol{\mu}_{g,F}, \boldsymbol{\Sigma}_{g,F}) \right] + \right.$$

$$\left. + \log \left[\phi(\mathbf{x}_{n,E} - \boldsymbol{G}_{E|F} \mathbf{x}_{n,F}; \boldsymbol{\mu}_{E|F}, \boldsymbol{\Sigma}_{E|F}) \right] \right)$$

$$\boldsymbol{\mu}_{E|F} = \boldsymbol{\mu}_E - \boldsymbol{G}_{E|F} \boldsymbol{\mu}_F, \quad \boldsymbol{\Sigma}_{E|F} = \boldsymbol{\Sigma}_E - G_{E|F} \boldsymbol{\Sigma}_{F,E}, \quad \boldsymbol{G}_{E|F} = \boldsymbol{\Sigma}_{E,F} \boldsymbol{\Sigma}_F^{-1}$$

ML subset selector

1. *Robust Initialization:*
   - Draw a random $(P+1)$-subset for each class g, $g = 1, \ldots, G$
   - $\zeta(\mathbf{x}_n) = 1$ if $\mathbf{x}_n$ belongs to any of such G subsets, otherwise $\zeta(\mathbf{x}_n) = 0$ (different strategy if $P >> p$)
2. *M-step:*

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

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

$$\hat{\boldsymbol{\mu}} = \frac{\sum_{n=1}^{N} \zeta(\mathbf{x}_n) \mathbf{x}_n}{\lceil N(1-\alpha_l) \rceil}.$$

$\hat{\boldsymbol{\Sigma}}_g$ and $\hat{\boldsymbol{\Sigma}}$ according to (Bensmail and Celeux 1996)

ML subset selector

3. *S-step:* Minimize the difference

$$h(F) = \sum_{g=1}^{G} \hat{\tau}_g \log \det \hat{\boldsymbol{\Sigma}}_{g,F} - \log \det \hat{\boldsymbol{\Sigma}}_F$$

w.r.t. the subset $\hat{F} \subseteq 1, \ldots, P$

4. *T-step:*

$$\hat{\boldsymbol{G}}_{\hat{E}|\hat{F}} = \hat{\boldsymbol{\Sigma}}_{\hat{E},\hat{F}} \hat{\boldsymbol{\Sigma}}_{\hat{F}}^{-1}, \;\; \hat{\boldsymbol{\mu}}_{\hat{E}|\hat{F}} = \hat{\boldsymbol{\mu}}_{\hat{E}} - \hat{\boldsymbol{G}}_{\hat{E}|\hat{F}} \hat{\boldsymbol{\mu}}_{\hat{F}}, \;\; \hat{\boldsymbol{\Sigma}}_{\hat{E}|\hat{F}} = \hat{\boldsymbol{\Sigma}}_{\hat{E}} - \hat{\boldsymbol{\Sigma}}_{\hat{E},\hat{F}} \hat{\boldsymbol{\Sigma}}_{\hat{F}}^{-1} \hat{\boldsymbol{\Sigma}}_{\hat{F},\hat{E}}$$

Update the value of $\zeta(\cdot)$, discarding $\lfloor N\alpha_l \rfloor$% units with lowest:

$$\sum_{g=1}^{G} l_{ng} \log \left[\hat{\tau}_g \phi(\mathbf{x}_{n,\hat{F}}; \hat{\boldsymbol{\mu}}_{g,\hat{F}}, \hat{\boldsymbol{\Sigma}}_{g,\hat{F}}) \right] + \log \left[\phi \left(\mathbf{x}_{n,\hat{E}} - \hat{\boldsymbol{G}}_{\hat{E}|\hat{F}} \mathbf{x}_{n,\hat{F}}; \hat{\boldsymbol{\mu}}_{\hat{E}|\hat{F}}, \hat{\boldsymbol{\Sigma}}_{\hat{E}|\hat{F}} \right) \right]$$

5. Iterate $2 - 4$ until $\zeta(\cdot)$ does not change.

Data adulteration

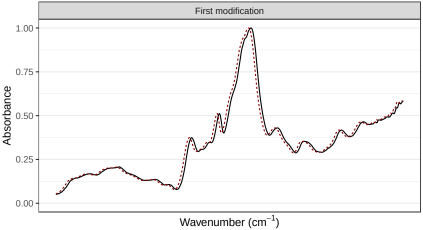
- Training set: 4 units with label noise

Data adulteration

- Training set: 4 units with label noise
- Test set: 4 modified units

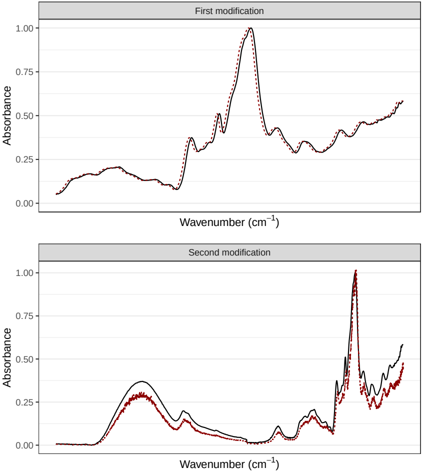
Data adulteration

- Training set: 4 units with label noise
- Test set: 4 modified units



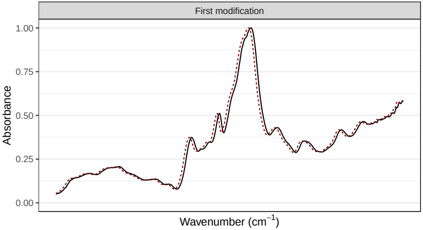
Data adulteration

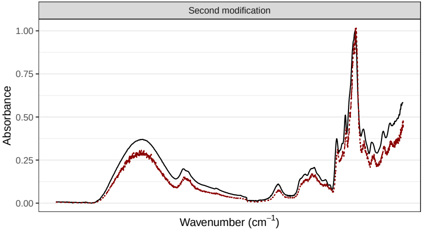
- Training set: 4 units with label noise
- Test set: 4 modified units

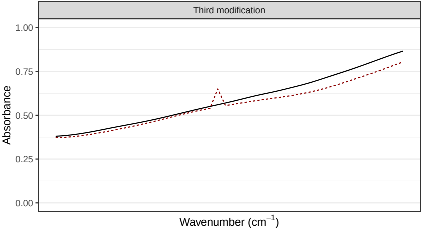


Data adulteration

- Training set: 4 units with label noise
- Test set: 4 modified units

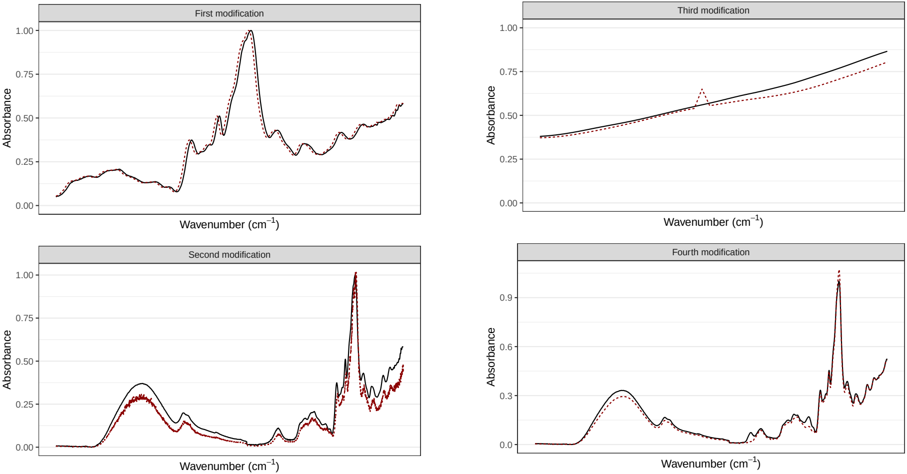






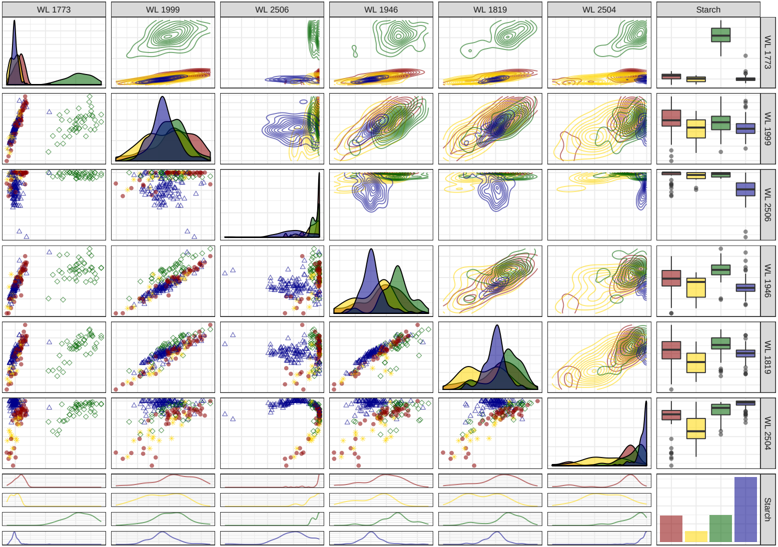
Data adulteration

- Training set: 4 units with label noise
- Test set: 4 modified units



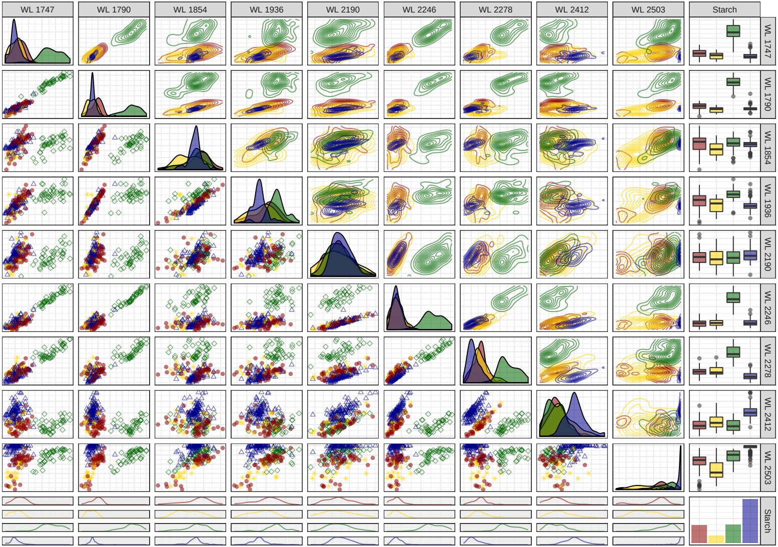
Results: Robust stepwise via TBIC

Selected WL: 1773, 1999, 2506, 1946, 1819, 2504



Results: ML subset selector

Selected WL: 1747, 1790, 1854, 1936, 2190, 2246, 2278, 2412, 2503



Results & adulteration detection

	REDDA (TBIC)	REDDA (ML subset)	SVM radial kernel	ROC+PLS+SVM
With outliers				
# correctly predicted	34	36	32	33
% correctly predicted	0.791	0.837	0.744	0.767
Without outliers				
#correctly predicted	32	34	31	31
% correctly predicted	0.821	0.872	0.795	0.795

Results & adulteration detection

	REDDA (TBIC)	REDDA (ML subset)	SVM radial kernel	ROC+PLS+SVM
With outliers				
# correctly predicted	34	36	32	33
% correctly predicted	0.791	0.837	0.744	0.767
Without outliers				
#correctly predicted	32	34	31	31
% correctly predicted	0.821	0.872	0.795	0.795

- Adulteration detection is performed considering:

$$\hat{p}(\mathbf{y}_{m,\hat{F}}; \hat{\tau}, \hat{\boldsymbol{\mu}}_{\hat{F}}, \hat{\boldsymbol{\Sigma}}_{\hat{F}}) = \sum_{g=1}^{G} \hat{\tau}_g \phi\left(\mathbf{y}_{m,\hat{F}}; \hat{\boldsymbol{\mu}}_{g,\hat{F}}, \hat{\boldsymbol{\Sigma}}_{g,\hat{F}}\right) \tag{3}$$

3 out of the 4 modified units possess lowest values of (3).

Conclusions

We have introduced two wrapper variable selection methods, resistant to outliers and label noise

- Robust stepwise via TBIC: robust model-based classifier within a greedy-forward algorithm
- ML subset selector: the subset of relevant variables is a parameter to be estimated

Future research direction

- Extension to the adaptive framework, where unobserved classes in the test set need to be discovered
- Development of dedicated R package

References

- Bellman, Richard (1957). *Dynamic Programming*. Rand Corporation research study. Princeton University Press.
- 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.
- 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`.
- Fernández Pierna, Juan Antonio and Pierre Dardenne (2007). "Chemometric contest at 'Chimiométrie 2005': A discrimination study". In: *Chemometrics and Intelligent Laboratory Systems* 86.2, pp. 219–223.
- Indahl, Ulf and Tormod Næs (2004). "A variable selection strategy for supervised classification with continuous spectroscopic data". In: *Journal of Chemometrics* 18.2, pp. 53–61.
- Kass, Robert E. and Adrian E. Raftery (1995). "Bayes Factors". In: *Journal of the American Statistical Association* 90.430, p. 773.
- McLachlan, Geoffrey J. (1992). *Discriminant Analysis and Statistical Pattern Recognition*. Vol. 544. Wiley Series in Probability and Statistics. Hoboken, NJ, USA: John Wiley & Sons, Inc.
- Neykov, N. et al. (2007). "Robust fitting of mixtures using the trimmed likelihood estimator". In: *Computational Statistics & Data Analysis* 52.1, pp. 299–308.
- Raftery, Adrian E and Nema Dean (2006). "Variable selection for model-based clustering". In: *Journal of the American Statistical Association* 101.473, pp. 168–178.
- Ritter, Gunter (2014). *Robust Cluster Analysis and Variable Selection*. Chapman and Hall/CRC.
- Rousseeuw, Peter J. and Katrien Van Driessen (1999). "A fast algorithm for the minimum covariance determinant estimator". In: *Technometrics* 41.3, pp. 212–223.
- Zhu, Xingquan and Xindong Wu (2004). "Class noise vs. attribute noise: A quantitative study". In: *Artificial Intelligence Review* 22.3, pp. 177–210.

Thank You!