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

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

- It simplifies parameter estimation and interpretation.
- It avoids loss in 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).
Variable selection in classification

The detection of $p$ relevant features (out of $P \gg p$) 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:
  \[(x, l) = \{ (x_1, l_1), \ldots, (x_N, l_N) \mid x_n \in \mathbb{R}^P, l_n \in \{1, \ldots, G\}\}\]
  $x_n$ is a $P$-dimensional predictor and $l_n$ its associated label

- Data generating process for genuine observations:
  \[G \sim Mult_G(1; \tau_1, \ldots, \tau_G)\]
  \[x | G = g \sim \mathcal{N}_P(\mu_g, \Sigma_g)\]

\[p(x_n, l_n; \theta) = p(l_n; \tau)p(x_n | l_n = g; \mu_g, \Sigma_g) = \prod_{g=1}^{G} [\tau_g \phi(x_n; \mu_g, \Sigma_g)]^{l_{ng}}\]

- $\phi(\cdot; \mu_g, \Sigma_g)$ multivariate normal density distribution
- $\tau_g$ prior probability of the $g$th class
- $\Sigma_g = \lambda_g D_g A_g D'_g$ (Bensmail and Celeux 1996)
REDDA protects the estimates against label noise and outliers defining a suitable trimmed mixture log-likelihood (Cappozzo, Greselin, and Murphy 2019)

\[
\ell_{\text{trim}}(\tau, \mu, \Sigma|X, l) = \sum_{n=1}^{N} \zeta(x_n) \sum_{g=1}^{G} \log \left( \tau_g \phi(x_n; \mu_g, \Sigma_g) \right)
\]

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

\[
f(x_n|l_{ng} = 1; \hat{\mu}_g, \hat{\Sigma}_g) = \phi \left( x_n; \hat{\mu}_g, \hat{\Sigma}_g \right) \quad n = 1, \ldots, N.
\]
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 \( \mathbf{x}_n = (\mathbf{x}^c_n, \mathbf{x}^p_n, \mathbf{x}^o_n) \) (Raftery and Dean 2006):

\[
\begin{align*}
\mathbf{x}^c_n & \text{ the variables currently included in the model} \\
\mathbf{x}^p_n & \text{ the variable proposed for inclusion} \\
\mathbf{x}^o_n & \text{ the remaining variables}
\end{align*}
\]

Grouping

\[
p(\mathbf{x}^c_n, \mathbf{x}^p_n | \mathbf{l}_n) p(\mathbf{x}^o_n | \mathbf{x}^p_n, \mathbf{x}^c_n)
\]

No Grouping

\[
p(\mathbf{x}^c_n | \mathbf{l}_n) p(\mathbf{x}^p_n | \mathbf{x}^p_n \subseteq \mathbf{x}^c_n) p(\mathbf{x}^o_n | \mathbf{x}^p_n, \mathbf{x}^c_n)
\]
Model comparison is carried out employing a robust approximation to the Bayes Factor (Kass and Raftery 1995):

\[ \mathcal{B}_{GR,NG} = \frac{p(x_n|M_{GR})}{p(x_n|M_{NG})} = \frac{\int p(x_n|\theta_{GR},M_{GR})p(\theta_{GR}|M_{GR})d\theta_{GR}}{\int p(x_n|\theta_{NG},M_{NG})p(\theta_{NG}|M_{NG})d\theta_{NG}} \]
Robust stepwise via TBIC

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

\[
B_{GR,NG} = \frac{p(x_n|\mathcal{M}_{GR})}{p(x_n|\mathcal{M}_{NG})} = \frac{\int p(x_n|\theta_{GR}, \mathcal{M}_{GR})p(\theta_{GR}|\mathcal{M}_{GR})d\theta_{GR}}{\int p(x_n|\theta_{NG}, \mathcal{M}_{NG})p(\theta_{NG}|\mathcal{M}_{NG})d\theta_{NG}}
\]

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

\[
2 \log (B_{GR,NG}) \approx TBIC(\text{Grouping}) - TBIC(\text{No Grouping}) \quad (2)
\]

Variable \(x_{i}^{p}\) with a positive difference in (2) is a candidate for being added (removed) to (from) the model
Robust stepwise via TBIC

\[ TBIC(\text{GR}) = 2 \sum_{n=1}^{N} \sum_{g=1}^{G} \eta_{ng} \log \left( \hat{\tau}_g^{cp} \phi(\mathbf{x}_n^c, \mathbf{x}_n^p; \mu_g^{cp}, \Sigma_g^{cp}) \right) + \]

2 trimmed log maximized likelihood of \( \eta(\mathbf{x}_n^c, \mathbf{x}_n^p) \)

\[ - \nu^{cp} \log(N^*) \]

\[ TBIC(\text{NG}) = 2 \sum_{n=1}^{N} \sum_{g=1}^{G} \eta_{ng} \log \left( \hat{\tau}_g^{c} \phi(\mathbf{x}_n^c; \mu_g^{c}, \Sigma_g^{c}) \right) - \nu^{c} \log(N^*) + \]

2 trimmed log maximized likelihood of \( \eta(\mathbf{x}_n^c) \)

\[ + 2 \sum_{n=1}^{N} \eta(\mathbf{x}_n^c, \mathbf{x}_n^p) \log \left[ \phi \left( \mathbf{x}_n^p; \hat{\alpha} + \hat{\beta}' \mathbf{x}_n^r, \hat{\sigma}^2 \right) \right] - \nu^{p} \log(N^*) . \]

2 trimmed log maximized likelihood of \( \eta(\mathbf{x}_n^c | \mathbf{x}_n^p \subseteq \mathbf{x}_n^c) \)
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)

\[
\ell_{\text{trim}}(\tau, \mu_F, \Sigma_F, G_{E|F}, \mu_{E|F}, \Sigma_{E|F}|x, 1) = \\
= \sum_{n=1}^{N} \zeta(x_n) \left( \sum_{g=1}^{G} l_{ng} \log \left[ \tau_g \phi(x_{n,F}; \mu_{g,F}, \Sigma_{g,F}) \right] + \\
+ \log \left[ \phi(x_{n,E} - G_{E|F}x_{n,F}; \mu_{E|F}, \Sigma_{E|F}) \right] \right)
\]

\[
\mu_{E|F} = \mu_E - G_{E|F} \mu_F, \quad \Sigma_{E|F} = \Sigma_E - G_{E|F} \Sigma_{F,E}, \quad G_{E|F} = \Sigma_{E,F} \Sigma_F^{-1}
\]
1. Robust Initialization:
   - Draw a random \((P + 1)\)-subset for each class \(g\), \(g = 1, \ldots, G\)
   - \(\zeta(x_n) = 1\) if \(x_n\) belongs to any of such \(G\) subsets, otherwise \(\zeta(x_n) = 0\) (different strategy if \(P >> p\))

2. M-step:
   
   \[
   \hat{r}_g = \frac{\sum_{n=1}^{N} \zeta(x_n) l_{ng}}{\lfloor N(1 - \alpha_l) \rfloor} \quad g = 1, \ldots, G
   \]

   \[
   \hat{\mu}_g = \frac{\sum_{n=1}^{N} \zeta(x_n) l_{ng} x_n}{\sum_{n=1}^{N} \zeta(x_n) l_{ng}} \quad g = 1, \ldots, G.
   \]

   \[
   \hat{\mu} = \frac{\sum_{n=1}^{N} \zeta(x_n) x_n}{\lfloor N(1 - \alpha_l) \rfloor}.
   \]

\(\hat{\Sigma}_g\) and \(\hat{\Sigma}\) according to (Bensmail and Celeux 1996)
3. **S-step**: Minimize the difference

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

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

4. **T-step**:

\[ \hat{G}_{\hat{E}|\hat{F}} = \hat{\Sigma}_{\hat{E},\hat{F}} \hat{\Sigma}_{\hat{F}}^{-1}, \quad \hat{\mu}_{\hat{E}|\hat{F}} = \hat{\mu}_{\hat{E}} - \hat{G}_{\hat{E}|\hat{F}} \hat{\mu}_{\hat{F}}, \quad \hat{\Sigma}_{\hat{E}|\hat{F}} = \hat{\Sigma}_{\hat{E}} - \hat{\Sigma}_{\hat{E},\hat{F}} \hat{\Sigma}_{\hat{F}}^{-1} \hat{\Sigma}_{\hat{F},\hat{E}} \]

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

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

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

![Graphs showing absorbance vs. wavenumber for different modifications](image)
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

<table>
<thead>
<tr>
<th></th>
<th>REDDA (TBIC)</th>
<th>REDDA (ML subset)</th>
<th>SVM radial kernel</th>
<th>ROC+PLS+SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>With outliers</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td># correctly predicted</td>
<td>34</td>
<td>36</td>
<td>32</td>
<td>33</td>
</tr>
<tr>
<td>% correctly predicted</td>
<td>0.791</td>
<td>0.837</td>
<td>0.744</td>
<td>0.767</td>
</tr>
<tr>
<td><strong>Without outliers</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td># correctly predicted</td>
<td>32</td>
<td>34</td>
<td>31</td>
<td>31</td>
</tr>
<tr>
<td>% correctly predicted</td>
<td>0.821</td>
<td>0.872</td>
<td>0.795</td>
<td>0.795</td>
</tr>
</tbody>
</table>

With outliers:

\[
\mathbb{p}(y_{m}; \mathbb{F}; \mathbb{G}; \mathbb{F}; \mathbb{G}) = \mathbb{G} \sum_{g=1}^{G} \phi(y_{m}; \mathbb{F}; g; \mathbb{F}; g; \mathbb{F})
\]  

3 out of the 4 modified units possess lowest values of (3).
Results & adulteration detection

<table>
<thead>
<tr>
<th></th>
<th>REDDA (TBIC)</th>
<th>REDDA (ML subset)</th>
<th>SVM radial kernel</th>
<th>ROC+PLS+SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>With outliers</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td># correctly predicted</td>
<td>34</td>
<td>36</td>
<td>32</td>
<td>33</td>
</tr>
<tr>
<td>% correctly predicted</td>
<td>0.791</td>
<td>0.837</td>
<td>0.744</td>
<td>0.767</td>
</tr>
<tr>
<td>Without outliers</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td># correctly predicted</td>
<td>32</td>
<td>34</td>
<td>31</td>
<td>31</td>
</tr>
<tr>
<td>% correctly predicted</td>
<td>0.821</td>
<td>0.872</td>
<td>0.795</td>
<td>0.795</td>
</tr>
</tbody>
</table>

Adulteration detection is performed considering:

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

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


Thank You!