Cellwise robust regularized discriminant analysis

Stéphanie Aerts (University of Liège)
Ines Wilms (Cornell University, KU Leuven)

ICORS, July 2018

Discriminant analysis

$\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ of dimension p,
splitted into K groups, each having n_k observations

Goal : Classify new data $\boldsymbol{x}$

π_k prior probability
$N_p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ conditional distribution

Discriminant analysis

$\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ of dimension p,
splitted into K groups, each having n_k observations

Goal : Classify new data $\boldsymbol{x}$

π_k prior probability
$N_p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ conditional distribution

Quadratic discriminant analysis (QDA) :

$$\max_k \left(-(\boldsymbol{x} - \boldsymbol{\mu}_k)^T \boldsymbol{\Theta}_k (\boldsymbol{x} - \boldsymbol{\mu}_k) + \log(\det \boldsymbol{\Theta}_k) + 2 \log \pi_k\right)$$

where $\boldsymbol{\Theta}_k := \boldsymbol{\Sigma}_k^{-1}$

Linear discriminant analysis (LDA) :
Homoscedasticity : $\boldsymbol{\Theta}_k = \boldsymbol{\Theta} \quad \forall k$

Discriminant analysis

In practice, the parameters $\boldsymbol{\mu}_k, \boldsymbol{\Theta}_k$ or $\boldsymbol{\Theta}$ are estimated by

the sample means $\bar{\boldsymbol{x}}_k$

the inverse of the sample covariance matrices $\widehat{\boldsymbol{\Sigma}}_k$

the inverse of the sample pooled covariance matrix $\widehat{\boldsymbol{\Sigma}}_{\mathrm{pool}} = \frac{1}{N-K}\sum_{k=1}^{K} n_k \widehat{\boldsymbol{\Sigma}}_k$

Example - Phoneme dataset

Hastie et al., 2009

$N = 1717$ records of a male voice
$K = 2$ phonemes : *aa* (like in *barn*) or *ao* (like in *more*)
$p = 256$: log intensity of the sound across 256 frequencies

Correct classification performance

s-LDA	s-QDA
77.7	62.4

Example - Phoneme dataset

Hastie et al., 2009

$N = 1717$ records of a male voice
$K = 2$ phonemes : *aa* (like in *barn*) or *ao* (like in *more*)
$p = 256$: log intensity of the sound across 256 frequencies

Correct classification performance

s-LDA	s-QDA
77.7	62.4

✗ $\widehat{\Sigma}_k^{-1}$ inaccurate when p/n_k is large, not computable when $p > n_k$
$\widehat{\Sigma}_{\text{pool}}^{-1}$ inaccurate when p/N is large, not computable when $p > N$

Objectives

Propose a family of discriminant methods that, unlike the classical approach, are

1. computable and accurate in high dimension
2. cover the path between LDA and QDA
3. robust against cellwise outliers

1. Computable in high dimension

Graphical Lasso QDA (Xu et al., 2014)

Step 1 : Compute the sample means $\bar{\boldsymbol{x}}_k$ and covariance matrices $\widehat{\boldsymbol{\Sigma}}_k$

Step 2 : *Graphical Lasso* (Friedman et al, 2008) to estimate $\boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_K$:

$$\max_{\boldsymbol{\Theta}_k} n_k \log\det(\boldsymbol{\Theta}_k) - n_k \mathrm{tr}\left(\boldsymbol{\Theta}_k \widehat{\boldsymbol{\Sigma}}_k\right) - \lambda_1 \sum_{i \neq j} |\theta_{k,ij}|$$

Step 3 : Plug $\bar{\boldsymbol{x}}_1, \ldots, \bar{\boldsymbol{x}}_K$ and $\widehat{\boldsymbol{\Theta}}_1, \ldots, \widehat{\boldsymbol{\Theta}}_K$ into the quadratic rule

Note : Use pooled covariance matrix for Graphical Lasso LDA

1. Computable in high dimension

Graphical Lasso QDA (Xu et al., 2014)

Step 1 : Compute the sample means $\bar{\boldsymbol{x}}_k$ and covariance matrices $\widehat{\boldsymbol{\Sigma}}_k$

Step 2 : *Graphical Lasso* (Friedman et al, 2008) to estimate $\boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_K$:

$$\max_{\boldsymbol{\Theta}_k} n_k \log\det(\boldsymbol{\Theta}_k) - n_k \mathrm{tr}\left(\boldsymbol{\Theta}_k \widehat{\boldsymbol{\Sigma}}_k\right) - \lambda_1 \sum_{i \neq j} |\theta_{k,ij}|$$

Step 3 : Plug $\bar{\boldsymbol{x}}_1, \ldots, \bar{\boldsymbol{x}}_K$ and $\widehat{\boldsymbol{\Theta}}_1, \ldots, \widehat{\boldsymbol{\Theta}}_K$ into the quadratic rule

Note : Use pooled covariance matrix for Graphical Lasso LDA

✗ QDA does not exploit similarities
LDA ignores group specificities

2. Covering path between LDA and QDA

Joint Graphical Lasso DA (Price et al., 2015)

Step 1 : Compute the sample means $\bar{\boldsymbol{x}}_k$ and covariance matrices $\widehat{\boldsymbol{\Sigma}}_k$

Step 2 : *Joint Graphical Lasso* (Danaher et al, 2014) to estimate $\boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_K$:

$$\max_{\boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_K} \sum_{k=1}^{K} n_k \log \det(\boldsymbol{\Theta}_k) - n_k \mathrm{tr}(\boldsymbol{\Theta}_k \widehat{\boldsymbol{\Sigma}}_k) - \lambda_1 \sum_{k=1}^{K} \sum_{i \neq j} |\theta_{k,ij}| - \lambda_2 \sum_{k<k'} \sum_{i,j} |\theta_{k,ij} - \theta_{k',ij}|,$$

Step 3 : Plug $\widehat{\boldsymbol{\Theta}}_1, \ldots, \widehat{\boldsymbol{\Theta}}_K$ into the quadratic discriminant rule

2. Covering path between LDA and QDA

Joint Graphical Lasso DA (Price et al., 2015)

Step 1 : Compute the sample means $\bar{\mathbf{x}}_k$ and covariance matrices $\widehat{\boldsymbol{\Sigma}}_k$

Step 2 : *Joint Graphical Lasso* (Danaher et al, 2014) to estimate $\boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_K$:

$$\max_{\boldsymbol{\Theta}_1,\ldots,\boldsymbol{\Theta}_K} \sum_{k=1}^{K} n_k \log\det(\boldsymbol{\Theta}_k) - n_k \mathrm{tr}(\boldsymbol{\Theta}_k \widehat{\boldsymbol{\Sigma}}_k) - \lambda_1 \sum_{k=1}^{K} \sum_{i \neq j} |\theta_{k,ij}| - \lambda_2 \sum_{k<k'} \sum_{i,j} |\theta_{k,ij} - \theta_{k',ij}|,$$

Step 3 : Plug $\widehat{\boldsymbol{\Theta}}_1, \ldots, \widehat{\boldsymbol{\Theta}}_K$ into the quadratic discriminant rule

✗ Lack of robustness

3. Robustness against cellwise outliers

Robust Joint Graphical Lasso DA

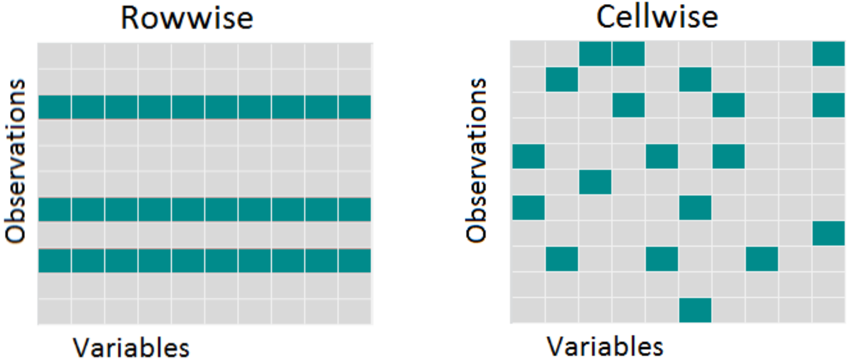
Step 1 : Compute *robust* $\boldsymbol{m}_k$ and $\boldsymbol{S}_k$ estimates

Step 2 : Joint Graphical Lasso to estimate $\boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_K$

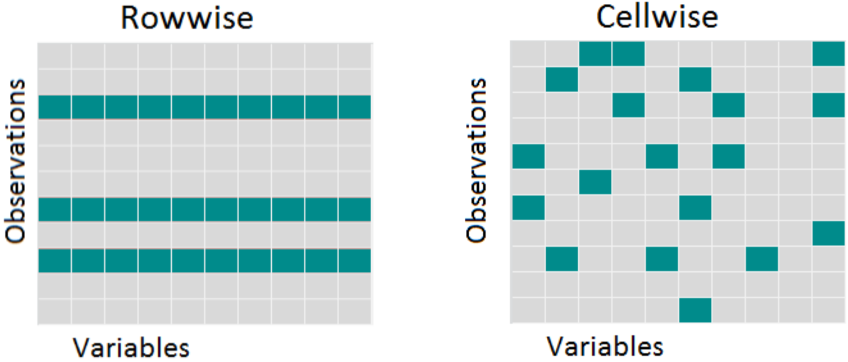
$$\max_{\boldsymbol{\Theta}_1,\ldots,\boldsymbol{\Theta}_K} \sum_{k=1}^{K} n_k \log\det(\boldsymbol{\Theta}_k) - n_k \mathrm{tr}(\boldsymbol{\Theta}_k \boldsymbol{S}_k) - \lambda_1 \sum_{k=1}^{K} \sum_{i \neq j} |\theta_{k,ij}| - \lambda_2 \sum_{k<k'} \sum_{i,j} |\theta_{k,ij} - \theta_{k',ij}|,$$

Step 3 : Plug $\boldsymbol{m}_1, \ldots, \boldsymbol{m}_K$ and $\widehat{\boldsymbol{\Theta}}_1, \ldots, \widehat{\boldsymbol{\Theta}}_K$ into the quadratic discriminant rule

Robust estimators



Robust estimators



$\boldsymbol{m}_k$: vector of marginal medians
$\boldsymbol{S}_k$: cellwise robust covariance matrices

Cellwise robust covariance estimators

$$\boldsymbol{S}_k = \begin{pmatrix} s_{11} & \dots & s_{1i} & \dots & s_{1p} \\ \vdots & & \vdots & & \vdots \\ s_{i1} & \dots & s_{ij} & \dots & s_{ip} \\ \vdots & & \vdots & & \vdots \\ s_{p1} & \dots & s_{pj} & \dots & s_{pp} \end{pmatrix}$$

$$s_{ij} = \widehat{\text{scale}}(X^i)\ \widehat{\text{scale}}(X^j)\ \widehat{\text{corr}}(X^i, X^j)$$

$\widehat{\text{scale}}(.)$: Q_n-estimator (Rousseeuw and Croux, 1993)
$\widehat{\text{corr}}(.,.)$: Kendall's correlation

$$\widehat{\text{corr}}_{\text{K}}(X^i, X^j) = \frac{2}{n(n-1)} \sum_{l<m} \text{sign}\left((x_l^i - x_m^i)(x_l^j - x_m^j)\right).$$

(see Croux and Öllerer, 2015 ; Tarr et al., 2016)

Simulation study

Setting

$K = 10$ groups
$n_k = 30$
$p = 30$
$M = 1000$ training and test sets

Classification Performance

Average percentage of correct classification

Estimation accuracy

Average Kullback-Leibler distance :

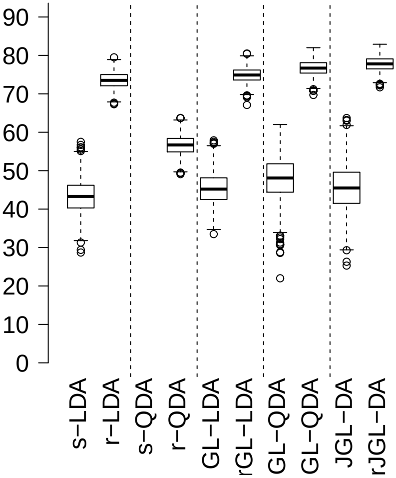
$$\mathrm{KL}(\widehat{\boldsymbol{\Theta}}_1, \ldots, \widehat{\boldsymbol{\Theta}}_K; \boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_K) = \left(\sum_{k=1}^{K} - \log \det(\widehat{\boldsymbol{\Theta}}_k \boldsymbol{\Theta}_k^{-1}) + \mathrm{tr}(\widehat{\boldsymbol{\Theta}}_k \boldsymbol{\Theta}_k^{-1}) \right) - Kp.$$

Uncontaminated scheme

Non-robust estimators		s-LDA	s-QDA	GL-LDA	GL-QDA	JGL-DA
$p = 30$	CC	77.7	NA	80.5	83.0	83.5
	KL	30.29	NA	21.87	40.41	5.03
Robust estimators		**r-LDA**	**r-QDA**	**rGL-LDA**	**rGL-QDA**	**rJGL-DA**
$p = 30$	CC	76.1	59.7	77.4	79.7	80.1
	KL	22.86	139.18	22.98	44.57	11.02

Contaminated scheme : 5% of cellwise contamination

Correct classification percentages, $p = 30$



Example 1 - Phoneme dataset

$N = 1717$
$K = 2$
$p = 256$

Correct classification performance

s-LDA	s-QDA	GL-LDA	GL-QDA	JGL-DA
77.7	62.4	81.4	74.9	78.4

r-LDA	r-QDA	rGL-LDA	rGL-QDA	rJGL-DA
81.1	74.7	81.7	76.0	76.7

$N_{\mathrm{train}} = 1030$, $N_{\mathrm{test}} = 687$, averaged over 10 splits

Conclusion

The proposed discriminant methods :

1. are computable in high dimension
2. cover the path between LDA and QDA
3. are robust against cellwise outliers
4. detect rowwise and cellwise outliers

Code publicly available

`http://feb.kuleuven.be/ines.wilms/software`

References

- S. Aerts, I. Wilms, Cellwise robust regularized discriminant analysis. *Statistical Analysis and Data Mining*, 10 : 436–447, 2017.

- C. Croux and V. Öllerer. Robust high-dimensional precision matrix estimation, *Modern Multivariate and Robust Methods*. Springer, 2015
- P. Danaher, P. Wang, and D. Witten. The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society, Series B*, 76 : 373–397, 2014.
- B. Price, C. Geyer, and A. Rothman. Ridge fusion in statistical learning. *Journal of Computational and Graphical Statistics*, 24(2) :439–454, 2015.
- P. Rousseeuw and C. Croux. Alternatives to the median absolute deviation. *Journal of the American Statistical Association*, 88(424) :1273–1283, 1993.
- G. Tarr, S. Müller, and N.C. Weber. Robust estimation of precision matrices under cellwise contamination. *Computational Statistics and Data Analysis*, 93 :404–420, 2016.
- B. Xu, K. Huang, King I., C. Liu, J. Sun, and N. Satoshi. Graphical lasso quadratic discriminant function and its application to character recognition. *Neurocomputing*, 129 : 33–40, 2014.