Robust Classifiers in Multivariate Statistics and Machine Learning

Jan Kalina

Institute of Computer Science CAS
& Institute of Information Theory and Automation CAS

Example: Credit approval

- Cases: clients
- Variables: personal information about credit cards and proprietors
  - Continuous
  - Categorial
- Aims:
  - Classification to two groups
  - Probability of belonging to a given group
- Logistic regression model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}, \quad i = 1, \ldots, n$$

- Possibly a large number of variables

Example: Cardiovascular genetic study

Center of Biomedical Informatics (2006–2011, prof. Zvárová)

Aim of the study:
Diagnostics of cardiovascular diseases.

Individuals (Municipal Hospital in Čáslav):

1. Acute myocardial infarction ($n = 98$)
2. Cerebrovascular stroke ($n = 46$)
3. Controls ($n = 169$)

Design:
Paired design based on risk factors (age, sex, hypertension, smoking).

Data:
Personal data. Clinical and biochemical measurements. Gene expressions across the whole genome from a sample of peripheral blood.

Example: Cardiovascular genetic study

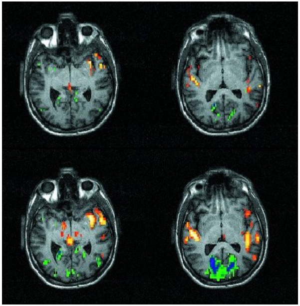
Table of gene expression values:

	Gene	24 patients with stroke				24 control persons			
		# 1	#2	···	#24	# 1	#2	···	#24
1	ADORA3	5.82	6.04	···	5.99	5.71	6.12	···	6.09
2	CPD	3.53	4.08	···	2.32	4.21	5.01	···	4.66
3	ECHDC3	2.50	2.71	···	3.17	2.99	3.52	···	3.01
4	VNN3	3.38	3.03	···	4.59	4.56	3.98	···	4.70
5	IL18RAP	4.03	4.91	···	5.81	5.12	5.01	···	5.23
6	ERLIN1	5.76	4.38	···	4.90	6.49	5.02	···	6.18
⋮	⋮	⋮	⋮	⋱	⋮	⋮	⋮	⋱	⋮
38 590	PHACTR1	5.21	4.99	···	5.06	5.15	5.53	···	5.20

High-dimensional data ($n < p$).

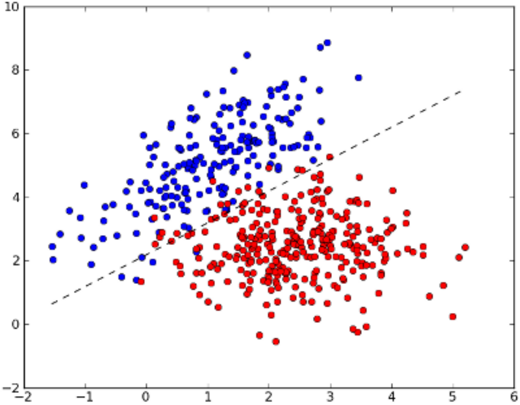
Example: Magnetic resonance of the brain

- Czech National Institute for Mental Health
- Aim: spontaneous brain activity (schizophrenia diagnostics)
- $n = 24$ patients
- $p = 4005$ brain features (correlations between brain parts)
- Classification task: resting state vs. a movie ($K = 2$)



NIMH
NATIONAL INSTITUTE OF MENTAL HEALTH

A classification task

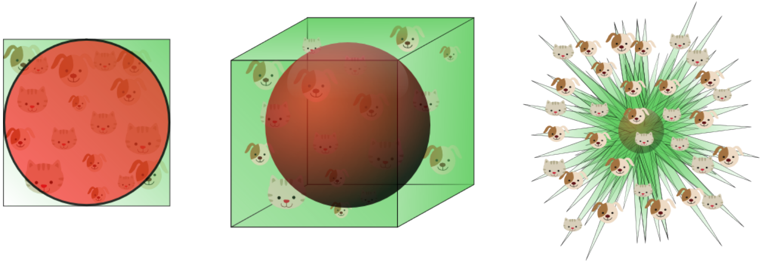


Classification into 2 groups (more generally: K groups).

Standard classification methods

- Linear discriminant analysis (LDA)
- Quadratic discriminant analysis (QDA)
- Logistic classification
- Support vector machines (SVM)
- Bayesian networks
- Classification trees/forests
- k-nearest neighbor
- Partial least squares

Curse of dimensionality



High-dimensional data

Examples of high-dimensional data in economics:

- Retail, advertising, insurance, online trade, portfolio optimization, customer analytics, ...

Analysis of high-dimensional data:

- Pre-processing
- Exploratory data analysis (EDA)
- Complexity reduction (dimensionality reduction)
- Some methods are unsuitable (e.g. neural networks)

Questions about dimensionality reduction:

- Is dimensionality reduction needed?
- Why **supervised** dimensionality reduction?
- Advantages and disadvantages: Interpretation, simplified computation, decorrelation of variables, easy visualization, ...
- Problem with repeated testing
- How many variables?

Reduction of dimensionality

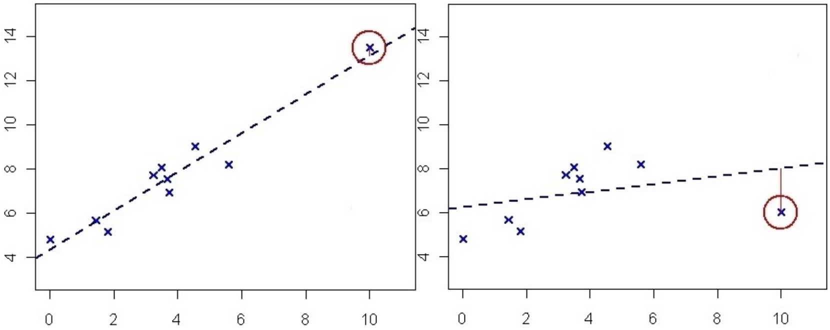
Variable selection:

- Tests (two-sample *t*-test)
- Variable selection based on maximal conditional entropy
- MRMR (Maximum Relevance Minimum Redundancy)
- Bayesian methods
- Intrinsic methods within a regression model

Feature extraction:

- Principal component analysis (PCA)
- Factor analysis
- Independent component analysis (ICA)
- Correspondence analysis
- Methods of information theory

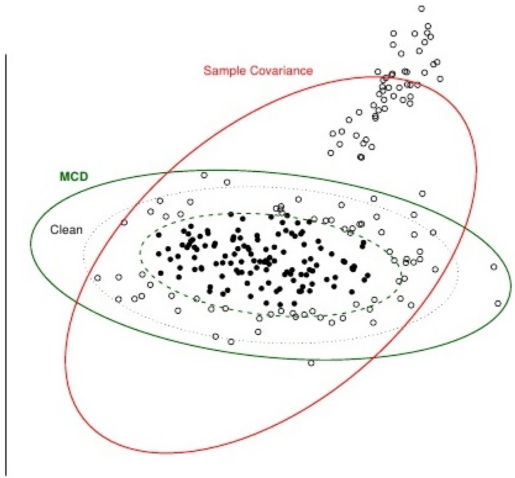
Outliers in linear regression



- Outliers vs. leverage points
- Outlier detection: masking and swamping effects

Outliers in multivariate estimation

Minimum Covariance Determinant (MCD) by Rousseeuw (1985): minimize determinant of sample covariance of 50% of data points:



Classification methods in a study of gene expressions

Robust optimization of mean

The concept of **robust optimization**

- Real numbers $X_1, \ldots, X_n$
- Model
$$X_i = \mu + e_i, \quad \mu \in \mathbb{R}, \quad i = 1, \ldots, n,$$
with i.i.d. random values $e_1, \ldots, e_n$
- The task
$$\operatorname*{argmin}_{a \in \mathbb{R}} \sum_{i=1}^{n} (X_i - a)^2$$
- Solution
$$\hat{a} = \frac{1}{n} \sum_{i=1}^{n} X_i$$
- What if the data are contaminated by measurement errors?

Robust optimization of mean

- We observe
$$X_i = \tilde{X}_i + \delta_i,$$
where $\delta = (\delta_1, \ldots, \delta_n)^T$ is the vector of measurements errors
- The optimization task is replaced by
$$\operatorname*{argmin}_{a \in \mathbb{R}} \max_{|\delta| \leq D} \sum_{i=1}^{n} (X_i - a)^2$$
$$= \operatorname*{argmin}_{a \in \mathbb{R}} \max_{|\delta| \leq D} \sum_{i=1}^{n} (\tilde{X}_i + \delta_i - a)^2,$$
where the requirement $|\delta| \leq D$ denotes
$$|\delta_1| \leq D, \ldots, |\delta_n| \leq D$$
for a fixed $D > 0$.

Robust optimization of mean

The solution has the form

$$\hat{a} = \bar{X} - D, \quad \text{if} \quad \bar{X} > D$$

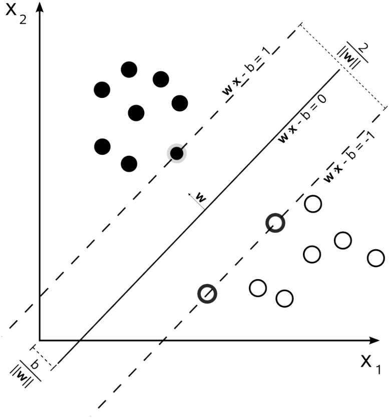
$$\hat{a} = 0, \quad \text{if} \quad -D \leq \bar{X} \leq D$$

$$\hat{a} = \bar{X} + D, \quad \text{if} \quad \bar{X} < -D$$

Some authors understand it as a robust estimator of μ (Tibshirani et al., 2003).

Principles of SVM

- p-dimensional continuous data $X_1, \ldots, X_n$ from two groups
- Response $Y_1, \ldots, Y_n \in \{-1, 1\}$
- We search for a hyperplane $f(x) = w^T x - b$ for classification to two groups, where $w \in \mathbb{R}^p$, $b \in \mathbb{R}$
- Maximal margin



SVM1: Linear SVM, separable case

Maximal margin

$$\min_{w,b} \left\{ \frac{1}{2} ||w||^2 \right\}$$

under the set of constraints

$$Y_i(w^T X_i - b) \geq 1, \quad i = 1, \ldots, n.$$

The solution is obtained as a saddle point of the Lagrange functional

$$\min_{w,b} \max_{\alpha \geq 0} \left\{ \frac{1}{2} ||w||^2 - \sum_{i=1}^{n} \alpha_i \left[Y_i(w^T X_i - b) - 1 \right] \right\}$$

Computation:

- Dual problem (quadratic programming) yields $\hat{\alpha}$
- $\Longrightarrow \hat{w} = \sum_{i=1}^{n} \hat{\alpha}_i Y_i X_i$ (& sparsity)
- $\Longrightarrow \hat{b}$
- A new observation $Z \in \mathbb{R}^p$ is classified according to

$$\mathrm{sgn}(\hat{f}(Z)) = \mathrm{sgn}(\hat{w}^T Z - \hat{b}) = \mathrm{sgn} \left(\sum_{i=1}^{n} \hat{\alpha}_i Y_i X_i^T Z - \hat{b} \right).$$

SVM2: Linear SVM, nonseparable case

The optimization task considers a penalization for violating separability

$$\min_{w,b} \left\{ \frac{1}{2}||w||^2 + C\sum_{i=1}^{n} \xi_i \right\} \quad \text{for a fixed } C > 0$$

under

$$Y_i(w^T X_i - b) \geq 1 - \xi_i, \quad i = 1, \ldots, n,$$

$$\xi_i \geq 0, \quad i = 1, \ldots, n.$$

Exploiting Lagrange multipliers

$$\min_{w,b,\xi\geq 0} \max_{\alpha\geq 0,\beta\geq 0} \left\{ \frac{1}{2}||w||^2 + C\sum_{i=1}^{n} \xi_i - \sum_{i=1}^{n} \alpha_i \left[Y_i(w^T X_i - b) - 1 + \xi_i \right] - \sum_{i=1}^{n} \beta_i \xi_i \right\}.$$

SVM3: Nonlinear SVM, nonseparable case

- We search for the hyperplane $f(x) = h(x)^T w - b$ for classification into two groups
  - $w \in \mathbb{R}^p$
  - $b \in \mathbb{R}$
  - h is a known nonlinear function
- Kernel trick

$$K(X_i, X_j) = h(X_i)^T h(X_j)$$

- Dual problem for the optimization task

$$\max_{\alpha} \left\{ \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j Y_i Y_j K(X_i, X_j) \right\}$$

under corresponding constraints

SVM3: Nonlinear SVM, nonseparable case

- $\hat{w} = \sum_{i=1}^{n} \hat{\alpha}_i Y_i h(X_i)$
- A new observation $Z \in \mathbb{R}^p$ is classified according to the hyperplane:

$$f(Z) = h(Z)^T \hat{w} - b = \sum_{i=1}^{n} \hat{\alpha}_i Y_i K(Z, X_i) - b$$

- Special case with a Gaussian kernel:

$$f(Z) = \sum_{i=1}^{n} \hat{\alpha}_i Y_i \exp\left\{ -\frac{||Z - X_i||^2}{2\sigma^2} \right\} - b \quad \text{for a fixed } \sigma > 0$$

Motivation for robust SVM:

- Measurement errors
- Rounding
- Random regressors
- Uncertainty in regressors

SVM4: Linear SVM, nonseparable case, robust approach

We observe

$$X_i = \tilde{X}_i + \delta_i, \quad i = 1, \dots, n$$

where δ_i is a p-dimensional vector of measurement errors for the i-th observation.

We assume

$$||\delta_i||_p \leq D_i, \quad D_i \in \mathbb{R}, \quad i = 1, \dots, n, \quad p \in [1, \infty].$$

The set of conditions from SVM2

$$Y_i(w^T X_i - b) \geq 1 - \xi_i, \quad i = 1, \dots, n$$

corresponds to

$$Y_i(w^T \tilde{X}_i - b) + Y_i w^T \delta_i \geq 1 - \xi_i, \quad i = 1, \dots, n.$$

SVM4: Linear SVM, nonseparable case, robust approach

This set of conditions is assumed for any $\delta_1, \ldots, \delta_n$:

$$\min_{||\delta_i||_p \leq D_i} \left\{ Y_i(w^T \tilde{X}_i - b) + Y_i w^T \delta_i \right\} \geq 1 - \xi_i, \quad i = 1, \ldots, n.$$

Now we assume a fixed w and search for the solution over δ_i:

$$\min_{||\delta_i||_p \leq D_i} \left\{ Y_i w^T \delta_i \right\}.$$

Hölder inequality yields

$$|Y_i w^T \delta_i| \leq ||w||_q ||\delta_i||_p \leq D_i ||w||_q,$$

where $||.||_q$ is a dual norm to $||.||_p$ and therefore

$$\min_{||\delta_i||_p \leq D_i} \left\{ Y_i w^T \delta_i \right\} = -D_i ||w||_q.$$

SVM4: Linear SVM, nonseparable case, robust approach

Thus, the resulting hyperplane is obtained as a solution of the same optimization task as in SMV2

$$\min_{w,b} \left\{ \frac{1}{2} ||w||^2 + C \sum_{i=1}^{n} \xi_i \right\}$$

but under the set of conditions

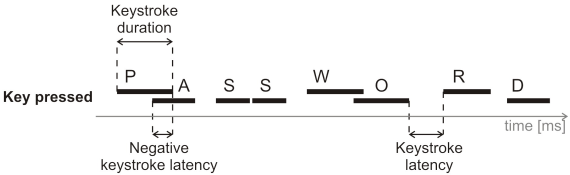
$$Y_i(w^T X_i - b) - D_i ||w||_q \geq 1 - \xi_i, \quad i = 1, \ldots, n,$$

$$\xi_i \geq 0, \quad i = 1, \ldots, n.$$

- The requirement on the norm of the error (in the primary task) yields a regularization of the (primary) task
- Complicated computation
- No implementation in R
- Other approaches: robust nonlinear SVM
- Other approaches: $\min ||w||_p$, $p \in [1, \infty]$

Keystroke dynamics

- 10 individuals
- $10\times$ slowly, $10\times$ quickly
- K–L–A–D–R–U–B–Y
- $p = 15$ variables [in milliseconds]
- Analysis: Semela (2016)



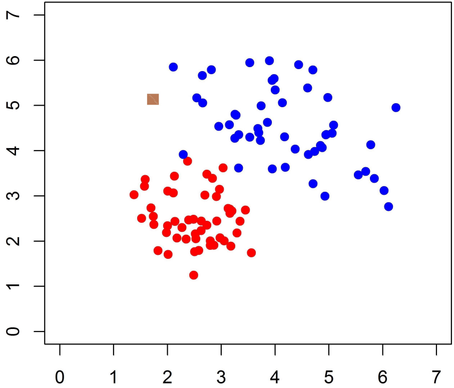
Keystroke dynamics

- First task: Classification of the typing style (speed)
- Second task: Classification of individuals
- Classification accuracy in a leave-one-study

	LDA	Linear SVM	Nonlinear SVM	Linear robust SVM
Classification of the typing style	0.595	0.615	**0.730**	0.645
Optimal value of C	–	0.160	3.000	0.700
Classification of individuals	0.830	0.835	**0.850**	0.715

Classification methods in a study of gene expressions

A classification task to K groups



Mahalanobis distance: $d(Z, \bar{X}_k) = \sqrt{(\bar{X}_k - Z)^T S^{-1} (\bar{X}_k - Z)}, \quad k = 1, \ldots, K$

Linear discriminant analysis (LDA)

Data: K different groups of p-dimensional data.

$$X_{11}, \ldots, X_{1n_1}$$
$$X_{21}, \ldots, X_{2n_2}$$
$$\vdots$$
$$X_{K1}, \ldots, X_{Kn_K}$$

Multivariate normality. Covariance matrix Σ.

An observation Z is classified to the k-th group, which has the maximal value of

$$-\frac{1}{2}(\bar{X}_k - Z)^T S^{-1}(\bar{X}_k - Z) + \log \pi_k,$$

where

- $\bar{X}_k =$ is the mean of the k-th group,
- $S =$ pooled empirical covariance matrix,
- $\pi_k =$ prior probability of the k-th group.

LDA

How LDA can be derived:

- Maximum likelihood for normal data
- $$\max_{a \neq 0} \frac{a^T B a}{a^T W a}$$

  (B variability between groups, W within groups)
- Bayesian approach: max posterior probability

Properties:

- Linear separability
- $\mathrm{P}(Z \in \text{group } 1), \ldots, \mathrm{P}(Z \in \text{group } K)$

Possible extension:

- Quadratic discriminant analysis

Regularized linear discriminant analysis (RDA)

p-dimensional observations in K different groups ($n < p$)

Classification of Z to the k-th group is based on

$$-\frac{1}{2}(\bar{X}_k - Z)^T S^{-1}(\bar{X}_k - Z) + \log \pi_k$$

$$\Downarrow$$

$$-\frac{1}{2}(\bar{X}_k - Z)^T \left(S^*\right)^{-1} (\bar{X}_k - Z) + \log \pi_k$$

Regularized **covariance matrix** for $\lambda \in (0, 1]$: $S^* = (1 - \lambda)S + \lambda T$

Most commonly:

- $T = \mathcal{I}_p$
- $T = \bar{s}\mathcal{I}_p$, where $\bar{s} = \sum_{i=1}^{p} S_{ii}/p$
- $T = \text{diag}\{S_{11}, \dots, S_{pp}\}$

Regularized mean estimation

Definition

- $$\bar{X}_k^{(2)} = (1-\delta^{(2)})\bar{X}_k + \delta^{(2)}\bar{X}, \quad \delta^{(2)} \in [0,1]$$
- $$\begin{aligned} \bar{X}_k^{(1)} &= \operatorname{sgn}(\bar{X}_k)\left(|\bar{X}_k| - \delta^{(1)}\right)_+ \\ &= \operatorname{sgn}(\bar{X}_k)\max\left\{|\bar{X}_k| - \delta^{(1)}, 0\right\}, \quad \delta^{(1)} \in \mathbb{R} \end{aligned}$$
- $$\bar{X}_k^{(0)} = \bar{X}_k \cdot \mathbb{1}\left[|\bar{X}_k > \delta^{(0)}|\right], \quad \delta^{(0)} \in \mathbb{R}$$

- Sparsity
- Choice of regularization parameters

Regularized LDA with different mean estimation

- RDA

$$\ell_k^* = (\bar{X}_k)^T (S^*)^{-1} Z - \frac{1}{2} (\bar{X}_k)^T (S^*)^{-1} \bar{X}_k + \log \pi_k$$

- RDA2

$$\tilde{\ell}_k^{(2)} = (\bar{X}_k^{(2)})^T (S^*)^{-1} Z - \frac{1}{2} (\bar{X}_k^{(2)})^T (S^*)^{-1} \bar{X}_k^{(2)} + \log \pi_k$$

- RDA1

$$\tilde{\ell}_k^{(1)} = (\bar{X}_k^{(1)})^T (S^*)^{-1} Z - \frac{1}{2} (\bar{X}_k^{(1)})^T (S^*)^{-1} \bar{X}_k^{(1)} + \log \pi_k$$

- RDA0

$$\tilde{\ell}_k^{(0)} = (\bar{X}_k^{(0)})^T (S^*)^{-1} Z - \frac{1}{2} (\bar{X}_k^{(0)})^T (S^*)^{-1} \bar{X}_k^{(0)} + \log \pi_k$$

- Which regularization to be used?
- Implementation in R: affine equivariance is lost!
- Regularization $\Longleftrightarrow$ robustness

LDA for $n < p$: Ye et al. (2006), Pekař (2015)

- p-dimensional observations $X_1, \ldots, X_n$ in K groups
- S = (pooled) covariance matrix
- $r = \text{rank}(S)$
- X_k = mean in the k-th group
- $$S_\tau^* = \tau S + (1-\tau)\mathcal{I}_p, \quad \tau \in (0,1)$$

We consider

- $$S = QDQ^T = \begin{pmatrix} Q_r & P \end{pmatrix} \begin{pmatrix} D_r & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} Q_r^T \\ P^T \end{pmatrix}$$
- $S_\tau^* = QD_\tau Q^T$
- $D_{r\tau} = \tau D_r + (1-r)\mathcal{I}_r$

Then

$$\arg \min_{j \in 1,\ldots,K} ||D_\tau^{-1/2} Q^T (Z - \bar{X}_k)|| = \arg \min_{k \in 1,\ldots,K} ||D_{r\tau}^{-1/2} Q_r^T (Z - \bar{X}_k)||.$$

- Ye J., Xiong T., Li Q., Janardan R., Bi J., Cherkassky V., Kambhamettu C. (2006): Efficient model selection for regularized linear discriminant analysis. *Proceedings International Conference on Information and Knowledge Management*, 532 – 539.

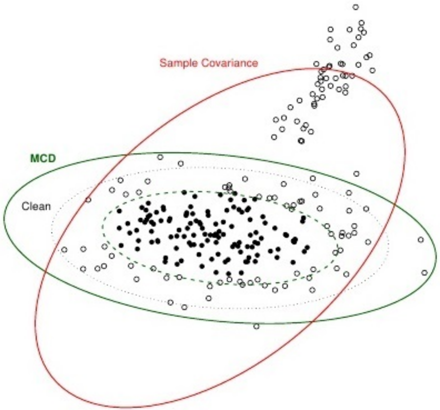
Classification methods in a study of gene expressions

1. Introduction
2. SVM
3. LDA
4. **Robust LDA**

- Duintjer Tebbens J., Kalina J.: A computationally inexpensive improvement of the C-step for the minimum covariance determinant estimator. Submitted to: *Computational Statistics & Data Analysis*.
- Kalina J., Hlinka J.: On coupling robust estimation with regularization for high-dimensional data. *Studies in Classification, Data Analysis and Knowledge Organization.* Accepted.
- Kalina J., Hlinka J.: Implicitly weighted robust classification applied to brain activity research. *Biomedical Engineering Systems and Technologies, Communications in Computer and Information Science.* Accepted.

Why robust statistics?

Minimum Covariance Determinant (MCD) by Rousseeuw (1985):
minimize determinant of sample covariance of 50% of data points:



The concept of robustness

Robust statistics

- Sensitivity of standard methods
- Contaminated normal distribution
- Breakdown point = minimal fraction of data that can drive an estimator beyond all bounds when set to arbitrary values
- Not robustness with respect to the model (data distribution)
- Robustification of standard methods

- Huber P.J. *Robust statistics.* Wiley, New York, 1981.
- Hampel F.R., Rousseeuw P.J., Ronchetti E.M., Strahel W.A. *Robust Statistics: The approach based on influence functions.* Wiley, New York, 1986.
- Rousseeuw P.J., Leroy A.M. *Robust regression and outlier detection.* Wiley, New York, 1987.
- Jurečková J., Sen P.K., Picek J. *Methodology in robust and nonparametric statistics.* CRC Press, Boca Raton, 2013.

Robust estimation of multivariate location and scatter

- $X_1, \dots, X_n$ i.i.d. p-dimensional
- $n > p$
- Elliptically symmetric unimodal distribution
  - $$f(x) = \frac{1}{(\det \Sigma)^{1/2}} g\left((x-\mu)^T \Sigma^{-1} (x-\mu)\right), \quad x \in \mathbb{R}^p$$
  - $\mu \in \mathbb{R}^p$
  - $\Sigma \in PDS(p \times p)$
  - g decreasing function
- Minimum Covariance Determinant (MCD)
  - Rousseeuw P.J., Leroy A.M. (1984): Least median of squares regression. *Journal of the American Statistical Association* **79**, 871 – 880.
- Minimum Weighted Covariance Determinant (MWCD)
  - Roelant E., van Aelst S., Willems G. (2009): The minimum weighted covariance determinant estimator. *Metrika* **70**, 177 – 201.

Minimum covariance determinant (MCD)

- Robust estimator of multivariate location and scatter
- H = subset of h observations
- $$\bar{X}_{MCD} = \sum_{i \in H} w_i X_i$$
- $$S_{MCD} = \delta \sum_{i \in H} (X_i - \bar{X}_{MCD})(X_i - \bar{X}_{MCD})^T,$$
  where δ is a consistency factor (to ensure Fisher consistency)
- $$\min \det(S_{MCD})$$
  over all h-subsets of observations
- Global & local robustness, affine equivariance, consistency, asymptotic normality

Minimum Weighted Covariance Determinant (MWCD)

- Weights $w_1 \geq w_2 \geq \cdots \geq w_n$; $\sum_{i=1}^{n} w_i = 1$.
- $$\bar{X}_{MWCD} = \sum_{i=1}^{n} w_i X_i$$
- $$S_{MWCD} = \delta \sum_{i=1}^{n} w_i (X_i - \bar{X}_{MWCD})(X_i - \bar{X}_{MWCD})^T$$
- $$\min \det(S_{MWCD})$$
  over all permutations of weights
- Approximate algorithm

Minimum Weighted Covariance Determinant (MWCD)

$$\begin{pmatrix} \bar{X}_{MWCD} \\ \tilde{S}_{MWCD} \end{pmatrix} = \underset{m,C;det\ C=1}{\operatorname{argmin}} \sum_{i=1}^{n} a_n(R_i) \underbrace{(X_i - m)^T C^{-1} (X_i - m)}_{d_i^2(m,C)}$$

- a_n = nonincreasing function
- $m \in \mathbb{R}^p$
- C = symmetric positive definite matrix $p \times p$
- R_i is the rank $d_i^2(m, C)$ among $d_1^2(m, C), \ldots, d_n^2(m, C)$.
- $S_{MWCD} = \delta \tilde{S}_{MWCD}$, where δ is a consistency factor

Weights for the MWCD estimator

Fixed magnitudes of weights:

- Linearly decreasing weights
- Properties of the estimator & corresponding functional

Adaptive (data-dependent) weights:

- $$w(t) = \frac{F_\chi^{-1}(t)}{\left(G_n^0\right)^{-1}(t)}, \quad t \in \left\{ \frac{1}{2n}, \frac{3}{2n}, \dots, \frac{2n-1}{2n} \right\}$$
  - F_χ^{-1} = quantile function of χ_p^2 distribution
  - $\left(G_n^0\right)^{-1}$ = empirical quantile function of $d_1^2(\hat{\mu}, \hat{\Sigma}), \dots, d_n^2(\hat{\mu}, \hat{\Sigma})$
- Approximate algorithm
- Asymptotic eficiency
- High **breakdown point**

Regularized MWCD estimator

- MWCD: Infeasible for a high dimension
- Regularized MWCD-covariance matrix S^*_{MWCD}:

$$\min \det\left((1-\lambda)S_w + \lambda \mathcal{I}_p\right), \quad \lambda \in (0,1]$$

- High robustness
- Regularized MWCD estimator (using M-estimation of Chen et al., 2011) $\Longrightarrow \bar{X}_{k,MWCD}, \ \tilde{S}_{MWCD}$

Proposal of MWCD-RDA, MWCD-RDA2, MWCD-RDA1, MWCD-RDA0.

Example: Cardiovascular genetic study

Classification to 2 groups:

- 24 patients vs. 24 controls
- $p = 38\,590$ gene expressions
- Leave-one-out cross validation
- Youden's index = sensitivity + specificity −1

Method	Youden's index
LDA	**1.00**
RDA1	**1.00**
SVM	**1.00**
Classification tree	0.94
Lasso-LR	0.97
Multilayer perceptron	Infeasible
MWCD-RDA	**1.00**
MWCD-RDA2	**1.00**
MWCD-RDA1	**1.00**
Dimensionality reduction	10 variables
PCA $\Longrightarrow$ LDA	0.15
PCA $\Longrightarrow$ MWCD-RDA1	0.62

Example: Brain activity

- Leave-one-out cross validation
- Contamination by $N(0, \sigma^2)$ noise

Method	Youden's index = sensitivity + specificity −1			
	Raw data	$\sigma = 0.1$	$\sigma = 0.2$	$\sigma = 0.3$
RDA1	**1.00**	**1.00**	**1.00**	0.99
SVM	**1.00**	0.99	0.98	0.96
Classification tree	0.96	0.95	0.91	0.92
Lasso-LR	0.99	**1.00**	0.97	0.94
MWCD-RDA	**1.00**	**1.00**	**1.00**	**1.00**
MWCD-RDA2	**1.00**	**1.00**	**1.00**	**1.00**
MWCD-RDA1	**1.00**	**1.00**	**1.00**	**1.00**
Dimensionality reduction	10 variables			
PCA $\Longrightarrow$ LDA	**1.00**	0.94	0.93	0.88
PCA $\Longrightarrow$ MWCD-RDA	**1.00**	0.95	0.94	0.89
PCA $\Longrightarrow$ MWCD-RDA2	**1.00**	0.95	0.94	0.89
PCA $\Longrightarrow$ MWCD-RDA1	**1.00**	0.95	0.94	0.89

Two other examples

Method	Youden's index Metabolomic profiles	Youden's index Keystroke dynamics
K	$K = 2$	$K = 2$
n	$n = 42$	$n = 32$
p	$p = 518$	$p = 15$
RDA1	0.91	0.80
SVM	**0.92**	**0.85**
Classification tree	0.84	0.11
Lasso-LR	0.87	0.82
MWCD-RDA	0.91	0.79
Dimensionality reduction	20 variables	4 variables
PCA $\Longrightarrow$ LDA	0.70	0.59
PCA $\Longrightarrow$ MWCD-RDA	0.72	0.59
MRMR $\Longrightarrow$ LDA	0.88	0.72
MRMR $\Longrightarrow$ MWCD-RDA	0.90	0.76

Discussion: robust classification

Advantages of MWCD-RDA (and other versions):

- Improvement for contaminated data
- No need for a prior dimensionality reduction
- Comprehensibility
- An efficient algorithm based on numerical linear algebra

Limitations of MWCD-RDA:

- Contaminated multivariate normal data
- The weights are assigned to individual observations
- Variability not substantially different across variables
- Intensive computations are required
- Regularization parameters should be small

Conclusions

- Introduction
- SVM
- LDA
- Robust LDA

Problems of common classifiers:

- Various data formats
- Computational demands
- Missing values
- Instability
- Dimensionality reduction?
- *"No free lunch"* theorems
- Design issues (how many observations?)

Conclusions

Machine learning:

- Universal classifiers?
- Linear separability for $n < p$ is guaranteed!
- SVM
  - Too many support vectors
  - $\Longrightarrow$ overfitting
  - No regularization
- Complicated for $K > 2$ (voting scheme etc.)
- Suboptimal solution
- Interpretation

$\Longrightarrow$ THANK YOU FOR YOUR ATTENTION $\Longleftarrow$