Statistical aspects of prediction models with high-dimensional data

Anne–Laure Boulesteix
Institut für Medizinische Informationsverarbeitung, Biometrie und Epidemiologie

February 15th, 2017

Structure

1. **Introduction and examples**
2. Overview of prediction methods for high-dimensional data
3. Estimation of prediction error: cross-validation and related methods
4. Dos and donts, good practice

Example: Predicting survival of CLL leukemia patients using gene expression data

Letter to the Editor

Leukemia (2011) **25,** 1639–1645; doi:10.1038/leu.2011.125; published online 31 May 2011

An eight-gene expression signature for the prediction of survival and time to treatment in chronic lymphocytic leukemia

T Herold[1], V Jurinovic[2], K H Metzeler[1], A-L Boulesteix[2], M Bergmann[1,5], T Seiler[1], M Mulaw[3], S Thoene[3], A Dufour[1], Z Pasalic[1], M Schmidberger[2], M Schmidt[4], S Schneider[1], P M Kakadia[1,3], M Feuring-Buske[5,6], J Braess[1], K Spiekermann[1,3], U Mansmann[2], W Hiddemann[1,3], C Buske[6,7] and S K Bohlander[1,3,7]

[1]Department of Internal Medicine III, University Hospital Grosshadern, Ludwig-Maximilians-University (LMU), Munich, Germany

[2]Institute for Medical Informatics, Biometry and Epidemiology (IBE), Ludwig-Maximilians-University (LMU), Munich, Germany

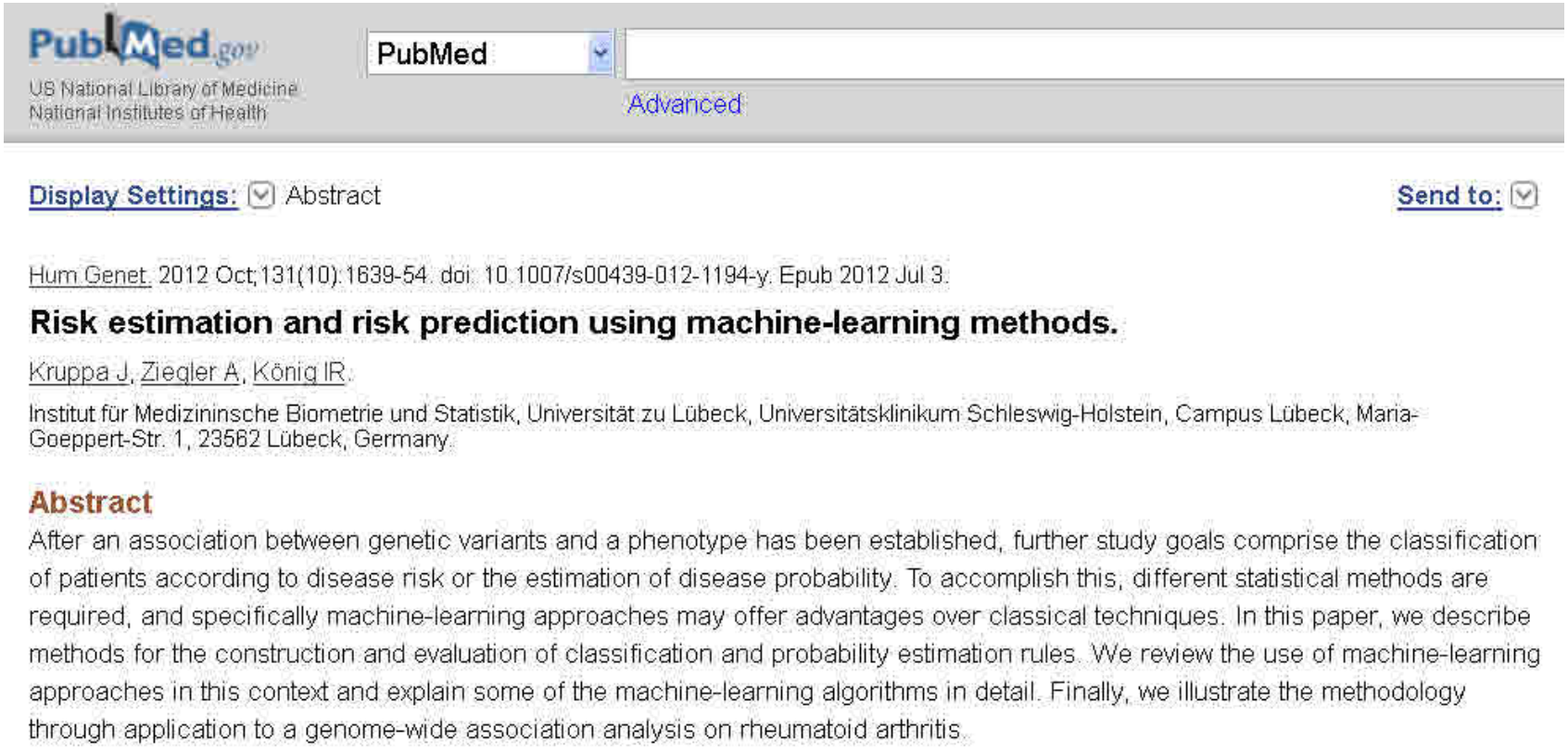
[3]Clinical Cooperative Group Acute Leukemia, Helmholtz Center Munich for Environmental Health, Munich, Germany

[4]Munich Cancer Registry (MCR) of the Munich Cancer Center (MCC) at the Institute for Medical Informatics, Biometry and Epidemiology (IBE), Ludwig-Maximilians-University (LMU), Munich, Germany

[5]Department of Internal Medicine III, University Hospital Ulm, Ulm, Germany

[6]Institute of Experimental Cancer Research, Comprehensive Cancer Center Ulm, University of Ulm, Ulm, Germany

Example: Predicting disease risk based on genome-wide SNP data



PubMed.gov
US National Library of Medicine
National Institutes of Health

PubMed

Advanced

Display Settings: Abstract

Send to:

Hum Genet. 2012 Oct;131(10):1639-54. doi: 10.1007/s00439-012-1194-y. Epub 2012 Jul 3.

Risk estimation and risk prediction using machine-learning methods.

Kruppa J, Ziegler A, König IR.

Institut für Medizininsche Biometrie und Statistik, Universität zu Lübeck, Universitätsklinikum Schleswig-Holstein, Campus Lübeck, Maria-Goeppert-Str. 1, 23562 Lübeck, Germany.

Abstract

After an association between genetic variants and a phenotype has been established, further study goals comprise the classification of patients according to disease risk or the estimation of disease probability. To accomplish this, different statistical methods are required, and specifically machine-learning approaches may offer advantages over classical techniques. In this paper, we describe methods for the construction and evaluation of classification and probability estimation rules. We review the use of machine-learning approaches in this context and explain some of the machine-learning algorithms in detail. Finally, we illustrate the methodology through application to a genome-wide association analysis on rheumatoid arthritis.

Further examples

- molecular diagnosis
- prediction of response to therapy (pharmacogenomics, personalized medicine)
- ...

High-dimensional

Classical data:

Index	Outcome	Sex	Age	Smoking	BMI
1	bad	F	54	yes	28
2	good	M	50	yes	22
...	...	...	...	...	

High-dimensional data:

Index	Outcome	Sex	Age	Smoking	BMI	V_1	...	...	V_{1000}
1	bad	F	54	yes	28	0.97	...	...	1.33
2	good	M	50	yes	22	1.25	...	...	2.87
...	...	...	...	...		...	...		

Terminology I
What we try to predict

- **Examples:** survival time, response to therapy, disease outcome, risk of developing disease,...
- **Types of variable:** time-to-event, class membership, continuous variable
- **Terminology in literature:** response variable, dependent variable, outcome, Y,...
- **Terminology in this lecture:** Y (because of usual notation in linear model: $Y = a + bX$)

Terminology II
Predictors

- **Examples we are considering here:** gene expression data, metabolomic data, SNP data, proteomic data,...
- **Types of variables:** categorical (SNPs), continuous
- **Terminology in the literature:** predictors, covariates, independent variables, variables, prognostic factors, features,...
- **Terminology in this lecture:** predictors $V_1, \ldots, V_{1000}$

Terminology III
Prediction rules

- **Prediction rule:** a function that takes as an argument the values of the predictors for a new patient and outputs a prediction of Y for this patient.

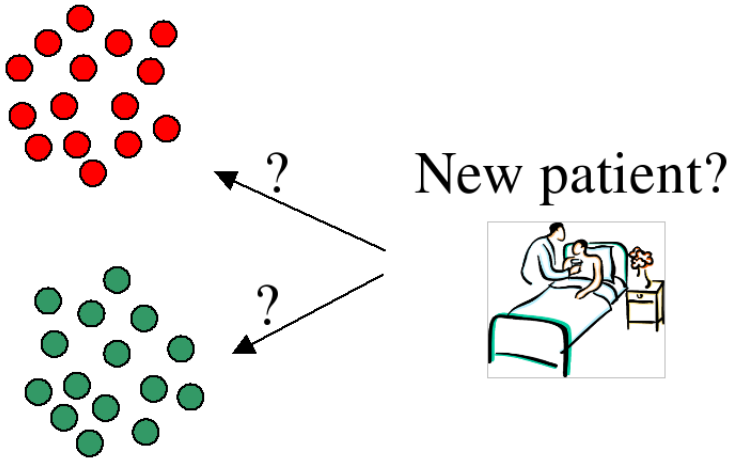
- **Terminology in the literature:** prediction model, prediction function, classifier, classification model, classification rule, prognostic model, learning rule, signature, predictive model,...

- **Terminology in this lecture:** prediction rule

How to derive a prediction rule?

- **Aim:** Construct a prediction rule that relates the predictors to the variable Y in order to make predictions for future patients.
- **How:** By "looking" at a data set containing the true values of Y and the values of the predictors for a relevant set of patients.
- In complex situations like those considered here, "looking at a data set" means running an algorithm on the data that automatically constructs the prediction rule.
- **Problem:** How should such an algorithm be designed in order to yield a good prediction rule? That is the research topic of methodological statisticians/machine learners.

How to derive a prediction rule?



Look at patients with known Y

Y	V_1	...	...	V_{1000}
bad	0.97	...	...	1.33
good	1.25	...	...	2.87
...				

and derive a rule that relates Y and $V_1, \ldots, V_{1000}$

Common problems

It is not so easy to design such an algorithm:

- too many predictors to use a classical method (logistic, Cox)
- poor priori knowledge on role of predictors
- high correlations between predictors
- no practical graphical representation

In this lecture we will give an overview of possible simple algorithms used in practice to derive prediction rules from a large set of predictors.

Structure

1. Introduction and examples
2. **Overview of prediction methods for high-dimensional data**
3. Estimation of prediction error: cross-validation and related methods
4. Dos and donts, good practice

Common approaches to build prediction models

- classical regression model (linear, logistic, Cox) built after **selection of the most informative predictors**
- **penalized** regression approaches
- PLS **dimension reduction**
- machine learning approaches: support vector machines, random forests

Selection of the most informative variables

Idea

- Pre-filter informative predictors based on a univariate criteria
- Use the top predictors in a classical regression method

Example

- For each variable $V_1, \ldots, V_{1000}$, compute p-value of t-test testing equality of means in groups $Y = 0$ and $Y = 1$.

- Select the 5 variables with the smallest p-values
  $\rightarrow V_{73}, V_{98}, V_{254}, V_{406}, V_{408}$.

- Fit logistic regression model

$$\log \frac{P(Y=1)}{P(Y=0)} = \beta_0 + \beta_{73} V_{73} + \beta_{98} V_{98} + \beta_{254} V_{254} + \beta_{406} V_{406} + \beta_{408} V_{408}$$

Issues with variable selection

- How many predictors?
- Convergence issues in logistic and Cox regression
- Univariate variable selection may be suboptimal (does not take correlations between variables into account)
- Penalized regression may be a solution

Penalized regression: Principle

- Consider classical regression model (linear, logistic, Cox) but impose a constraint on the coefficients to keep them small.
- By doing that, we circumvent the dimension problem mentioned in the introduction lecture.
- **Advantage 1:** Familiar methods, simple linear predictor
- **Advantage 2:** You do not have to bother about variable selection or dimension reduction. The method automatically takes into account the fact that not all predictors are equally important for the prediction problem and makes some kind of intrinsic variable selection.

Penalized regression: Technical aspects

- **Mathematical formulation:** minimize $-\ell(\boldsymbol{\beta}) + S(\boldsymbol{\beta}, \lambda)$ with, e.g. $S(\boldsymbol{\beta}, \lambda) = \lambda \sum_j |\beta_j|$ (lasso regression) or $S(\boldsymbol{\beta}, \lambda) = \lambda \sum_j \beta_j^2$ (ridge regression)

- Implemented in R packages `penalized` and `glmnet`

- Choice of λ usually done by cross-validation

- Advantage of Lasso: Many coefficients are shrunken to zero = intrinsic variable selection

PLS dimension reduction

- **Idea:** Construct linear combination (PLS components) of $V_1, \ldots, V_{1000}$ of the form

$$a_1 \cdot V_1 + \cdots + a_p \cdot V_{1000}$$

that have maximal covariance with response Y.

- These PLS components can then be used as predictors in a classical regression method.

- Simple and computationally efficient algorithms are available for finding the linear combinations that have maximal covariance with Y.

Other approaches

- **Support Vector Machines** (especially for classification): good performance but black box and a parameter (cost) to choose
- **Random Forest** (for both classification and regression): good performance but black box and several parameters to choose
- **Elastic net** = mixture of lasso regression and ridge regression combining their advantages

Structure

1. Introduction and examples

2. Overview of prediction methods for high-dimensional data

3. **Estimation of prediction error: cross-validation and related methods**

4. Dos and donts, good practice

Prediction error

- For a binary Y: error rate, sensitivity, specificity, etc

	$Y = 0$	$Y = 1$
$\hat{Y} = 0$	$\ldots$	$\ldots$
$\hat{Y} = 1$	$\ldots$	$\ldots$

- For a continuous Y: mean squared error mean$((Y_i - \hat{Y}_i)^2)$

Estimation of prediction error

- Suppose you have estimated a prediction rule $\hat{f}$ based on the available data set D, for instance

$$\hat{Y} = 1 \text{ if } 0.256 + 0.118 \cdot V_{25} + 0.078 \cdot V_{546} - 0.055 \cdot V_{887} > 0, \quad \hat{Y} = 0 \text{ otherwise}$$

- The **question** is now how good it is at predicting new patients.

- If we had another set of new patients with known Y, we would apply the prediction rule to these patients and look whether the prediction is correct by comparing to the true value Y.

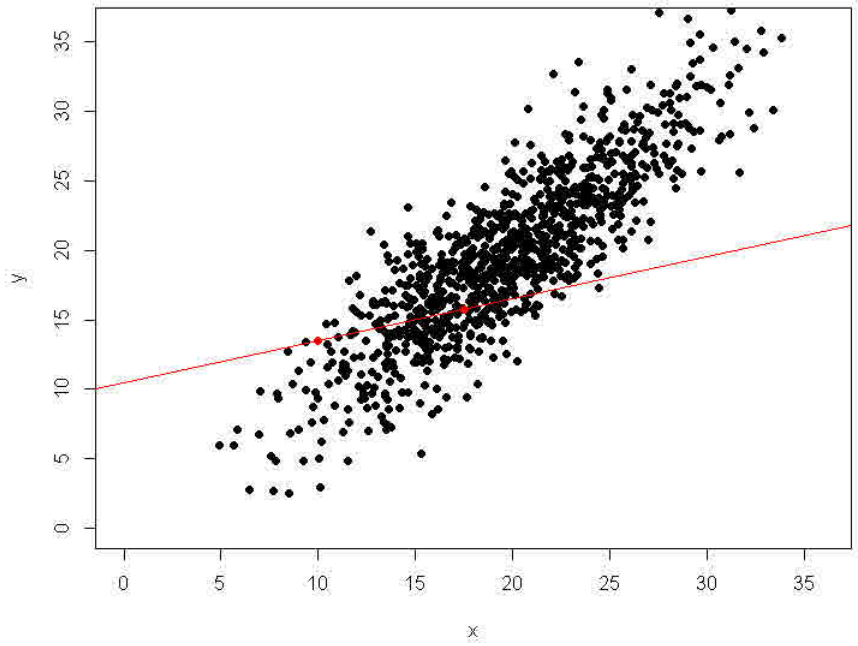
- The **problem** is that we have no other set of patients to use for this evaluation.

Estimation of prediction error based on training data

- The **naive approach** consists to use the same patients again for the evaluation of the prediction rule, i.e. to apply the prediction rule to this set of patients and compare the true and predicted Y.
- This is a WRONG approach, at least in high-dimensional settings!
- By doing that, we substantially underestimate the error.
- That is because $\hat{f}$ has been constructed especially to predict these patients well. But it would probably be much worse on new patients.

Example 1: regression with one predictor

Illustration with only $n = 2$, one predictor x, a continuous Y and linear regression as prediction method:



(red patients are used to construct the prediction rule, black patients are new)

Example 2: classification

- Binary response Y
- $n = 6$: 3 with $Y = 0$, 3 with $Y = 1$
- 100 binary predictors
- selection of the best predictor

It is likely that at least one of the 100 candidate predictors separates the two classes $Y = 0$ and $Y = 1$ perfectly $\rightarrow$ estimating prediction error on training data would yield an unrealistic accuracy of 100%

Overfitting

- In high-dimensional settings, there are (infinitely) many possible prediction rules. The algorithm selects one of those that best fit to the data set at hand.

- Since this selection is done among many possible prediction rules, it usually fits the data set at hand very well... much better than it will on future data sets!

- This mechanism is called **overfitting** and is particularly relevant to high-dimensional data.

- Roughly speaking, **overfitting** is the reason why **one should not estimate prediction error with the data used to derive the prediction rule.**

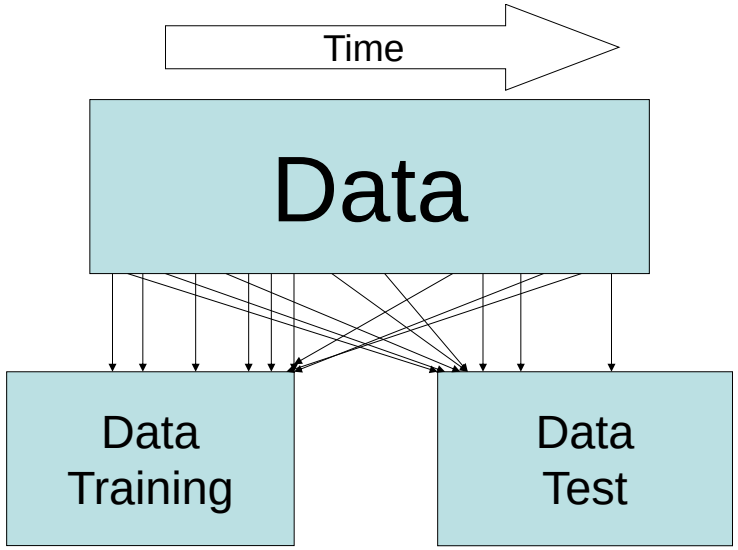
Solutions?

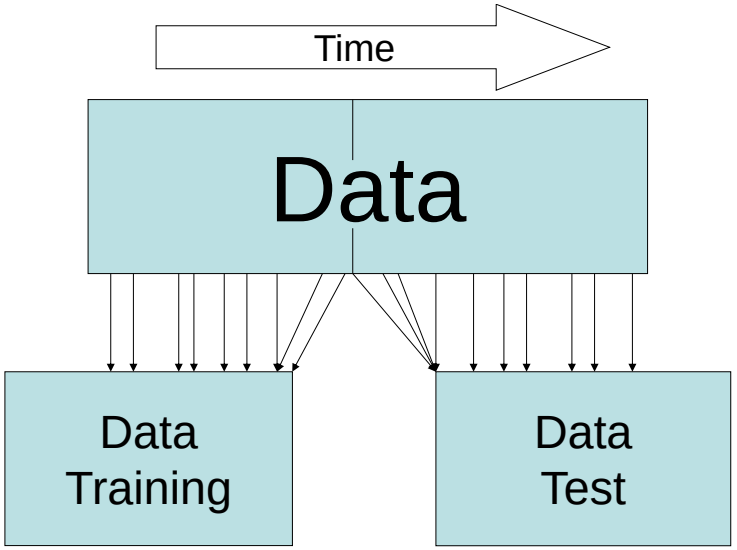
- A solution is to artificially split the data set into a **training set** (used to construct the prediction rule) and a **test set** (used to evaluate prediction error by comparing the true and predicted Y).

- This is possible only if the data set is large enough.

- The test set should NEVER be used for the construction of the prediction rule.

Different splitting approaches





Problems of single splitting and solution

- Estimated error depends a lot on the particular splitting into training and test data.
- **Solution:** Consider several splittings successively and build the average. This is called **repeated splitting** or **repeated subsampling**.
- **In practice:** at least 50 splittings, training:test ratio of 2:1, 4:1 or 9:1.

Other solution: cross-validation

- Partition the data into K subsets $D_1, \ldots, D_K$.
- For $k = 1, \ldots, K$, repeat:
  1. Exclude D_k (blue) and construct prediction rule based on the remaining data (orange).
  2. Apply the prediction rule to D_k and compute its prediction error.

  and then build the average prediction error over the K iterations.
- This approach is strongly related to repeated splitting.
- Special case: "leave-one-out cross-validation" means that each subset contains only patient.

Structure

1. Introduction and examples
2. Overview of prediction methods for high-dimensional data
3. Estimation of prediction error: cross-validation and related methods
4. **Dos and donts, good practice**

Good practice – does and donts

ARTICLE

Critical Review of Published Microarray Studies for Cancer Outcome and Guidelines on Statistical Analysis and Reporting

Alain Dupuy, Richard M. Simon

Background Both the validity and the reproducibility of microarray-based clinical research have been challenged. There is a need for critical review of the statistical analysis and reporting in published microarray studies that focus on cancer-related clinical outcomes.

Methods Studies published through 2004 in which microarray-based gene expression profiles were analyzed for their relation to a clinical cancer outcome were identified through a Medline search followed by hand screening of abstracts and full text articles. Studies that were eligible for our analysis addressed one or more outcomes that were either an event occurring during follow-up, such as death or relapse, or a therapeutic response. We recorded descriptive characteristics for all the selected studies. A critical review of outcome-related statistical analyses was undertaken for the articles published in 2004.

Results Ninety studies were identified, and their descriptive characteristics are presented. Sixty-eight (76%) were published in journals of impact factor greater than 6. A detailed account of the 42 studies (47%) published in 2004 is reported. Twenty-one (50%) of them contained at least one of the following three basic flaws: 1) in outcome-related gene finding, an unstated, unclear, or inadequate control for multiple testing; 2) in class discovery, a spurious claim of correlation between clusters and clinical outcome, made after clustering samples using a selection of outcome-related differentially expressed genes; or 3) in supervised prediction, a biased estimation of the prediction accuracy through an incorrect cross-validation procedure.

Conclusions The most common and serious mistakes and misunderstandings recorded in published studies are described and illustrated. Based on this analysis, a proposal of guidelines for statistical analysis and reporting for clinical microarray studies, presented as a checklist of "Do's and Don'ts," is provided.

Don't use the test data for variable selection

- A very common error that leads to completely false prediction error estimates!
- If one selects the most relevant predictors from a set of candidate predictors, this should be considered as part of the construction of the prediction rule and **done with the training data only**.
- Violating this rule may lead to prediction errors of 95% even if the data are completely random!
- See Ambroise & Mc Lachlan (PNAS 2002).

Don't use the test data for variable selection

If cross-validation is performed, it means that:

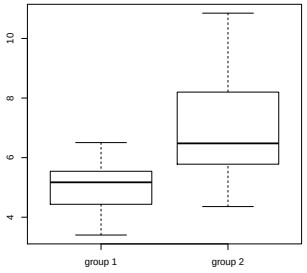
- one should repeat the variable selection process for each iteration
- or equivalently: one should not perform any "preliminary filtering" before doing the cross-validation.

See the dos and donts by Dupuy & Simon:

26	Don't	Use the same feature selection for all iterations.
27	Don't	Perform a cross-validation procedure on a selection of outcome-related differentially expressed genes.

Do not overinterpret plots drawn from the training data

- **Example 1:** If you select 5 most differentially expressed genes (between two conditions) out of 2000 genes, it is not surprising that the boxplots are different in the two conditions. They would look different even if the data were completely random.
- **Example 2:** If you perform PLS dimension reduction on a data set consisting of two groups, it is not surprising that the PLS components separate the two groups well. They have been designed for that and would also apparently separate the groups even if the data were completely random.



Good practice in clustering

Clustering is often performed on omics data to identify new groups of patients (see lecture by Prof. Mansmann). Typical flaws should be avoided (Dupuy and Simon, JNCI 2007):

14	Don't	Use class discovery methods if you are interested in classifying new samples in the future.
15	Don't	Use a selection of outcome-related differentially expressed genes if you intend to correlate cluster-defined classes with the outcome.
16	Don't	Select the clustering method that gives the best result.
21	Don't	Attempt to predict cluster-defined classes.
18	Don't	Use conventional statistical tests for computing the statistical significance of genes that are differentially expressed between two clusters.

Do not fish for significance

28	Do	Report the estimates for all the classification algorithms if several have been tested, not just the most accurate.

- Otherwise you get substantially biased (optimistic) results.
- This problem does not only affect prediction studies but also any type of statistical analysis.
- This problem is particularly relevant in high-dimensional data analysis because in this setting the results are very variable, thus yielding a high probability that at least one of the results looks good by chance.