Over-optimism in biostatistics and bioinformatics

Anne-Laure Boulesteix

joint with M. Jelizarow, V. Guillemot, A. Tenenhaus, K. Strimmer, C. Strobl

Institut für Medizinische Informationsverarbeitung, Biometrie und Epidemiologie
Ludwig-Maximilians-Universität München

Berlin, October 6th 2010
Workshop on "Validation in Statistics and Machine Learning"

An example: prediction models based on high-dimensional molecular data

$\mathbf{Y}$	$\mathbf{X_1}$	...	$\mathbf{X_p}$
0	...	...	...
0	...	...	...
...	...	...	...
1	...	...	...
1	...	...	...
...	...	...	...

Available data:

- an outcome variable $\mathbf{Y}$ which has to be predicted, such as survival time, responder/ non-responder, etc
- high-dimensional molecular predictors $\mathbf{X_1}, \ldots, \mathbf{X_p}$ such as gene expression data, metabolomic data, proteomic data, etc

Challenge: $n \ll p$

Spoilt for choice?

- There is no gold standard prediction method for $n \ll p$ data and no easy diagnostic tool to choose the method.
- A few methods are known to work well in general:

M_1	SVM	$\rightarrow$	$CV(M_1)$
M_2	Random Forests	$\rightarrow$	$CV(M_2)$
M_3	Nearest shrunken centroids	$\rightarrow$	$CV(M_3)$
...	...	...	...
M_K	Lasso	$\rightarrow$	$CV(M_K)$

Simon and Dupuy (JNCI, 2007): "*Do* report the [cross-validation] estimates for all the classification algorithms if several have been tested, not just the most accurate."

→ Reporting all error rates may be confusing.

→ Trying only one method may lead to bad accuracies.

- In practice, one often **tries several methods successively** in a cross-validation framework and presents only the results obtained with the most accurate method.
- However, selecting the method *a posteriori* on the basis of the obtained results may introduce **a severe bias** as shown in our empirical study.

Boulesteix and Strobl, 2009. BMC Med. Res. Meth. 9:85.

Empirical assessment of the bias via data-driven simulations

Design of our study:

1. Generate realistic data sets with non-informative predictors by **balanced permutation of the class label in real data sets** (colon cancer, prostate cancer).
2. Compute CV error rates for each data set using different "acceptable" methods for each data set: *kNN*, LDA, FDA, DLDA, PLS+LDA, NN (combined with different variable selection schemes), and SVM, Lasso, RF, NSC.
3. Select the minimal error rate over the different methods for each data set.

Empirical assessment of the bias via data-driven simulations

Results:

The median minimal error rate is as low as 31% in the colon data and 41% in the prostate data... although we are sure that the response and the predictors are not associated!

A.-L. Boulesteix, C. Strobl, 2009. Optimal classifier selection and negative bias in error rate estimation: An empirical study on high-dimensional prediction. *BMC Medical Research Methodology* 9:85.

Bias correction:

See the talk by Christoph Bernau tomorrow.

Bias in methodological research

- When developing statistical methods, researchers often think of several possible variants (called "methods' characteristics" here).
- If they choose the methods' characteristics a posteriori (i.e. because they obtain nice results with these characteristics), the results of the new method are also optimistically biased!

Here we present an empirical study to illustrate this bias and the need for validation with independent data.

Jelizarow et al, 2010. Over-optimism in bioinformatics: an illustration. Bioinformatics 26:1990–1998.

Optimization mechanisms

- Optimization of the data sets: Try the new method on different data sets... and report only the best results...
- Optimization of the competing methods: Omit the best state-of-the-art competing methods in the comparison study.
- **Optimization of the settings**: Try the new method in combination with different variable selection or preprocessing steps... and report only the best results...
- **Optimization of the methods' characteristics**: Consider several variants of the new method... and report only the best results...

Jelizarow et al, 2010. Over-optimism in bioinformatics: an illustration. Bioinformatics 26:1990–1998.

The history of our project

- We had an idea that we considered as promising.
- The idea was to incorporate biological knowledge on the gene structure into linear discriminant analysis (see the next slides).
- This idea turned out to yield bad results in terms of prediction accuracy.
- ... But it is possible to report seemingly good results by "fishing for significance".

Here, we present a quantitative study on optimization mechanisms in methodological research based on this example.

Jelizarow et al, 2010. Over-optimism in bioinformatics: an illustration. Bioinformatics 26:1990–1998.

A "promising" method

Discriminant function in linear discriminant analysis:

$$d_r(\mathbf{x}) = \mathbf{x}^\top \Sigma^{-1} \boldsymbol{\mu}_r - \frac{1}{2} \boldsymbol{\mu}_r^\top \Sigma^{-1} \boldsymbol{\mu}_r + \log(\pi_r),$$

Problem: The sample estimator $\hat{\Sigma}$ of the covariance matrix Σ is not invertible when $n \ll p$!

Solution: Use a regularized estimator of Σ instead of the sample covariance $\hat{\Sigma}$, for instance the shrinkage estimator by Schäfer and Strimmer (2005):

$$\hat{\Sigma}^* = \lambda \hat{\Sigma} + (1 - \lambda) T,$$

where T is an adequately chosen target and λ a shrinkage parameter.

A "promising" method

Idea: Define T using priori knowledge on the gene functional groups (GFG) from the KEGG database:

Target D

$$t_{ij} = \begin{cases} s_{ii} & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$

Target G

$$t_{ij} = \begin{cases} s_{ii} & \text{if } i = j \\ \bar{r}\sqrt{s_{ii}s_{jj}} & \text{if } i \neq j, i \sim j \\ 0 & \text{otherwise} \end{cases}$$

Problem: How should we deal with genes that are in no GFG, genes that are in several GFG, negative correlations within GCG, non-significant correlations?

$\rightarrow$ 10 candidate variants

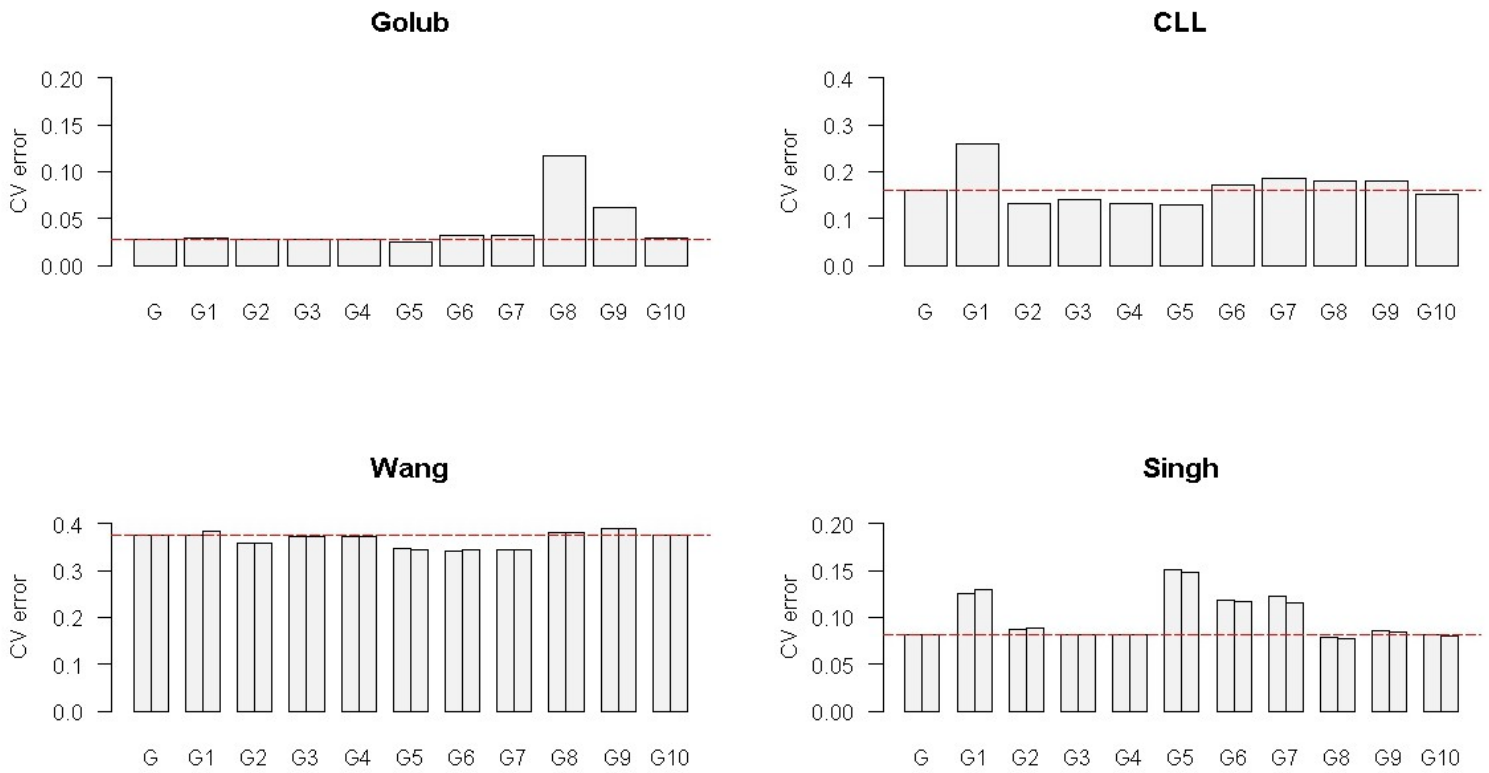
A "promising" method

Design of the study:

- Data: Four "$n \ll p$" microarray data sets with binary response variable
- The data sets are first analysed separately.
- For each data set, we look for the best variable selection setting (out of 12) and the best variant (out of 10).
- Then we look at the performance of this "best combination" on the other 3 data sets $\approx$ validation.

Jelizarow et al, 2010. Over-optimism in bioinformatics: an illustration. Bioinformatics 26:1990–1998.

Selecting the methods' characteristics optimally



The error rate can be decreased by optimizing the "methods' characteristics" (i.e. by choosing the optimal variant for a particular data set).

Selecting the methods' characteristics optimally

	M_{opt}	s_{opt}	Golub	CLL	Wang	Singh
Golub	rlda.TG$^{(5)}$	s_{opt} = (200, Limma)	**0.025**	0.180	0.345	0.152
CLL	rlda.TG$^{(5)}$	s_{opt} = (200, Wilcoxon test)	0.079	**0.129**	0.363	0.141
Wang	rlda.TG$^{(6)}$	s_{opt} = (200, t-test)	0.029	0.221	**0.342**	0.115
Singh	rlda.TG$^{(8)}$	s_{opt} = (100, Limma)	0.033	0.274	0.384	**0.078**

- Seemingly good results are obtained by "fishing for significance" (i.e. optimizing the variable selection setting and the methods' characteristics).
- These seemingly good results cannot be validated based on other data sets.

Sources of the problems

Results presented in statistical bioinformatics papers are sometimes the product of intense optimization: optimization of the settings and optimization of the methods characteristics.

- **Problem 1:** Error rate estimators have high variance in $n \ll p$ settings, hence the opportunity for optimization. If we had a very large data set, we would not have this problem.

$\longrightarrow$ This is basically the same problem as in the introductory example with SVM, lasso, etc.

Sources of the problems (ctd.)

- **Problem 2:** In methodological research we are interested in the error rate of the method for any data set, not just for the data set at hand.

$\longrightarrow$ Several data sets are needed.

$\longrightarrow$ In this context, nested cross-validation is not a perfect solution.

Some (partial) solutions

- Internal cross-validation?
  - → not for the methods' characteristics
  - → would not address the (most important) variability between data sets
- Check the superiority of the new method using other "validation" data sets. ... But the unbiased selection of appropriate data sets is a non-trivial task!
- Pay more attention to the substantive context.
- Publish negative results? (Boulesteix, Bioinformatics 2010)

Thanks for your attention!

Thanks to V. Guillemot, M. Jelizarow, K. Strimmer (University Leipzig), C. Strobl, A. Tenenhaus (Ecole Supélec).

The papers:

- M. Jelizarow, V. Guillemot, A. Tenenhaus, K. Strimmer, A.-L. Boulesteix, 2010. Over-optimism in bioinformatics: an illustration. *Bioinformatics* 26:1990–1998.
- A.-L. Boulesteix, 2010. Over-optimism in bioinformatics research. *Bioinformatics* 26:437–439.
- A.-L. Boulesteix and C. Strobl, 2009. Optimal classifier selection and negative bias in error rate estimation: An empirical study on high-dimensional prediction. *BMC Medical Research Methodology* 9:85.