Apprentissage statistique en grande dimension et données issues de la biologie moléculaire

MÉMOIRE D'HABILITATION À DIRIGER LES RECHERCHES

soutenu le 4 novembre 2011 à l'Université d'Evry Val d'Essonne
devant un jury composé de

Pr. Christophe Ambroise	
Pr. David Causeur	
Pr. Florence d'Alché-Buc	*rapporteuse*
Pr. Yves Grandvalet	*rapporteur*
Pr. Bernard Prum	
Pr. Stéphane Robin	*rapporteur*
Prof. Dr. Korbinian Strimmer	

par

Anne-Laure Isabeau Boulesteix

16 novembre 2011

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

Remerciements

Mes remerciements vont tout d'abord à Christophe Ambroise, qui a eu la bienveillance de me soutenir dans ma démarche d'habilitation externe. Merci beaucoup Christophe. Je remercie ensuite Stéphane Robin, Yves Grandvalet et Florence d'Alché-Buc d'avoir bien voulu être mes rapporteurs ainsi que David Causeur, Anestis Antoniadis, Bernard Prum et Korbinian Strimmer d'avoir accepté d'être membres de mon jury.

Je remercie également les personnes qui m'ont fait confiance en me permettant de travailler au sein de leur groupe/institut dans de bonnes conditions, en particulier Korbinian Strimmer qui a largement contribué à éveiller ma vocation de chercheuse. Un merci tout particulier aussi au directeur de l'institut de d'informatique médicale, biométrie et épidémiologie de la LMU, Ulrich Mansmann, qui me permet de travailler aujourd'hui dans un cadre particulièrement stimulant, de façon indépendante tout en bénéficiant de son soutien.

J'ai eu le plaisir de collaborer avec de nombreuses personnes durant ces années post-doctorales, et je tiens à remercier ici tous mes co-auteurs, partenaires de coopération et étudiants. Sans eux, ce mémoire ne serait pas ce qu'il est, et j'ai eu avec chacun d'eux beaucoup de plaisir à travailler. Merci aussi aux autres collègues de l'institut de statistiques de la LMU, de l'institut de statistiques médicales et épidémiologie de la TUM, du Sylvia Lawry Centre et de l'institut d'informatique médicale, biométrie et épidémiologie de la LMU d'avoir contribué à créer une environnement de travail agréable. Je suis également reconnaissante à toutes les personnes de l'UEVE qui m'ont réservé un bel accueil et m'ont aidée dans l'organisation du jour J.

Enfin, je tiens à remercier mes parents et mon mari qui m'ont soutenue tout au long de ma jeune carrière, ainsi que Victor, Marius et Emil qui me communiquent quotidiennement leur joie de vivre et leur énergie.

Table des matières

1 Introduction **3**

1.1 Biostatistician ? 3

1.2 The content of this synthesis 4

2 Over-optimism in statistical learning **5**

2.1 Supervised classification with high-dimensional omics data 5

2.2 Spoilt for choice ? 12

2.3 Optimistic bias of the optimal error rate 15

2.4 Optimization in methodological research 22

3 Relative importance of predictors **33**

3.1 Preliminary 33

3.2 Added predictive value of high-dimensional data to low-dimensional clinical data 34

3.3 Bias in decision trees and random forests 43

3.4 Extreme values and their influence 48

3.5 Research perspective : "Top-top-ranking" variables and very sparse models 52

4 Concluding remarks **55**

4.1 New methods versus other projects 55

4.2 Dirty details ? 57

Chapitre 1

Introduction

1.1 Biostatistician ?

I define myself as a biostatistician with focus on both methodological statistical research and biomedical applications. As such, I am trying to address some problems encountered in biometrical practice from a statistical point of view. This includes the development of new statistical methods to address unsolved or partially unsolved problems, but also the critical assessment and the comparison of existing methods or practices. My main research area being the statistical analysis of high-dimensional "omics" data, I am working at the intersubsection between statistics, computational medicine and bioinformatics.

For the sake of simplicity and consistency, I always denote researchers like me as "statisticians" and the cooperation partners – who produce the data and are interested in the results on their data set – as "biomedical scientists". Biomedical scientists are interested in the results of the statistical methods, while statisticians are interested in the methods themselves. Note that many (academic) researchers with statistical background, for instance statistical consultants, are not interested in the methods themselves but rather in the results they produce on data from the biomedical cooperation partner or consulting customer. I take this point of view in some of my application projects. This part of my activities, however, is not be described in this thesis.

1.2 The content of this synthesis

The assessment of prediction models and the related problem of over-optimism in applied statistical research are among the subjects I am interested in. Prediction models are a key concept in biometrical practice. An important part of medical research is devoted to prediction problems in a broad sense : improved diagnostic, assessment of personalized disease risk, prediction of the response to a therapy, to cite only a few examples. Prediction problems are supervised problems. From a statistical point of view, supervised problems have the advantage that they allow the statistical analyses to be objectively assessed in real data analysis by confronting the obtained predictions with the truth. This holds both for methodological research (development of new prediction methods) and for applications (development of concrete prediction models for biomedical cooperation partners). An issue that is relevant to most fields of applied statistical research in general and prediction models in particular is over-optimism of the assessment of performance. In the first section of this manuscript, I give an overview of prediction problems with high-dimensional omics data and the assessment of their performance with a particular emphasis on over-optimism issues.

Another problem related to prediction models is the assessment of the relative importance of the predictors included in these models. This problem can also be seen as supervised, because the importance of predictors can be measured in terms of prediction accuracy. However, it is not as easy as it might seem at first view to define the "importance" of a predictor. Plenty of definitions are conceivable, making the development and assessment of importance measures a highly complex matter. Four different aspects of this issue are addressed in the second section of this work : the problem of the added predictive value of a high-dimensional group of predictors, various types of bias affecting importance measures derived from decision trees and ensemble methods, the influence of extreme values on the assessment of predictors, and the selection of a very small number of predictors for clinical applications.

A large part of these two sections is adapted from articles published in international referred journals as a first or supervising author.

Chapitre 2

Over-optimism in statistical learning

2.1 Supervised classification with high-dimensional omics data

Settings

Although over-optimism issues are essentially relevant to all areas of applied statistics, we focus our attention on the special case of supervised classification with high-dimensional continuous data, which is particularly affected by this kind of problems.

Let us consider a data set with p "omics" variables as predictors, including a total of n observations corresponding to independent patients. In this synthesis, the term "omics variables" may refer, e.g., to the expression levels of transcripts or to the abundance of metabolites or proteins. For simplicity, we assume that these variables are all continuous and that the number p of these variables is large, hence the term *high-dimensional omics data*. For example, in microarray gene expression studies, the number of predictors p is huge compared to n (typically, $5000 \leq p \leq 50000$ and $20 \leq n \leq 300$), which makes standard statistical prediction methods inapplicable.

The term *response class* refers to the categorical variable that has to be predicted based on omics data. It can be, e.g., the presence or absence of disease, the tumor subtype, or the response to a therapy (responder/non-responder). The number of classes may be higher than two, although binary class prediction is by far the most frequent case in practice. Omics data may also be used to predict survival times, ordinal scores or continuous parameters. In this overview and in most of the following subsections, however, we focus on the special case of a binary response denoted as Y and taking values 0 and 1.

From a statistical point of view, supervised classification can be described as follows (the framework and notations presented here are borrowed from the technical report by Bernau et al. (2011)). On the one hand we have a finite set of classes whose labels are denoted by $\mathcal{Y} = \{0, .., C-1\}$ where C represents the number of different classes, here $C = 2$. On the other hand we have predictors taking value in $\mathcal{X} \subset \mathbb{R}^p$ that are used for constructing a decision function. Predictors and response class follow a joint distribution on $\mathcal{X} \times \mathcal{Y}$ denoted by $P(\mathbf{x}, y)$. The observed i.i.d. sample of size n is denoted as $s_0 = (\mathbf{x}_1, y_1)...(\mathbf{x}_n, y_n)$. In our case the data are high-dimensional, which mean that p may be much larger than n. The classification task consists in building a decision function $\hat{f}$ that maps elements of the predictor space $\mathcal{X}$ into the response space $\mathcal{Y}$:

$$\hat{f} : \mathcal{X} \mapsto \mathcal{Y}$$

$$\mathbf{x} \mapsto \hat{f}(\mathbf{x}).$$

Many classification methods have been proposed in the literature to derive such a function f from the available sample in the so-called "$n \ll p$" setting. Most of them involve a parameter that adjusts their complexity to the sample at hand. A brief overview of such methods is given in the next subsection. From now on, we denote as "methods $1, \ldots, K$" the considered combinations of methods and tuning parameter values. As an example, method 1 may stand for SVM with cost$=1$, method 2 for SVM with cost$=10$, method 3 for kNN with $k = 5$ neighbors, and so on. As a special case, methods $1, \ldots, K$ might represent different parameter values of the same method. No matter the signification of methods $1, \ldots, K$, the decision function obtained by fitting the prediction method k to the sample s_0 is denoted as $\hat{f}_k^{s_0}$.

Different approaches to the dimensionality problem

Hundreds of classification methods have been suggested in the biostatistics, bioinformatics and machine learning literature to derive a decision function $\hat{f}$ in high-dimensional settings. It is out of the scope of this introduction to give an extensive overview of these methods. Instead, let us give an overview of four common approaches to the dimensionality problem as described in more details in Boulesteix et al. (2008b). This four approaches partly overlap and can sometimes be combined. However, they are helpful to briefly sketch the different types of approaches.

Variable selection
The most intuitive approach consists of first selecting a small subset of predictors and then applying a traditional classification method to the reduced data set. By traditional methods, we mean well-known statistical methods handling a rather limited number of predictors, such as discriminant analysis methods including linear and quadratic discriminant analysis or Fisher's linear discriminant analysis, classical logistic regression or k-nearest-neighbors. In principle, the latter could be applied to a high number of predictors but performs poorly on noisy data.

Variable selection methods can be classified as *univariate* and *multivariate* approaches. Univariate approaches consider each predictor separately : they are based on the marginal utility of each predictor for the classification task. Predictors are ranked according to some criterion reflecting their association to the phenotype of interest. After ranking, the first predictors of the list are selected for further analysis. Many criteria are conceivable, for instance usual test statistics like Student's t-statistic or nonparametric statistics such as Wilcoxon's rank sum statistic. In the context of differential expression detection, several regularized variants of the standard t-statistic have been proposed in the last few years.

Univariate methods are fast and conceptually simple. However, they do not take correlations or interactions between predictors into account, resulting in a subset of predictors that may not be optimal for the considered classification task. This is obvious in the extreme case where, say, the 10 first predictors correspond to the same transcript, yielding a strong correlation structure. It is then suboptimal to select these 10 redundant predictors instead of predictors with a worse univariate criterion value but giving non-redundant information.

Multivariate variable selection approaches for microarray data have been the subject of a few tens of rather theoretical articles. They take the preceding argument seriously that the subset of the predictors with best univariate discrimination power is not necessarily the *best subset of predictors*, due to interactions and correlations between predictors. Therefore, multivariate variable selection methods do not score each predictor individually but rather try to determine which combinations of predictors yield high prediction accuracy. A multivariate variable selection method is characterized by i) the criterion used to score the considered subsets of predictors and ii) the algorithm employed to search the space of the possible subsets, an exhaustive enumeration of the 2^{p-1} possible subsets being computationally unfeasible. A very simple algorithm to address ii) consists in ranking the predictors and selecting the top of the list. For example, this is the algorithm recommended when using the essentially multivariate "cat-score" approach by Zuber and Strimmer (2009) which is based on the discriminative power of the predictors in the context of discriminant analysis.

Variable selection is the subject of the papers Boulesteix (2007) and Bernau and Boulesteix (2010). Boulesteix (2007) suggests a fast algorithm for efficiently ranking predictors according to the p-values of the two-sample Wilcoxon-test in different learning sets, while Bernau and Boulesteix (2010) investigate the effect of a preliminary variable selection on the internal cross-validation performed for parameter tuning.

Dimension reduction

A major shortcoming of variable selection when applied in combination with classification methods requiring the sample size n to be larger than the number p of predictors is that only a small part of the available information is used. For example, if one applies logistic regression to a data set of size $n = 50$, the model should include at most about 10 predictors, which excludes possibly interesting candidates. Note that a model based on few predictors may be preferred in some cases. For instance, a practitioner might be interested in an extremely sparse model for application in clinical settings. In this context, 10 predictors are better than 100 predictors, because they can be measured more easily in the clinical lab. We discuss this point more extensively in Subsection 3.5. In many cases, however, more than 10 predictors include interesting information that we want to take into account while derive a prediction rule.

Moreover, correlations between predictors are not taken into account by univariate variable selection approaches and can even pose a problem in model estimation, the more as omics data are known to often be highly correlated. An option to circumvent these problems is dimension reduction, which aims at "summarizing" the numerous predictors in form of a small number of new components (often linear combinations of the original predictors). Well-known examples are Principal Component Analysis (PCA), Partial Least Squares (PLS, Boulesteix, 2004; Boulesteix and Strimmer, 2007) and its generalizations (Fort and Lambert-Lacroix, 2005; Ding and Gentleman, 2005). A concise overview of dimension reduction methods that have been used for classification with microarray data is given in Boulesteix (2006b).

After dimension reduction, one can basically apply any classification method to the constructed components, for instance logistic regression or discriminant analysis. However, as opposed to the original omics predictors, the components constructed with dimension reduction techniques themselves may not be interpretable any more. Note that dimension reduction methods may be used in combination with variable selection, i.e. applied on a preliminarily selected subset of predictors. Preliminary variable selection is especially beneficial to unsupervised dimension reduction methods.

Penalization and shrinkage methods
Instead of reducing the data to a small number of (either constructed or selected) predictors, statistical methods based on penalization or shrinkage methods essentially perform an intrinsic regularization, i.e. they determine internally which predictors should be given more importance in the final decision function. This category includes penalized logistic regression (Zhu and Hastie, 2004), the Prediction Analysis of Microarrays (PAM) method based on shrunken centroids (Tibshirani and Efron, 2002), Support Vector Machines (SVM) (Vapnik, 1995) or regularized linear discriminant analysis (Guo et al., 2007). Such methods usually involve one or several penalty or shrinkage parameter(s) reflecting the amount of regularization. Note that they can be used in combination with variable selection. For example, in huge dimensional data including, say, 10^6 predictors, methods such as penalized regression might benefit from a pre-filtering of the predictors, at least in terms of computation time and storage requirement.

Recursive partitioning and ensemble methods
Some algorithms borrowed from the machine learning communities also perform an intrinsic variable selection, although in a completely different way. Classification trees (Breiman et al., 1984) perform an intrinsic variable selection since they select the most relevant predictor at each split. Random forests (Breiman, 2001), which are based on the combination of a large number of these trees, can also be viewed as a method performing intrinsic variable selection. More generally, the idea of ensemble methods is to combine predictions from a large number of relatively simple prediction rules called “base learners”. Base learners may be, e.g., classification trees or linear models with only one predictor. They are usually

based on a relatively small number of predictors – that have been somehow selected. In this sense, such methods are related to variable selection. Note that ensemble methods have partially unknown behaviors which are the subject of recent or current research projects (Strobl et al., 2007b, 2008; Boulesteix et al., 2011a).

Estimating the error

Suppose we have estimated a decision function $\hat{f}$ using any of the classification methods outlined in the above subsection. The true error that is to be estimated can be written as

$$\varepsilon[\hat{f}] = \mathbf{E}_P \left[L\left(\hat{f}(\mathbf{x}), y\right)\right] = \int_{\mathcal{X}\times\mathcal{Y}} L\left(\hat{f}(\mathbf{x}), y\right) dP(\mathbf{x}, y), \qquad (2.1)$$

where $\mathbf{E}_P$ stands for the mean over the joint distribution P and $L(.,.)$ is an adequate loss function. In this synthesis, we consider the indicator loss function, yielding the error rate, but most concepts can be generalized to other loss functions such as, e.g. the area under the curve (AUC). The rest of this subsection and Subsection 2.1 are largely borrowed from the technical report by Bernau et al. (2011).

For notational simplicity, the true error $\varepsilon[\hat{f}_k^{s_0}]$ of method k is denoted as $\varepsilon[\hat{f}_k^{s_0}] = \varepsilon(k \parallel S = s_0)$. Note that $\varepsilon(k \parallel S)$ should be seen as a random variable, where S stands for a random sample that follows the distribution P^n. It is usually denoted as *conditional error*, where the term “conditional” refers to the considered sample S. The mean

$$\varepsilon^n(k) = \mathbf{E}_{P^n}\left[\varepsilon(k \parallel S)\right]$$

of the random variable $\varepsilon(k \parallel S)$ is usually denoted as the *unconditional* true error rate of method k. It depends only on the method k, on the size n of the sample S and on the joint distribution P. Unlike the conditional error $\varepsilon(k \parallel S)$, the unconditional error $\varepsilon^n(k)$ is not a random variable in our framework.

Since the joint distribution $P(\mathbf{x}, y)$ is unknown in practice, the conditional errors $\varepsilon(1 \parallel S = s_0), \ldots, \varepsilon(K \parallel S = s_0)$ and the unconditional errors $\varepsilon^n(1), \ldots, \varepsilon^n(K)$ have to be somehow estimated. Standard estimation approaches are based on cross-validation (CV) or repeated subsampling, see Boulesteix et al. (2008b) for an overview. For notational simplicity, we focus on the repeated subsampling method in this synthesis, although many of the discussed concepts can be directly generalized to classical CV. These resampling procedures are implemented in our package 'CMA' (Slawski et al., 2008) in combination with tuning and variable selection, that are always performed based on learning data only.

In repeated subsampling, the whole data set is randomly split into a learning set and a test set a large number of times, e.g. $B = 100$ times. Each learning set L_b, $b = \{1, \ldots, B\}$ of size n_L (with $n_L < n$) is used to estimate a decision function that is subsequently evaluated on the corresponding test set $S \setminus L_b$. For each iteration $b = \{1, \ldots, B\}$ and each method k, $k = \{1, \ldots, K\}$, one obtains an estimated error :

$$e(k \parallel L_b, S \setminus L_b),$$

where the term $L_b, S \setminus L_b$ means that method k was fitted to the learning set L_b and evaluated on the test set $S \setminus L_b$. Note that we use the notation e for estimators and ε for true errors. For each method k, these test errors are eventually combined into an error rate estimate by averaging over the iterations $b = 1, \ldots, B$, yielding

$$e(k \parallel S) = \frac{1}{B} \sum_{b=1}^{B} e(k \parallel L_b, S \setminus L_b), \tag{2.2}$$

which obviously may depend on the random choice of the partitions $\{L_b, T_b\}$, $b = 1, \ldots, B$. Note that $e(k \parallel S)$ is an unbiased estimator of $\varepsilon^{n_L}(k)$ but an upwardly biased estimator of $\varepsilon(k \parallel S)$ and $\varepsilon^n(k)$, because the decision functions are estimated based on n_L observations instead of n, with $n_L < n$.

The "best" method

Let us further denote the method yielding the smallest error rate based on S as method $k^*(S)$, i.e.

$$k^*(S) = \arg\min_k e(k \parallel S). \tag{2.3}$$

Note that the random variable $k^*(S)$ depends not only on the sample S but, in general, also on the considered learning sets L_b, $b = 1, \ldots, B$. In the following considerations, we ignore the variability due to the choice of the learning sets and assume that they are fixed (this may be the case when all possible partitions are considered).

For a given sample s_0, the error estimate $e(k^*(s_0) \parallel s_0)$ obtained by repeated subsampling is a downwardly biased estimator of the true conditional error $\varepsilon(k^*(s_0) \parallel s_0)$, because $k^*(s_0)$ is chosen based on s_0, i.e. such that $e(k^*(s_0) \parallel s_0)$ is minimal. In other words, if one simply chooses the method yielding the minimal error rate $(e(k^*(s_0) \parallel s_0))$, this minimal error rate underestimates the true error rate $(\varepsilon(k^*(s_0)) \parallel s_0)$ of the chosen method. This bias is related to the problem of multiple comparisons. If several classifiers are tried, the chance that one of them randomly performs well rises. This bias can also be seen as the result of the variability of the estimates $e(k||S)$. In an ideal scenario with very large sample sizes and low variability, no optimization would take place, and the method $k^*(S)$ with minimal $e(k \parallel S)$ would also be the method with minimal true conditional error $\varepsilon(k \parallel S)$ and with minimal true unconditional error $\varepsilon^n(k)$. A quantitative assessment of this bias in the special case of binary classification with small sample high-dimensional gene expression data is given in Boulesteix and Strobl (2009) and summarized in the Subsection 2.3.

2.2 Spoilt for choice ?

Multiplicity of methods

In this part, we take the perspective of a statistician whose task is to analyse a concrete data set in the context of a collaboration with biomedical scientist. In most common statistical problems, a broad range of more or less appropriate methods is available to the informed user. The number of “candidate” methods that may be chosen usually depends on :

The explorative character of the study
Whereas statisticians usually restrict to a small number of well-established standard procedures in confirmative studies, the willingness to try recent experimental approaches increases with the explorative character of the study. Clinical trials, whose analysis is subject to guidelines from the regulatory authorities, are an extreme example of studies with a conservative profile. In practice, the range of available methods is restricted to those methods that have good chance to pass the regulatory processes. At the other extreme, exploratory studies of descriptive nature are the ideal place for experimenting new procedures. They are correspondingly characterized by a wide range of potential methods. Between these two extremes, many parameters play a role, including the biomedical scientist’s personality and his readiness to explore new horizons beyond well-established standard methods, the research area’s conventions and culture, or, from a purely strategic point of view, the editorial line of the target journals.

The history of the considered type of data
The choice may be very limited for a completely new data type and a novel

research problematic. Conversely, the user often has to choose between a large number of methods if the statistical community had the time and opportunity to work on the considered question. An extreme example of the latter situation is supervised classification with high-dimensional data, that has been the subject of hundreds of methodological publications in bioinformatics, (bio-)statistics and machine learning journals and conference proceedings in the last ten years. In this context, the space of potential methods is virtually infinite.

In parallel, the community tends to establish standards and guidelines as time goes by. In an ideal world, these standards are the results of well-done comparative studies and consensus from independent teams. However, other factors might contribute to promote a particular method, including the reputation of the authors, the renown of the journals the method was published in, well-documented and user-friendly implementations or an application of this method in one of the few leading scientific journals that other scientists tend to imitate. A statistician cannot reasonably support the strategy consisting to choose a method solely based on a previous publication (that may be erroneous or consider data with different characteristics), or based on the expected reaction of reviewers who may systematically reject new approaches or conversely systematically recommend complex methods over more simple procedures. However, methods used in a previous publication have the major advantage that they allow a more consistent comparison between the two studies. And from a pragmatic point of view, it is natural to take into account the expected reaction of reviewers in addition to purely scientific considerations, all other things being equal.

No matter how standards are established and whether the process is satisfying or not, the natural increase of the number of available is in practice counterbalanced by the emergence of non-written and vaguely defined standards and guidelines. When plotted against time, the number of available methods that can potentially be used in a concrete study finally often takes the form of an umbrella. The number of potential methods first increases, especially in the case of a scientific "hype". It subsequently decreases as standards are established, since it becomes harder and harder to justify the use of a non-standard "exotic" approach. This mechanism is described in the next paragraph in the special case of supervised classification for high-dimensional data.

Supervised classification for high-dimensional data

As already mentioned above, hundreds of methods for supervised classification with high-dimensional have been suggested in the literature by statisticians, computer scientists or biomedical scientists with computational background. At the beginning of the “microarray era”, i.e. in the early 2000s, virtually any method would probably have gone through the review process of most journals – if applied correctly (or not...) to a good study. The number of candidate methods was steadily increasing. Later on, in the middle 2000s, some methods grew to standard methods – for various reasons that we have already briefly mentioned. Drawing a list turns out to be highly subjective task and all lists may be subject of controversy. Let us only mention a few well-known examples like Lasso logistic regression, nearest shrunken centroids and other forms of linear discriminant analysis, or support vector machines.

Nowadays, it is probably more difficult than 5 or 10 years ago to publish an analysis using a non-standard “exotic” approach in a high-ranking biomedical journal. It would probably require a thorough justification which is essentially impossible to give, since it is extremely hard to show that the standard methods are less appropriate than the new exotic method. It is easy to show that a standard regression problem with three predictors departs from linearity and that a non-linear method like splines regression or fractional polynomials is thus more appropriate. But it is by far less easy, if not hopeless, to show that method A is more appropriate than method B in high-dimensional settings. Readers may legitimately wonder why the authors of the study of interest used this strange non-standard method. Because they know only this method ? This would indicate lack of expertise, and the study’s results should be considered with caution. Because they know/like this method much better than existing standard methods, perhaps even as developers of this new method ? It is admittedly a bad idea to use a method one is not familiar with. And it is natural to apply in practice an in-house method one has worked on for years. However, these arguments are certainly not sufficient to exclude standard methods with a well-known behavior that would allow easier comparison to previous studies.

Last but not least, the authors might have used a “strange non-standard” method simply because it yielded better results than standard methods. Again, this argument seems natural and at first view valid. However, such a strategy is extremely likely to hide a “fishing for significance” or a “data dredging” mechanism. By “fishing for significance”, we mean the strategy to try different methods and to select – or “fish” – for the most favorable results, for instance the lowest CV error rate in the special case of supervised classification. In the situation sketched above, the researchers would try various methods including a few standard methods, but find better results with a non-standard method, and report only these better results ! An article presenting classification results obtained with an experimental non-standard might thus be suspected of “data dredging” or “fishing for significance, except if it gives really good reasons for not using standard methods.

Note that data dredging and fishing for significance are by far not limited to studies using non-standard approaches, as we will quantitatively outline in the next subsection. Fishing for significance can occur in any study, provided that several more or less appropriate methods are available for the problem at hand.

2.3 Optimistic bias of the optimal error rate

An empirical study

In biometric practice researchers often apply a large number of different methods in a “trial-and-error” strategy to get as much as possible out of their data and, due to publication pressure or pressure from the consulting customer, present only the most favorable results. This is particularly true in the context of supervised classification with high-dimensional data, which is characterized by a lack of standards and guidelines.

Using the notations defined in subsections 2.1 and 2.1, it means that the researchers would report $e(k^*(s_0) \parallel s_0)$ only, where s_0 denotes the particular considered data set. Obviously, $e(k^*(s_0) \parallel s_0)$ is a downwardly biased estimate of the true unconditional error $\varepsilon_{n_L}(k^*(s_0))$ of the most favorable method $k^*(s_0)$.

We conducted an empirical study based on two real microarray data sets (colon cancer data by Alon et al. (1999) and prostate cancer data by Singh et al. (2002)) to assess this bias quantitatively in realistic settings (Boulesteix and Strobl, 2009). In our study we considered a total of 124 variants of classification methods (possibly including variable selection or tuning steps) within a 5-fold cross-validation evaluation scheme.

In the first study, the classification methods were applied to permuted versions of the two considered data sets that are simply obtained by randomly permuting

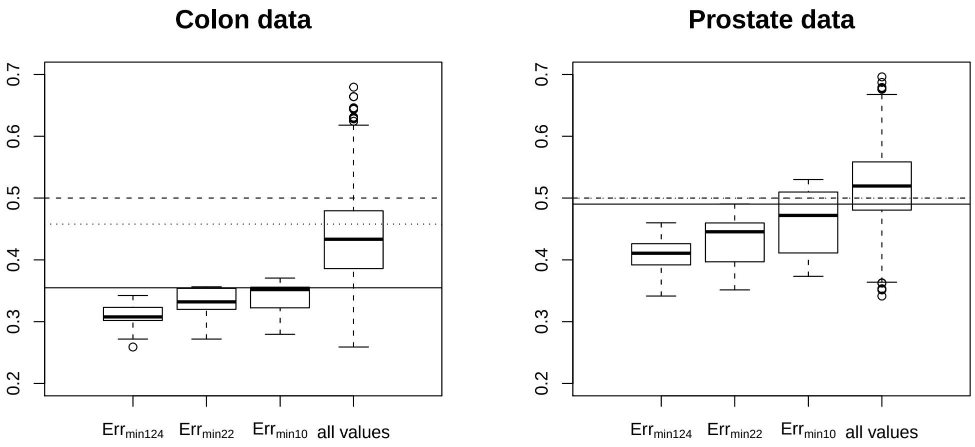

FIGURE 2.1 – Permutation-based analyses. Boxplots of the minimal error error over 1) 124 combinations of classification method, variable selection method and parameter value (1st boxplot), 2) 22 combinations of classification methods and variable selection method, parameter values being kept fixed (2nd boxplot), 3) 10 classification methods, variable selection and parameter values being kept fixed (3rd boxplot). For comparison, the fourth boxplot represents the errors estimated with all classification methods without optimization process. **Left :** colon cancer data. **Right :** prostate cancer data.

the response class in a balanced way. For 20 random permutations, Figure 2.1 displays boxplots of the minimal error over 1) 124 combinations of classification method, variable selection method and parameter value (first boxplot), 2) 22 combinations of classification methods and variable selection method, parameter values being kept fixed (2nd boxplot), 3) 10 classification methods, variable selection and parameter values being kept fixed (3rd boxplot). For comparison, the fourth boxplot represents the errors estimated with all classification methods without optimization process. Figure 2.1 clearly shows that the optimization bias cannot be ignored in practice. In the colon cancer data set, a median minimal error rate as low as 31% is obtained – although there is no association between response and variables !

On the difficulty of error rate estimation in $n \ll p$ settings

Over-optimism due to the various optimization mechanisms results from insufficient sample size. If the sample size were very large, we would not have any optimization bias. Optimization biases are observed because resampling-based error estimates have large unknown variance, and are even virtually uncorrelated with the actual error (Hanczar et al., 2007) in small sample settings. Thus, the methods/variable selection methods/parameter values yielding the smallest error rates with a particular data set do not necessarily have the smallest true error rates, hence the risk of over-optimization and the discrepancy between error rates obtained on training and validation data sets. This explains why optimization biases, which are relevant to all statistical research areas, particularly affect the analysis of small sample high-dimensional data.

The curse of the multiplicity of methods

The so-called *curse of dimensionality* has been abundantly discussed in biostatistics and bioinformatics literature in the context of high-dimension prediction. Various types of optimistic biases pointed out in the literature can be seen as the consequence of bad practice enhanced by the curse of dimensionality. For instance, performing variable selection using both learning set and test set can be seen as bad practice (Ambroise and McLachlan, 2002) but it has less dramatic consequences in low dimensions than in high dimensions. In this sense, the curse of dimensionality acts as a distorting mirror for bad practice.

The results presented in Boulesteix and Strobl (2009) and summarized in the last subsection show that the curse of dimensionality also combines with a mechanism that we denote as *curse of multiplicity of methods* to produce strongly biased error estimates. As outlined above, the multiplicity of methods itself is not sufficient to produce a bias. Without the high variability of error estimation, which itself results from the curse of dimensionality, there would be no bias.

Note that the bias of $e(k^*(s_0) \parallel s_0)$ as an estimator of $\varepsilon(k^*(s_0) \parallel s_0)$ can also be paralleled to the selection bias outlined by Ambroise and McLachlan (2002) in the sense that both biases result from a bad separation between the learning set and the test set. As far as the variable selection bias is concerned, the test data are used in the learning process to select variables. Similarly, when estimating $\varepsilon(k^*(s_0) \parallel s_0)$ by $e(k^*(s_0) \parallel s_0)$, one implicitly uses the test sets T_b to select the best method $k^*(s_0)$, i.e. the tests sets are used to choose the final decision function $\hat{f}^{s_0}_{k^*(s_0)}$.

From the empirical study by Boulesteix and Strobl (2009), it is clear that one should definitely not report only the best result $e(k^*(s_0) \parallel s_0)$, because this strategy generates a considerable optimistic bias. In practice, the bias due to the optimal selection of the tuning parameters of a fixed classification method is often addressed by nested cross-validation (Varma and Simon, 2006) as implemented in our Bioconductor package 'CMA' (Slawski et al., 2008). Within each cross-validation iteration, the best parameter value is determined based on an inner cross-validation procedure and the error rate is then computed for this parameter value. The final error rate estimate is obtained by averaging over the cross-validation iterations. Note that, doing that, one averages error rates obtained with different parameter values.

Going one step further, one could theoretically consider the classification method as a (nominally scaled) tuning parameter and also address the choice of the classification method using nested cross-validation. However, besides substantial interpretation problems, this approach would be extremely computationally expensive and difficult to apply in practice for methods involving a tuning parameter that also has to be tuned via inner cross-validation.

Research perspective : bias correction

In a current research project, Christoph Bernau aims at correcting the bias of $e(k^*(s_0) \parallel s_0)$ as an estimator of $\varepsilon(k^*(s_0) \parallel s_0)$, because, roughly speaking, we are interested in the expected performance of the "best method" $k^*(s_0)$ on independent data. The objective of this current project is two-fold. Firstly, we develop a new method for tuning bias correction by embedding the tuning problem into a decision theoretic framework. The method is based on the decomposition of the unconditional error rate involving the tuning procedure. The corrected error estimator can be written as a weighted mean of the errors obtained using the different parameter values. It can be interpreted as a smooth version of nested cross-validation, where the weighting scheme guarantees intuitive bounds for the corrected error. Secondly, we suggest to also use bias correction methods – like our new approach, NCV, or the approach by Tibshirani and Tibshirani (2009) – to address the bias resulting from the optimal choice of the classification method, which has never been done in the literature to our knowledge. The results suggest that the news correction method yields similar results as NCV but at a much lower computational price. This approach is presented in the technical report by Bernau et al. (2011).

The estimation of the variance of the unconditional error rate plays a crucial role in this correction procedure. It is a non-trivial problem (Bengio and Grandvalet, 2004) without completely satisfying solutions to date. We also aim to examine this estimation problem more precisely, to compare existing estimator, and to assess the impact of variance estimation procedures on the comparison of classification methods in methodological research.

Research perspective : other types of bad separation between learning set and test set

Prediction models aim to make predictions for future patients. These future patients are not yet in the available data set. To estimate the error rate on future patients, e.g. by repeated subsampling (see Subsection 2.1), one thus has to derive decision functions on the learning sets L_b without seeing T_b (for $b = 1, \dots, B$). The separation between L_b and T_b should be perfect. In particular, variable selection should be performed using the learning sets L_b only (Ambroise and McLachlan, 2002) as already discussed above. Similarly, one should not estimate $\varepsilon(k^*(s_0) \parallel s_0)$ as $e(k^*(s_0) \parallel s_0)$ because it would imply that the test sets T_b are used both for the estimation of the error $\varepsilon(k^*(s_0) \parallel s_0)$ and for the selection of the decision function $\hat{f}^{s_0}_{k^*(s_0)}$. The same data are thus used to choose the decision function and to estimate its error : separation between learning data and test data is not perfect. These are only two examples of bad separation among others. Many important separation problems have not been studied extensively in the literature.

For example, it is common practice to normalize microarray data using the whole available data set. In the context of repeated subsampling, almost nobody would imagine normalizing the data anew at each of the B iterations using the learning set L_b only. And even if we would do that, how should be normalized the test set T_b subsequently ? Normalizing the test set T_b separately is a bad idea if T_b is small, and impossible in the case of leave-one-out cross-validation. Approaches for normalizing the test data using the normalization parameters estimated from the learning data have been recently proposed (Kostka and Spang, 2008), but they are not yet well-established. In practice, all statisticians who estimate the error rate via repeated subsampling normalize their microarray only once, using the whole data set. Doing so, they implicitly use the test sets in the learning process at each subsampling iteration, since the test set may affect the decision function via the normalization procedure.

Other similar examples are pre-processing steps such as the imputation of missing values, scaling of the predictors, or dichotomization of the predictors. Such pre-processing steps are almost always performed using the whole data set in practice and not at each subsampling iteration anew. It is unclear whether this “bad separation” of learning set and test set affects the estimation of the error rate, and if yes, how we should proceed to accurately estimate the error rate that we would obtain using *independent* data that have not been seen before – neither by the statistician and nor by the algorithm.

The case of unsupervised learning

Over-optimism in the context of unsupervised statistical problems is even more difficult to address, because the truth is unknown. Whereas one may compare the minimal error rate of 31% in the permuted data sets to the expected error rate of 50%, such a study cannot be performed in this form for unsupervised problems.

However, optimization mechanisms may also be at work in unsupervised problems, and they may be even more dangerous since it is less easy to quantify the potential bias. Let us consider the example of different criteria used to rank predictors according to their association with an outcome of interest. A typical example is the assessment of differential gene expression across two conditions for thousands of genes whose expression levels are measured using microarrays. Plenty of ranking criteria have been used or/and suggested to rank the genes in these context, including the straightforward p-value of the two-sample t-test, ranked-based approaches, or various regularized versions of the t-statistic. Some of these statistics are implemented in user-friendly software, such that an informed data analyst might apply several criteria to his/her data set in a very short time. Most importantly, these criteria usually yield very different lists of top-genes in small sample high-dimensional settings, as extensively discussed in Boulesteix and Slawski (2009).

The multiplicity of available methods may lead to a substantial optimistic bias in this context in the sense that some researchers might choose their ranking criterion in an iterative optimization process. For example, if the “favorite feature” is not identified as top-ranking by a particular criterion, biomedical scientists are more likely to urge their statistical consultant to try out another statistic than if they find their favorite feature in the top-list. By favorite feature, we mean for instance a feature that is expected to be relevant based on biological knowledge, the feature that is expected to rank best based on a previous study, or conversely a feature which has not yet been identified as relevant in this context and may thus yield an innovative ground-breaking marker.

The same holds in the more complicated case of network analyses. A plethora of methods are available for network reconstruction from high-dimensional data. Similarly to the problem of predictor ranking even similar variants of the same method often yield completely different results. For instance, in the paper by Krämer et al (2009) we consider a specific family of methods all based on the graphical gaussian networks (GGM) methodology. The considered methods can be seen as different estimation procedures for the same underlying model. Despite this common framework, the five considered methods (based on the shrinkage estimator of the covariance, on Lasso regression, on adaptive Lasso regression, on ridge regression, and on Partial Least Squares regression respectively) yield substantially different results, i.e. a moderate overlap of the sets of common edges.

Such differences may also indirectly lead to a bias, in the sense that the biologist is more likely to report in his/her paper a network that, e.g. shows a characteristic hub around a particular transcription factor of interest. In the previous subsection, we have seen that the choice of the prediction method through optimization of the estimated error rate may yield substantially optimistic results due to the high variability of the procedures. Similarly, if the network reconstruction method is chosen a posteriori (because it yields a network that is in some way “more interesting” than the networks output by other methods), this result is in a way also optimistically biased. More concretely, suppose that four network reconstruction methods identify no hub and the fifth one identifies a hub. The strategy consisting to choose the method that identifies the hub just because it identifies a hub is biased. Similarly to the case of supervised learning, this bias essentially grows with the between-method variability and with the variability of estimation.

Similar problems may also occur in clustering analyses. Although this is often ignored (at least in final published papers), clustering analyses are known to be highly instable in the sense that different clustering algorithms may yield completely different results and that the results tend to be sensitive against changes in the data set. In a current research project with a master student, I am interested in the quantification of this instability via resampling analyses. More precisely, the student is developing resampling-based importance measures for each object and each variable. By “object”, we simply mean the objects that have to be clustered, while in this context the term “variable” denotes the variables characterizing the objects.

2.4 Optimization in methodological research

Optimization mechanisms and biases as outlined in Subsection 2.3 may also be at work in methodological research, although this aspect is by far less acknowledged in the literature.

Evaluation criteria for newly developed methods

The main goal of methodological statistical research is the development of new methods. Other types of works, like comparison studies and reviews, are considered as less exciting and less useful by many researchers and by most journal editors. Note that this is in contrast with biomedical research, where review articles or meta-analyses are usually considered as extremely valuable research outputs. In this subsection, we focus on the research projects dealing with *new* statistical methods, and we illustrate the presented ideas based on examples referring to the development of new prediction methods, in particular variants of discriminant analysis.

New methods are supposed to outperform existing methods in some sense. Which criteria should be used to evaluate new methods ? This apparently simple question does not have any unique answer, all the more since the answers strongly depend on the specific problem at hand, on the considered sub-area of applied statistics, and on the field of application. Indeed, each community, each team and even each researcher may develop its own non-written rules.

From a theoretical point of view, properties such as e.g. the bias, consistence or variance of an estimator, the power of a statistical test, or the mean and variance of the error rate of a statistical learning method are important criteria. These theoretical properties can often merely be proven under (very) restrictive assumptions, for instance when the size of the sample goes to infinity and/or when the distribution of the random variable of interest fulfills specific requirements. From a practical point of view, agreement between the assumptions and the reality of the data in the considered field plays a major role. A theoretical property may be considered as a strong advantage in a field of application where the assumptions are fulfilled, but less interesting in another field. Assumptions are often not fulfilled in omics data analysis because of high-dimensionality, high correlation between the variables, etc. It is of course advantageous to know the behavior of a new method under simplifying assumptions. All other things being equal, a new method fulfilling theoretical properties under unrealistic assumptions is certainly better than a new method with theoretically unknown behavior. The question whether one should favor theoretical properties over other advantages is almost a philosophical one and strongly depends on the considered context. We do not intend to answer this question here but rather to give an overview of possible criteria for judging a new method.

From a pedagogical point of view, one should not forget that the method is destined to be used... by experts or non-expert users. Therefore, the simplicity of a new method constitutes an important advantage, all other things being equal. A complicated "good method" is unlikely to be correctly understood and used by an audience of non-specialists – and sometimes even by specialists. Users will thus more likely give up using it and settle for a more simple (but possibly inappropriate) method. They are also more likely to use the new complicated method incorrectly, for instance i) because they do not correctly understand the implied assumptions and use it in a case where it should not be used, ii) because they make an implementation error, or iii) because they choose the parameters inappropriately. In this sense, all other things being equal a simple method should be preferred to a more complicated method. Note that this criterion is rarely invoked in methodological articles presenting new methods. In some sub-areas of applied statistics, simplicity might rightly or wrongly raise suspicion.

The absence or small number of critical parameters is a particular aspect of simplicity. On the one hand, it limits the risk that inexperienced users choose parameters inadequately. On the other hand, it limits the risk that developers and expert users overfit the data they analyse by optimizing the parameters based on the final results of the different parameter combinations. The latter problem is related to the "fishing for significance" problem discussed in Subsection 2.3.

From a technical point of view, particular attention may be devoted to computational aspects such as computation time, storage requirements, the influence on initial values in an iterative algorithm, or more generally the dependence on a random generator. Users might give up using a "good method" because of a few successive computer bugs due to memory requirement problems or because they cannot afford to wait five days for the results (they would probably wait if they had to, but they would not wait if they just wanted to "try" the method as one of several candidates).

Other important aspects include the ability to generalize to other related problems and, conversely, the embedding of the method in a more general framework. Such criteria are important because related methods often share related behaviors or related characteristics, and because results obtained within the same framework are often easier to compare. The ability to generalize and to fit into a more general framework is also important from the pedagogical point of view mentioned above. That is because potential users are more likely to understand and use a method correctly if they are familiar with another method within the same framework.

Last but not least, new methods are of course evaluated based on objective criteria telling us us whether they do what they were developed for. In the context of supervised classification, it can for instance be the error rate or the area under the receiver operating characteristic (ROC) curve. In the context of hypothesis testing, one may assess the power under various kinds of alternative hypotheses. These criteria are most often assessed in simulations, especially in the context of statistical problems with "unknown truth", for instance the ranking of variables. In the context of statistical learning for omics data, however, simulations usually do not play such a major role. That is because simulation designs are often believed to poorly reflect the complexity of the data. In this context, it is common practice to apply the newly developed method to one or several "real data sets", for instance data sets extracted from public data repositories. A better error rate is then usually considered as an argument in favor of the new method. Indeed, many abstracts contain a sentence like "our method performed better than existing methods on real data".

On the one hand, this approach for assessing a new method seems natural : all researchers are more interested in methods that "work" than in methods that "do not work". However, this approach also raises important epistemological problems that are surprisingly widely ignored in the literature. Firstly, in a testing perspective, two or three data sets as often considered in practice are not sufficient to establish the superiority of a new method. Secondly, the term "performs better" is widely used in the literature but not well-defined from an epistemological point of view. These problems require further research, especially in the context of high-dimensional data. I will try to address them in future research. Thirdly, quantitative criteria are often affected by a strong optimistic bias in the same spirit as the bias discussed in Subsection 2.3. Virtually all methodological research projects might end up with biased results in the sense that the results are tuned so as to optimize the quantitative criterion of interest. We will see an example of such optimization mechanisms in Subsection 2.4, and present quantitative results based on a concrete example in Subsection 2.4. These two subsections are adapted from the article by Jelizarow et al. (2010).

Sources of bias in methodological research

The reported results on the performance of new statistical learning algorithms are known to be over-optimistic, as discussed in my letter to the editors of *Bioinformatics* (Boulesteix, 2010) in the particular context of statistical bioinformatics. Several sources of bias can be identified.

The first and perhaps most obvious reason for over-optimism is that researchers sometimes randomly search for a specific data set such that their new method works better than existing approaches, yielding a so-called "data set bias". While a method cannot reasonably be expected to yield "universally better" results in all data sets, it would be wrong to report only favorable data sets without mentioning and/or discussing the other results. This strategy induces an optimistic bias. This aspect of over-optimism is quantitatively investigated in the study by Yousefi et al. (2010) and termed as "optimization of the data set" here.

The second source of over-optimism, which is related to the optimal choice of the data set mentioned above, is the optimal choice of a particular setting in which the superiority of the new algorithm is more pronounced. For example, researchers could report the results obtained after a particular feature filtering which favors the new algorithm compared to existing benchmark approaches. This mechanism, which is strongly related to data overfitting, is termed as "optimization of the settings" here.

The third source of over-optimism is related to the choice of the existing benchmark methods applied for comparison purposes. Researchers are supposed to compare their new algorithm to state-of-the-art methods, but may consciously or subconsciously choose suboptimal existing methods and exclude the best competing methods from the comparison for any reason, e.g. because running the software demands very particular knowledge, because previous authors excluded these methods as well, because the methods induce high computational expense or because they belong to a completely different family of approaches and thus do not fit in the considered framework. Then the new algorithm artificially seems better than competing approaches and over-optimistic results on the superiority of the new algorithm are reported – because the best competing approaches are disregarded. Since the definition of state-of-the-art methods is often ambiguous, such problems may occur even when researchers are decided to perform a fair comparison. This mechanism, also known as "straw-man phenomenon" is termed as "optimization of the competing methods" here.

Finally, researchers often tend to optimize their new algorithms to the data sets they consider during the development phase (Boulesteix, 2010). This mechanism essentially affects all research fields related to data analysis such as statistics, machine learning, or bioinformatics. Indeed, the trial-and-error process constitutes an important component of data analysis research. As most inventive ideas have to be improved sequentially before reaching an acceptable maturity, the development of a new method is per se an unpredictable search process. The problem is that, as stated by the *Bioinformatics* editorial team Rocke et al. (2009), this search process leads to an artificial optimization of the method's characteristics to the considered data sets. Hence, the superiority of the novel method over an existing method (as measured, e.g. through the difference between the cross-validation error rates) is sometimes considerably overestimated. In a concrete medical prediction study, fitting a prediction model and estimating its error rate using the same training data set yields a downwardly biased error estimate commonly termed as apparent error. In the same spirit, computing cross-validation error rates with different classifiers and systematically selecting the classifier variant with the smallest error rate yields a substantial optimization bias (Boulesteix and Strobl, 2009). Similarly, developing a new algorithm (i.e. selecting one of many variants) and evaluating it by comparison to existing methods using the same data set may lead to optimistically biased results in the sense that the new algorithm's characteristics overfit the used data set. This source of over-optimism is termed as "optimization of the method's characteristics" here.

The four mechanisms discussed above may lead to over-optimistic conclusions regarding the superiority of the new method compared to existing methods. The importance of validation with independent data has recently gained much attention in biomedical literature. For instance, we refer to the empirical study by Daumer et al. (2008) which points out the usefulness of a pre-publication validation strategy based on data-splitting. To our knowledge, no such study was performed in the context of statistical research and this issue is largely ignored in the literature.

An empirical study

We performed an empirical study to fill this gap (Jelizarow et al., 2010). It reviews and illustrates the problem of validation and false research findings through a concrete example from a current research field : the incorporation of prior biological knowledge on gene functional groups into high-dimensional microarray-based classification. The "promising idea" we pursue here is to extend the shrinkage correlation estimator of Schäfer and Strimmer (2005) to incorporate prior knowledge on gene functional groups with the aim to improve the performance of linear discriminant analysis. This approach combines a simple and well-established statistical method, regularized discriminant analysis, with the incorporation of prior biological knowledge on gene functional groups, a popular concept that has attracted a lot of attention in the last few years.

While this method does not yield any improvement in terms of prediction error rate, it is straightforward to produce over-optimistic results via any of the four mechanisms discussed above. Based on this example, we demonstrate quantitatively that optimization of the data set, optimization of the settings, optimization of the competing methods and, most importantly, optimization of the method's characteristics can lead to substantially biased results and over-optimistic conclusions on the superiority of the new method. This study is deliberately of empirical nature. We neither model the different sources of over-optimism theoretically nor do we derive analytical expressions of the resulting bias for simplified situations, because we feel it would not reflect the complexity of the addressed mechanisms. Instead, we stick to concrete observations to illustrate what consciously or subconsciously happens in virtually all methodological projects – possibly including our own projects. We are convinced that most biased results are presented by mistake and that the involved researchers are disposed to make efforts towards better practice.

In a nutshell, we consider 10 variants of our new statistical learning algorithm together with 12 preliminary variable selection procedures. These methods are applied to four real microarray data sets. For each data set successively, we select the variant and the variable selection scheme yielding the smallest estimated error rate. This estimated error rate is typically small, since it is the result of an optimization. If the same combination of variant and variable selection is applied to another of the four data sets, however, the obtained error rate of this combination is not small anymore, as can be seen from Table 2.1. These results illustrate the optimization mechanisms discussed above (optimization of the settings and optimization of the method's characteristics).

Prediction method	CV error in data set 1	CV error in data set 2	CV error in data set 3	CV error in data set 4
optimal variant according to data set 1	**0.025**	0.180	0.345	0.152
optimal variant according to data set 2	0.079	**0.129**	0.363	0.141
optimal variant according to data set 3	0.029	0.221	**0.342**	0.115
optimal variant according to data set 4	0.033	0.274	0.384	**0.078**

Table 2.1 – Cross-validation (CV) error of the optimal variant/variable selection scheme selected in each of the four data sets. For example, the variant yielding the smallest CV error rate 2.5% in data set 1 yielded an error rate of 18% in data set 2. The figures outside the diagonal can be understood as "validation error rates".

Other sources of bias

As illustrated above, the four investigated sources of over-optimism may yield substantially over-optimistic results. Beyond these four mechanisms, various other sources of over-optimism may also affect the reported results. For instance, one might optimize the evaluation criterion : the sensitivity and specificity may yield other results than the error rate, especially in case of strongly unequal class sizes. The applied normalization technique may also affect the results and yield optimization potential. Another indirect source of over-optimism is related to technical problems : if an implementation problem occurs with the competing approaches and slightly worsens their results, researchers often tend to spontaneously accept these inferior results. Conversely, they would probably obstinately look for the error if such problems occur with their new algorithm. Note that the validation strategy recommended here would not help in this case, since the error in the competing methods would also affect the validation phase.

On cross-validation (CV) as a potential solution

Our results demonstrate that validation based on independent data sets (i.e. looking at the error rates outside the diagonal in Table 2.1) avoids hasty over-optimistic conclusions and automatically corrects for the optimization of the settings and optimization of the methods' characteristics. A natural question is whether a CV procedure (or related approach) might be used in place of validation with independent validation data.

CV is useful to choose the best number of genes and the best variable selection scheme for each method considered in the comparison study. Such a CV correctly addresses the "optimization of the settings" mechanism and is sometimes used in methodological studies, as recommended in Ambroise and McLachlan (2002) for the number of genes. From a theoretical point of view, CV could also be applied to select the methods' characteristics (i.e. to select among the 10 considered variants of the new method). In this case, however, the application of a CV procedure is much more problematic because the different variants of the new method are usually not investigated simultaneously in practice. Researchers typically begin with the most intuitive variant. Having realized the latter's sub-optimality (e.g. in terms of error rates) they investigate a few alternative variants, which often requires up to several months. Presenting first results at conferences often leads to fruitful discussions with other researchers, resulting in further variants of the original method, and so on.

While the ten variants are considered simultaneously in the present study, this process typically drags on in practice, and the variants are investigated rather successively than simultaneously. Therefore, researchers cannot be expected to perform an internal CV to choose between variants they have explored (and rejected) at the beginning of their project. An advantage of validation with fresh data over cross-validation is that it ensures a more stringent separation between data used for development and data used for evaluation. Cross-validation might be incomplete in practice, for instance if researchers forget some of the variants they have tried some time ago. In statistical learning terminology, we would say that they select a "tuning parameter" (here : the methods' characteristics) using the whole training set instead of repeating the selection procedure in each iteration. Such human errors cannot occur if validation is performed with a fresh data set after having developed a method. Moreover, validation based on other independent data sets has the considerable advantage that it takes the variability between data sets into account, a very important aspect discussed in the next subsection.

Finally, CV induces substantial computational expense. Using a complete embedded CV procedure involving three layers to i) estimate the error rate, ii) select the number of genes, the variable selection scheme and some additional tuning parameters of the method internally, iii) select the best variant of the method (among the 10 candidate variants) internally rapidly becomes computationally intractable and, in general, cannot be recommended in practice.

On the difficulty of error rate estimation

Most importantly, over-optimism due to the various optimization mechanisms results from insufficient sample size. If sample sizes were in the hundreds of thousands, the problems would be solved because they result from imprecision of the error estimates (Yousefi et al., 2010; Hanczar et al., 2010). Optimization biases occur because cross-validation error estimates have large unknown variance (Braga-Neto and Dougherty, 2004). These estimates are even virtually uncorrelated with the actual error (Hanczar et al., 2007) in small sample settings. Thus, the methods/variants/settings yielding the smallest error rates with a particular data set do not necessarily have the smallest true error rates, hence the risk of over-optimization and the discrepancy between error rates obtained on training and validation data sets. This explains why optimization biases, which are relevant to all statistical research areas, particularly affect the analysis of small sample high-dimensional data.

The real problem is thus the absence of suitable means of error estimation based on a single data set. When comparing prediction methods, we would like to reject the “null hypothesis” that a newly proposed prediction algorithm has an error rate higher than or equal to the error rate of competing approaches. However, this possibility is killed at the outset by using CV on a single data set because the internal variance (i.e. the variance within a single data set) can be estimated but not the external variance (i.e. the variance between data sets). In a way, this external variance is taken into account when applying the algorithms to validation data. Note that the external variance could be potentially taken into account by using several training data sets. However, the estimation of external variability based on a small number of data sets is also a non-trivial issue.

On simulations as a potential solution

Another way to take this “between-data sets” variance into account is to perform simulation studies. However, while simulations are often extremely useful (Mehta et al., 2004), some aspects of the developed methods can only be evaluated through real data studies. A general problem in high-dimensional data analysis is that it is very difficult to generate realistic data sets. Our example of a new statistical learning algorithm incorporating prior biological knowledge can be seen as an extreme case, since it involves a complex cluster structure with clusters of different sizes that potentially overlap. An additional difficulty is that the performance of our promising idea essentially depends on two components : 1) the ability of the new method to adequately incorporate knowledge on the structure of the predictors into the statistical learning algorithm and 2) the quality of the available biological information. While simulations may address the first aspect at the price of simplifying assumptions on the data structure, the second aspect can only be assessed through real data studies. Finally, we point out that simulation studies are potentially also affected by conscious or subconscious optimization mechanisms.

What next ?

Applying the new method to “validation data sets” that have not been used during the development phase is certainly a good idea and may in some cases moderate hasty over-optimistic conclusions. However, it is important to note that 1) a method cannot reasonably be expected to always perform better than all other methods, 2) even if there were such a universally better method for a considered problem, (too) many validation data sets would be required to establish this superiority considering the high variability across data sets. Consequently, a “solution” to over-optimism problems in methodological research is probably to admit that, from an epistemological point of view, we simply cannot establish the superiority of a new method on real data. Real data analyses should probably be rather seen as illustrations or as proofs of concepts than as a strict evaluation criteria for the new methods. Last but not least, other criteria such as those discussed in Subsection 2.4 should probably be given more attention. Editors/reviewers may partially relax the request for performance improvement on real data so that researchers would be less encouraged to “fish for significance” (Boulesteix, 2010). Furthermore, I believe that the occasional publication of well-designed studies on promising sensible ideas with disappointing quantitative results may in the long run contribute to a less optimistically biased literature.

Chapitre 3

Relative importance of predictors

3.1 Preliminary

In biometric practice, the development of prediction models discussed in the previous section is only one part of the problem. Biomedical researchers are interested in prediction models, but they keep in mind that their prediction model has only poor chance to get applied in practice by other scientists. With a few exceptions, prediction models presented in biomedical publications are themselves less important than the predictors they include. In other words, the motivation of biomedical researchers is not only to obtain a good prediction model, but also to find out which predictors are more relevant to the prediction problem. That is what we denote as “relative importance” of the predictors in statistical learning.

At first view, this problem seems more easy to address than the development of a prediction model. After all, the development of prediction rules takes the relative importance of predictors into account. The determination of the relative importance of predictors in statistical learning, however, is not simply of a “subset” of the supervised learning task. Above all, it is a badly defined problem involving various issues related to statistical learning. It may have different meanings depending on the considered sub-field of applied statistics, on the field of application, on the investigated substantive research question and perhaps even on the philosophy of the statistician. In this part, we highlight the variety of concepts related to the relative importance of predictors in statistical learning by successively taking four different views illustrated by concrete methods and applications.

The focus of Subsection 3.2 is on added predictive value of high-dimensional (omics) predictors to low-dimensional (clinical) predictors. This problem relates to the relative importance of *groups* of predictors. In simple (intentionally vague) words, the question is “is group B interesting given that I already have group A ?”. We will see that this question is actually multifaceted, hence the vague formulation at this stage. Subsection 3.3 also addresses the relative importance of predictors in the context of supervised learning, but from a completely different point of view. The focus is not on groups, but on individual predictors, that are ranked according to their relevance for prediction in the context of recursive partitioning methods. Subsection 3.4 addresses the problem of extreme values that may affect the assessment of the importance of predictors. Finally, Subsection 3.5 introduces a future research project on the selection of very small subsets of predictors for clinical applications.

3.2 Added predictive value of high-dimensional data to low-dimensional clinical data

This subsection first gives an overview of the problem of added predictive value and its validation, following the lines of the recently published synthesis article by Boulesteix and Sauerbrei (2011).

Overview

While omics data such as microarray gene expression data have been used for disease outcome prediction or diagnosis purposes for more than ten years in biomedical research, the question of the added predictive value of such data given that classical clinical predictors are already available has long been underconsidered in the bioinformatics literature. This issue can be summarized as follows. For a given prediction problem (for example tumor subtype diagnosis or long-term outcome prediction), two types of predictors are considered. On the one hand, conventional clinical predictors such as, e.g. age, sex, disease duration or tumor stage are available as potential predictors. They have often been extensively investigated and validated in previous studies. On the other hand, we have omics predictors which are generally much more difficult to measure and collect than conventional clinical predictors that are of variable utility and often not well-established. In the context of translational biomedical research, biomedical scientists are often interested in the *added* predictive value of such predictors over classical clinical predictors. Clinical predictors may be given as a list of individual factors or in form of a well-established index such as the International Prognostic Index (IPI) for lymphoma or the Nottingham Prognostic Index (NPI) for breast cancer. Here we do not distinguish between the case of individual clinical predictors and the case of an aggregated index. From a statistical point of view, an aggregated index can be seen as a clinical predictor.

Strategies to derive combined prediction models

Prediction models combining clinical with omics data are important to assess the added predictive value of omics predictors. That is because some methods for assessing added predictive value are based on the comparison of the accuracy of prediction models with and without omics predictors. However, the concept of combined models is not clearly defined and different strategies have been adopted in the literature.

Strategy 1 ("naive")
The perhaps most naive approach consists in building a combined prediction model by treating clinical and omics predictors in the same way. This approach is very general. It can be applied to any prediction method that can handle predictors of the considered types, for instance a mixture of continuous omics predictors and categorical clinical predictors. In this approach, individual clinical predictors may "get lost" within the numerous omics predictors and thus not be fully exploited – especially when clinical information is available in form of a single aggregated score. If the clinical predictors have good predictive value, such naive prediction models are expected to underestimate the accuracy of combined models. The estimated added predictive value then tends to be small – not because the omics predictors are bad but because the combined rule does not fully exploit the clinical predictors (that are lost within a large amount of noise).

Strategy 2 ("residuals")
The other extreme strategy consists in deriving a fixed clinical prediction model, for instance using logistic regression or Cox regression. The resulting linear predictor is then considered as an offset and updated using omics predictors,

for instance via lasso regression (Tibshirani, 1996) or boosting regression (Bühlmann and Hothorn, 2007). This approach yields a linear predictor in which the coefficients of the clinical predictors are not affected by the omics predictors. It is adequate to test the added predictive value (Boulesteix and Hothorn, 2010) since the focus is here on the residual variation of the outcome. However, it may be sub-optimal in terms of prediction accuracy. Depending on the correlation between clinical and omics predictors, accuracy may be improved by adapting the coefficients of the clinical predictors (Binder and Schumacher, 2008).

Strategy 3 ("favoring")
An intermediate strategy between strategies 1 and 2 is to fit a prediction model to clinical and omics predictors simultaneously while somehow "favoring" the clinical predictors, since they are more or less "established" prognostic factors. A comparative study of some of these approaches is given in Bovelstad et al. (2009) in the context of survival prediction. For instance, clinical predictors might be favored in terms of prior in Bayesian settings or through a different penalty in penalized regression. The R package **penalized** (Goeman, 2010) provides an implementation of L_1 and L_2 penalized regression with so-called unpenalized coefficients. Such methods, in particular L_2 penalized regression, have been shown to perform well in terms of prediction in a comparative study on survival prediction from combined models (Bovelstad et al., 2009). In the same vein, the CoxBoost approach (Binder and Schumacher, 2008) forces clinical predictors into the prediction model.

Strategy 3 better exploits the predictive potential of clinical predictors than Strategy 1, since they are "favored" in the model building process. In contrast to strategy 2, however, the influence of clinical predictors in the prediction model is affected by omics predictors. A critical question is how much clinical predictors are/should be favored. Obviously, that should depend on clinical knowledge. It is difficult to give clear recommendations on this heterogeneous family of methods. If clinical predictors are much favored, Strategy 3 is similar to Strategy 2 and the prediction accuracy of the combined model is possibly sub-optimal. If they are not enough favored, however, Strategy 3 has the same pitfall as Strategy 1.

Strategy 4 ("dimension reduction")
Dimension reduction approaches constitute an important special case of methods favoring clinical predictors. We are considering them separately here. They include methods like the PLS+RF procedure (standing for Partial Least Squares followed by Random Forest) (Boulesteix et al., 2008a) discussed in further details

in Subsection 3.2 or the supervised principal component approach (Bair and Tibshirani, 2004), that are all based on two successive steps. The omics predictors are first summarized in form of new components in a dimension reduction step. A prediction rule including these new components and the clinical predictors as covariates is then built using, e.g. the classical Cox model (Bovelstad et al., 2009) or random forests (Boulesteix et al., 2008a). A critical aspect of these methods is over-fitting that should be avoided while constructing the new components. For instance, over-fitting can be avoided through a pre-validation procedure applied to the dimension reduction step (Boulesteix et al., 2008a; Tibshirani et al., 2002). Otherwise, the new components are likely to be strongly correlated with the outcome even in the case of non-informative omics predictors, and thus yield sub-optimal combined models.

Validation of added predictive value using an independent data set

Validation of prediction models using independent data is important from a clinical point of view, because it measures the accuracy of the prediction model based on a possibly different patient population and thus assesses its generalizability. Model calibration may be required in this context (van Houwelingen, 2000). Good discrimination in new data is an important pre-requisite for a good prediction model. In the context of translational research, the *validation of added predictive value* is perhaps even more important than the validation of the prediction accuracy of the prediction model. Some approaches have been proposed for assessing added predictive value based on a single training data while avoiding overfitting problems. In this subsection, however, we address the assessment of added predictive value based on independent validation data. Note that, from a technical point of view, an independent validation data set can be generated artificially from a large data set by random splitting. Compared to data from a new setting (external data) this internal validation approach has disadvantages. External validation is a more stringent procedure necessary for evaluating whether the predictive model will generalize to populations other than the one on which it was developed. In the following we review/present four approaches (A,B,C,D) for the assessment the added predictive value of omics predictors in a validation data set.

Approach A

A first possible approach (denoted as "approach A" in the following) is to fit two prediction models based on the training data : a clinical prediction model and a combined prediction model. Note that the combined prediction model should be fitted using strategies 2, 3 or 4. Otherwise the results cannot be correctly interpreted. The two models are then applied to make a prediction for the observations from the validation data set, and the predicted and true outcomes are compared for both models. Depending on the type of outcome (right-censored time-to-event or class) and on the point of view of the researcher, different assessment criteria are available. For time-to-event outcomes the (integrated) Brier-score and related methods such as prediction error curves (Graf et al., 1999; Gerds and Schumacher, 2006) are popular measures, but other may also be used depending on the main focus of the study. The problem of the choice of a suitable measure to assess the added value is similar for other approaches like approach D discussed below, that is also based on the accuracy of prediction models. For class prediction we can compute, e.g., the specificity, the sensitivity or the error rate of the two prediction models and compare them using standard statistical tests.

Approach B

In some cases, however, prediction models resulting from the training phase cannot be directly applied to the validation data. For instance, this may be the case if the training data set was collected within a case-control design while the validation data set stems from a population study with a (much) smaller percentage of cases. The probabilities output by the prediction model from the training phase do not make sense for the validation data set. In this case, re-calibration may be considered. Another option is to validate the discriminative ability of the score underlying the prediction model rather than the prediction model itself. A perhaps more characteristic example for which it makes sense to consider the discriminative ability instead of the prediction is the case of omics predictors that are measured at a different scale in the training and validation data sets. For instance, gene expression may have been measured using microarrays in the training set but using the low-throughput reverse transcription quantitative polymerase chain reaction (RT-PCR) technique in the validation set. The unit of measurement is then not the same for the two data sets. It thus makes no sense to apply the model coefficients derived from the training set to the validation set. In this case, it may be useful to look at the values of the score in the validation data and its association with the outcome ("approach B") rather than at the accuracy of the prediction model. One then needs criteria to assess and compare the scores underlying the prediction models instead of the prediction models themselves. ROC curves including tests of equality of the area under the curve (AUC) or the c-index can be considered in the case of class prediction. For survival analysis the association between the two scores and the outcome can be assessed using Cox regression, for instance based on quantile survival curves or other measures of discriminative ability.

Approach C
Approaches A and B are not widely used in practice, probably because combined prediction models and combined scores are tricky and not yet well-established. Moreover, practitioners often prefer to establish their score in form of a omics score that does not involve clinical predictors. Last but not least, the required clinical predictors are sometimes not available for the training data. The two other approaches we are reviewing here are devoted to procedures that do not necessitate the use of combined scores. The training phase outputs solely a omics score whose added predictive value is then determined in the validation data set, thus taking into account the clinical predictors of the validation data.

This omics score may have been constructed while taking the clinical predictors of the training data into account or not. The SuperPC approach (Bair and Tibshirani, 2004) is an example of method deriving a omics score while taking the clinical predictors into account. The idea is to derive the omics score by applying principal component analysis to predictors that are correlated with the outcome in the training data *after adjustment for clinical predictors* (Bovelstad et al., 2009). No matter how the omics score is derived, we assume that it can be computed for all observations from the validation data set. It is in a way considered as a "new predictor".

The most natural way to assess the score's association with the outcome while adjusting for clinical predictors is to fit a prediction model based on the validation data using the omics score as well as the clinical predictors as predictors ("approach C"). One can then perform a suitable test to check whether the regression coefficient of the score differs significantly from zero. Since the score does not overfit the validation data set, this approach is unbiased in the sense that it does not systematically over-estimate the added predictive value of the omics predictors. It has been widely used in prognostic studies involving high-dimensional omics data (Metzeler et al., 2008; Yao et al., 2008).

However, approach C tells nothing about the predictive value in terms of prediction error. Furthermore, p-values get smaller with increasing sample size – independently of the gained prediction accuracy. As stated by Altman and Royston (2000) *"usefulness is determined by how well a model works in practice, not by how many zeros there are in the associated p-values"*. In other words, small p-values may be observed even if the gained prediction accuracy is poor. For a binary outcome Pepe et al. (2004) illustrate that in the case of binary classification the odds ratio of a binary marker has to be extremely high (e.g. 10 or more) in order to improve the performance of a classification rule substantially. Even a "large" odds ratio, e.g. 3, does not give sufficient strength for a suitable classification tool. They also discuss this issue in the context of the added value of a marker.

Approach D

Approach D is similar to approach C, but it consists in comparing the prediction accuracy of prediction models with and without omics score via cross-validation or related resampling methods rather than via significance testing. Therefore it addresses the important pitfall of approach C which was based on p-values only. Note that these prediction models can be constructed via logistic or Cox regression or by any other model building approach. Like in approach C, the omics score is considered as a new predictor. While approach C assesses this new predictor based on the p-value obtained in a multivariate regression, approach D explicitly evaluates the gain of accuracy yielded by the new predictor by cross-validation. More precisely, the validation data are divided into a number k of cross-validation folds, for instance $k = 10$. In the kth iteration, the kth fold is excluded from the data and two prediction models are fitted to the remaining $k - 1$ folds : one model with clinical predictors only and one model with both the score and the clinical predictors. The two models are applied to the kth fold and evaluated based on a suitable criterion like the Brier-score (Graf et al., 1999; Gerds and Schumacher, 2006) (for both survival analysis and class prediction), the error rate or the AUC (for class prediction only). Approach D is used in a paper by Herold et al. (2011) on the prediction of leukemia outcome.

Further variants

Different variants of approaches A, B, C and D may also be of interest in particular situations. For example, they may be applied in subgroups defined by clinical predictors, hence implicitly accounting for interaction effects between omics and clinical predictors. They can also be applied to fixed classical clinical scores already described in the literature instead of clinical scores fitted from the data. On the whole, a multitude of methods are conceivable and sensible, but they are often not described adequately in medical literature, and often even not used adequately. Furthermore, there has been no study on the respective behavior and merits of these methods. We are trying to fill this gap in a current project.

PLS-RF combined classifier based on PLS dimension reduction and random forests

The PLS-RF proposed by Boulesteix et al. (2008b) is one of the methods addressing the construction of combined models via dimension reduction ("strategy 4") as described in Subsection 3.2. The basic idea is to summarize omics predictors through Partial Least Squares (PLS) and to build a random forest (Breiman, 2001) using the clinical predictors and the PLS component(s) as predictors. Since PLS dimension reduction is per se supervised, the constructed PLS are expected to be strongly correlated with the response – even in the absence of true association between response and omics predictors ! This approach thus tends to over-evaluate the relative importance of omics predictors for the prediction problem.

To avoid this systematic bias, we suggest in Boulesteix et al. (2008b) to "pre-validate" the PLS components, i.e. to compute them within cross-validation. The data are first partitioned into non-overlapping subsets. At each cross-validation iteration, one of the subsets is excluded and PLS dimension reduction is performed on the rest of the data, considered as training data for this iteration. The PLS components are then computed for the excluded subset using the weights derived from the training data and denoted as *pre-validated* PLS components. Pre-validated PLS components are computed in the same way for all cross-validation iterations, i.e. for all observations in the data set. Finally, these pre-validated PLS components are used to build the random forest in place of the original (overfitting) PLS components.

The principle of pre-validation originally proposed outside the context of dimension reduction by Tibshirani and Efron (2002) can be virtually applied to any dimension reduction method and could benefit to other methods related to strategy 4 beyond the particular case of random forests and PLS.

Globalboosttest : a permutation-based test for added predictive value of high-dimensional data

In Subsection 3.2 we reviewed methods assessing added predictive value of high-dimensional omics data based on a validation data set. Methods to assess added predictive value of high-dimensional omics data based on a single training data set have also been proposed in the literature. The present subsection outlines one of these methods, a global testing approach, that was suggested by Boulesteix and Hothorn (2010). The general idea of global testing approaches is to test the coefficients of omics predictors in a generalized linear model based on both clinical and omics predictors. The linear predictor has the form

$$\eta = \beta_0 + \beta_1 Z_1 + \cdots + \beta_q Z_q + \beta_1^* X_1 + \cdots + \beta_p^* X_p,$$

where $Z_1, \ldots, Z_q$ stand for the clinical predictors while $X_1, \ldots, X_p$ are the omics predictors. In the considered setting, the number of omics predictors p tends to be much larger than the number of omics predictors q. In the example of logistic regression, the linear predictor η is linked to the probabilities of the two classes $Y = 0$ and $Y = 1$ through the logistic function. In Cox regression, the linear predictor corresponds to the hazard ratio. Global tests test the null-hypothesis

$$H_0: \quad \beta_1^* \ = \ \ldots \ = \ \beta_p^* \ = \ 0$$

i.e. that $X_1, \ldots, X_p$ have no added predictive value in the considered generalized linear model. The testing procedure described in the paper by Boulesteix and Hothorn (2010) addresses this hypothesis through a two-step procedure involving permutations. In the first step of this procedure, a logistic regression is fit to the clinical predictors $Z_1, \ldots, Z_q$ while ignoring the omics predictors $X_1, \ldots, X_p$. In the second step of the procedure, the linear predictor fitted in the first step is considered *as an offset*, and a combined model based on both $Z_1, \ldots, Z_q$ and $X_1, \ldots, X_p$ is fitted using a regularized regression technique for $n \ll p$ data. Several regularized regression techniques are conceivable for the second step. Boulesteix and Hothorn (2010) consider boosting regression (see Bühlmann and Hothorn, 2007, for a review), but possible alternative methods include, e.g., Lasso regression or L_2 penalized regression. After completing the second step, a model fit criterion (for example the likelihood) is computed for the combined model consisting of the offset from the first step and a term involving omics predictors. More details are given in Boulesteix and Hothorn (2010).

The second step is repeated a large number of times (e.g. $B = 1000$) after permutation of the predictors $X_1, \ldots, X_p$, while the clinical predictors are left unchanged and the offset thus also remains unchanged. Strictly speaking, permutation of the variables $X_1, \ldots, X_p$ corresponds to the null-hypothesis

$$H_0: \quad \beta_1^* \ = \ \ldots \ = \ \beta_p^* \ = \ 0 \ \text{ and } \ X_1, \ldots, X_p \perp\!\!\!\perp Z_1, \ldots, Z_q.$$

Hence, if we do not have $X_1, \ldots, X_p \perp Z_1, \ldots, Z_q$, such a permutation procedure could in theory yield rejection of the null-hypothesis even if $\beta_1 = \cdots = \beta_p = 0$. However, in our procedure the variables $Z_1, \ldots, Z_q$ enter the model in form of a fixed offset. In a simulation corresponding to the extreme case $X_1 = Z_1, \dots, X_5 = Z_5$ with $\beta_1^* = \cdots = \beta_p^* = 0$, the distribution of the p-value is found to be uniform – as desired since the omics predictors do not have any added predictive value in this case.

In various simulation settings, the power of our testing procedure is at least as good as Goeman's global test (Goeman et al., 2004, 2005), and superior in the case of a small number of strong omics predictors. The procedure is implemented in the publicly available R package **globalboosttest**. The major inconvenience of our procedure is the high computation time associated with the permutation scheme. However, like many permutation-based procedures it is easily parallelizable.

3.3 Bias in decision trees and random forests

While the previous subsection focused on the problem of added predictive of a high-dimensional block of variables given that other variables are already available, the present subsection addresses a completely different issue : the ranking of single predictors using recursive partitioning methods. The ranking of single predictors might at first view seem much easier than the assessment of groups of predictors. Properties of variable importance measures, however, are often not well-known beyond classical linear models. In particular, variable importance measures derived from decision trees become increasingly popular in many scientific fields including "omics" sciences, but a thorough investigation of their behavior in particular conditions often reveals more or less "hidden problems". By "problem", we mean in a very general sense that in simulations the variable importance measure does not rank the variables as we would have expected. More specifically, we address the problem of biases, which we define as follows. A bias occurs when a particular type of predictor is systematically ranked higher than another type of predictor – although none of the investigated predictors is associated with the response.

Decision trees and ensemble methods

Decision trees (Breiman et al., 1984) and random forests (Breiman, 2001) are attractive methods that become increasingly important in supervised learning applications. An advantage of random forests is that they yield, in addition to the prediction itself, a measure of variable importance (VIM) for each of the involved predictors. In contrast to standard measures of variable importance output by classical methods like linear models, random forest VIMs are not subject to particular assumptions, they may capture non-linear relationships between predictors and response, and they can be computed even with highly correlated $n \ll p$ data. In the eyes of some biomedical scientists, a random forest resembles a "magic blackbox" returning valid results whatever type of data you give them. Things are not so simple, however. Even though random forests present major advantages for the analysis of high-dimensional complex and not well-characterized data, there are still many open questions regarding their behavior in non-standard and even in standard settings. Examples are presented in the rest of this subsection.

Bias and number of candidate splittings in predictors in decision trees

The first and perhaps most obvious problem of standard decision trees is their bias in favor of predictors with more potential splits. Before each splitting, a standard tree as defined in the seminal book by Breiman et al. (1984) considers all possible splits of all predictors, and selects the predictor and the split yielding the best value of a particular splitting criterion. In the case of a binary response and trees with binary splits, the most widely used criterion is the so-called Decrease of Gini Impurity (DGI), which can be shown to be equivalent to the chi-square statistic in terms of predictor ranking (Grabmeier and Lambe, 2007).

Imagine we want to predict a binary disease status based on two predictors : "sex" and "month of birth". The binary predictor "sex" is obviously more likely to be associated with the disease status than the predictor "month of birth", because many diseases affect more men than women or conversely. In a small data set, however, the nominally scaled predictor "month of birth" might be more likely to yield a good split by chance. A nominally scaled predictor with as many as 12 categories yields $2^{12} - 2$ potential splits ! Obviously, the multiplicity of potential splits advantages this predictor in the sense that, under the null-hypothesis of no association between predictors and response, it is much more likely to get selected and to be used for splitting than a predictor with less potential splits.

There are various ways to cope with this problem. In the case outlined above, an obvious solution would be to consider a single global χ^2 test for $k \times 2$ instead of several χ^2 tests for 2×2 tables. This approach considers the predictors globally instead of examining all the possible splits of the predictors successively. Another approach is to examine all the possible splits successively and select the best split for each predictor, *but* to address the resulting multiple testing issue by adjusting the splitting criterion within each predictor to take into account the fact that the split is selected optimally among several possible splits.

For the sake of clarity, we assume from now on that the chosen splitting criterion has to be maximized, i.e. a higher value of the criterion corresponds to a better split. Adjustment can be done based on the distribution of the maximally selected criterion under the null hypothesis of independence between the response and the predictor. The original splitting criterion is then replaced by the "probability that the maximally selected criterion exceeds the observed value under the null-hypothesis of independence between the response and the predictor".

Procedures to derive the exact distribution of maximally selected chi-square statistics
This probability, however, may be difficult to derive. Enumeration-based methods rapidly become intractable with increasing sample size. Permutation-based methods are also time consuming, especially if one wants to compute small p-values precisely. Asymptotic methods can be applied to large samples only, a condition that is obviously not fulfilled in classification trees involving small leaves at the bottom of the tree. This motivated the development of computationally efficient procedures for deriving the exact distribution of maximally selected chi-square statistics under the null hypothesis of independence between response class and predictor.

The twin papers Boulesteix (2006a) and Boulesteix (2006b) suggest such procedures that can be applied to the case of a binary response class and a categorical predictor with K categories. The paper Boulesteix (2006b) addresses the easier case of an ordinally scaled predictor, i.e. the K categories are naturally ordered. In this case, the suggested method to derive the exact distribution of the maximally selected chi-square statistic can be seen as a generalization of Koziol's method (Koziol, 1991). The other paper (Boulesteix, 2006a) is itself a generalization of the method for ordinal predictors and, roughly speaking, considers all possible orderings of the categories successively. Both procedures are conditional in the sense that the distribution is computed conditionally on the marginal frequencies of the binary response and ordinal/nominal predictor. They are both based on combinatorial considerations.

In the same vein, a method for deriving the asymptotical distribution of the maximally selected chi-square statistic in the case of two interacting ordinal predictors is presented in Boulesteix et al. (2007). In particular, this method can be applied to the assessment of SNP-SNP interactions in the context of genetic association studies with a binary outcome.

The method for ordinal predictors (Boulesteix, 2006b) can also be applied to metric predictors with ties. In fact, metric predictors with ties are treated in the same way as ordinal predictors in decision trees. In particular, they both have strictly less than $n-1$ possible cutpoints, in contrast to metric predictors without ties that have exactly $n-1$ possible cutpoints.

The case of missing values

Metric predictors without ties, which are very common in omics data analysis, have the advantage that they all have the same number $(n-1)$ of possible cutpoints. This is an advantage in terms of the variable selection bias outlined above : none of the candidate predictors is systematically favored. In the case of missing values in the predictors, however, things get more complicated. On the one hand, predictors with many missing values yield less possible cutpoints, thus reducing the chance to get a good cutpoint by chance. This effect tends to favor predictors without missing values compared to predictors with many missing values. On the other hand, the splitting criterion (for instance the chi-square statistic) has an increased variance in the case of missing values because of the reduced sample size. This increased variance tends to increase the maximal value of the criterion over the possible cutpoints.

Hence, we face two contradicting effects : while the "multiple comparisons effect" tends to favor predictors with few missing values, the "variance effect" tends to favor predictors with many missing values. All in all, simulations presented in Strobl et al. (2007b) show that predictors with many missing values are selected much more often, suggesting that the variance effect strongly dominates the multiple comparisons effect. Note that this difference in variable selection frequency is observed even for predictors with values missing completely at random. This is in contrast with most properties on missing values with classical statistical methods, where missing values do not produce any bias provided the values are missing completely at random. Moreover, the identified bias is in the "wrong direction". Indeed, all other things being equal one might prefer to select predictors with less missing values, because their assessment is based on more data and thus more reliable, and because a high number of many missing values may reflect, e.g., measurement problems that might also occur for future observations.

Bias and bootstrap sampling in random forests

To address the variable selection bias discussed in the previous subsection, Hothorn et al. (2006) propose a new class of random forests based on unbiased criteria. Describing the whole procedure in detail would go beyond the scope of this subsection. Let us just say that, at each split, the splitting predictor is selected according to the probability that the maximal criterion over the candidate splits exceeds the observed value under the null-hypothesis of no association to the response. The predictor with the smallest probability is selected, and the best split based on this predictor is used for splitting. This procedure is unbiased in the sense outlined above. Under the null-hypothesis, predictors with many categories do not have more chance to get selected that predictors with few categories.

However, simulations presented in Strobl et al. (2007a) show that, against their expectations, predictors with many categories are selected more often for splitting in random forests and receive higher Gini variable importances than predictors with few categories even if none of the predictors is associated with the response. Further simulations show that this systematic bias in favor of predictors with many categories can be removed by constructing the trees based on subsamples of the original data set rather than on bootstrap samples. In other words, the bias disappears if we draw the data without replacement instead of with replacement. Since there are to our knowledge no inconveniences in the use of subsamples in this context, we consequently propose to always construct random forest trees based on subsamples. In a future research project, I intend to investigate the “bootstrap bias” from a more general point of view independently of the random forest application.

Bias in random forests variable importance favoring variables with approximately equally sized categories

Random forests have recently grown to a standard statistical analysis tool in genetics. They are used in many genetic studies to rank candidate single nucleotide polymorphisms (SNPs) with respect to their association with a disease or trait of interest via the variable importance measures (VIM), or to investigate the prediction power of genetic data while possibly taking into account complex non-linear patterns. As outlined in Subsection 3.3, the Gini VIM is biased in favor of categorical variables with more categories (Strobl et al., 2007a). In the context of genetic association studies, some authors argue that this kind of bias is not relevant, because most SNPs - if not all - have three categories (“AA”, “Aa” and “aa”). Further, Calle and Urrea (2011) point out that the Gini VIM shows better stability than the permutation VIM in resampling analyses and consequently recommend the use of the Gini VIM. However, in a subsequent study on the stability of VIMs, Nicodemus (2011) shows that the higher stability of Gini VIM compared to permutation VIM may be due to the fact that the Gini VIM is biased in favor of SNPs with high minor allele frequency (MAF).

However, the main focus of Nicodemus’ study is not the bias in favor of large MAFs but rather the stability, performance and behavior of VIMs in various settings. In a recent paper (Boulesteix et al., 2011a), we present a study with the three following goals : 1) to assess the bias in favor of large MAFs quantitatively, independently of stability issues, for different variants of random forests and VIMs, 2) to highlight the mechanisms leading to this bias, 3) to summarize the known properties of random forest variable importance measures in the context of SNP data analysis and to make recommendations regarding the choice of the random forest variant.

3.4 Extreme values and their influence

Omics data are known to include many outlying values that may considerably affect the results of high-level analyses, but this problem has surprisingly not focused much attention in the literature. It is the subject of the paper Boulesteix et al. (2011b), which is summarized in this section. In the microarray literature, the term “outlier” most often refers to outlying arrays. In the context of class prediction, mislabeled arrays are a special type of outliers. Such mislabeled arrays can be seen as outliers with respect to their class and can often be easily detected in the context of prediction, since they are usually consistently misclassified by standard classification algorithms.

The problem of extreme values in the predictors is a different one, and probably more difficult to handle than outlying samples in high-dimensional settings. In the context of differential gene expression, a few methods have been proposed that allow to identify genes with extreme values. For instance, Tibshirani and Hastie (2007) suggest the related “outlier sum” (OS) statistic for identifying genes with extreme values. The focus of these procedures is on the identification of genes that have *several* extreme values in one of the two groups (the cancer group in their example), in the hope that these extreme values form a biologically interesting subgroup.

While subgroups are often assumed to be "biologically interesting", single extreme values may rather be the consequence of undesired events such as measurement error, technical errors in the lab, or the particular characteristics of a single patient that are not relevant to the disease. The standard traditional way to cope with such single outliers in multivariate analyses is to detect them and put them aside before performing further analyses, as already recommended by Edgeworth (1887) in the context of least squares regression : *"The method of least squares is seen to be our best course when we have thrown overboard a certain portion of our data – a sort of sacrifice which has often to be made by those who sail upon the stormy seas of Probability"*. The problem is that, in the context of high-dimensional data analysis, it is most often impossible to "throw overboard" all the observations that have an extreme value for at least one feature. Doing that, we would eliminate too many if not all observations.

Robust statistical procedures form the second main family of methods handling extreme values, as summarized by Rousseeuw and Leroy (2003) in low-dimensional settings. Many statistical methods can be adapted to achieve robustness against extreme values. However general agreement and clear guidance are still missing. The field is still in its infancy as far as high-dimensional data are concerned. A further problem is that robust methods have then to be used at all stages of the analysis, i.e. we would have to use, say, a robust t-statistic for differential gene expression analysis, a robust prediction method for multivariate model building, a robust method for the estimation of correlation graphs, etc. While it may be easy to compute a robust t-statistic, methods addressing the other issues are far less developed.

In Boulesteix et al. (2011b) we suggest to apply a very simple logistic-like transformation to each feature and to compare the results of statistical analyses before and after the transformation. This transformation was originally proposed by Royston and Sauerbrei (2007) to smoothly "switch off" extreme values in the context of regression with fractional polynomials. If x_{ij} denotes the value of the jth predictor for the ith individual, the transformed version x^*_{ij} of x_{ij} (for $j = 1, \dots, p$) is given as

$$x^*_{ij} \;=\; \left[\ln\left(\frac{\Phi(z_{ij}) + \epsilon}{1 - \Phi(z_{ij}) + \epsilon}\right) + \varepsilon^*\right] / (2\varepsilon^*), \qquad (3.1)$$

where Φ stands for the standard normal cumulative distribution function, z_{ij}

is simply the standardized form of x_{ij} (i.e. $z_{ij} = (x_{ij} - \overline{x}_j)/s_j$, with $\overline{x}_j$ and s_j denoting the sample average and standard deviation of X_j, respectively), ε is a parameter, and $\varepsilon^* = \ln[(1+\varepsilon)/\varepsilon]$. Royston and Sauerbrei (2007) recommend the choice $\epsilon = 0.01$.

Let us consider the example of a ranking in a two-group setting, e.g. differential expression analysis using the t-statistic. In our analyses using nine real microarray data sets, we first compute the ranking of the features before transformation, then transform the data and compute the ranking again. Features showing a high discrepancy between the rankings before and after the transformation are typically those with extreme values. We find features with such outlying values in many of the investigated data sets. The discrepancy between the results before and after the transformation is sometimes dramatic, especially in our analyses implementing multivariate model selection. To sum up, we feel that outlying values probably deserve more attention than they usually receive during statistical analyses. Moreover, the standard pre-processing techniques that we use in our study (RMA, VSN, GC-RMA and MAS5) produce data sets that still contain extreme values with strong influence on the results of statistical analyses.

Our somewhat "naive approach" identifies extreme values and proposes a simple way to handle them. If no extreme value is identified (as measured by the discrepancy between the ranks before and after the transformation), it confirms that extreme values do not seriously influence the univariate ranking of the relevant features. In many of the data sets of our study (mainly with a survival outcome), however, extreme values induce the selection of features that would not have been selected otherwise or conversely lead us to overlook important features. Ignoring extreme values in statistical analyses can be seen as a simple and common error (Baggerly and Coombes, 2009). In the context of multivariable analyses differences between models with and without the transformation are larger. This will hardly influence prediction accuracy, but may lead to different interpretations of the selected models. The specific implications of differences in results have to be discussed with subject-matter knowledge in mind. Since the transformation is easy to apply, we propose to use it at least as a sensitivity analysis. This is in line with the proposal in the original paper (Royston and Sauerbrei, 2007) where the transformation was developed to improve robustness of fractional polynomial functions.

Of course, one could argue that extreme values may have biological relevance and should not be switched off. However, they first have to be identified anyway.

The transformation with the corresponding rank discrepancy measure can be used for this purpose. It could be interesting to investigate why the considered observation has an extreme value for the considered feature. Indeed, in practice features with extreme values may be particularly interesting as outlined by Wu (2007). Even for a biomedical expert, it is usually difficult to find out whether the observed extreme values are "aberrant" irrelevant values or rather indicate a potentially interesting subgroup. While a group of, say, five similar extreme values may be worth further investigations, isolated extreme values are less likely to be biologically important and should definitely not have a strong influence on the results of statistical analyses. In this context, the proposed transformation leads to more robust results without completely eliminating the incriminated observation or feature.

Our aim is *not* to generally recommend to use transformed data in place of the original data set. We merely claim that, in the case of biologically uninteresting extreme values (or if one does not know if they are interesting), it is worth considering the results obtained from transformed data parallel to the original results. Statisticians are not expected to say whether the extreme values are interesting or not. Their role is only to identify extreme values and provide robust analyses that are not strongly affected by these extreme values. Our transformation can be useful for both tasks. Note that other transformations may show similar effects. We consider the transformation of Eq. (3.1) because it is very simple and appropriate for all types of statistical analyses, and because its usefulness has already been shown in a different type of multivariate data analysis in the field of medical statistics.

The particularity of our transformation is its simplicity – with several positive consequences. Firstly, it is able for new test data collected later. One simply has to apply the transformation defined by a single equation to the new test observations with the mean and standard deviation estimates from the training set used to do the transformation. Secondly, our transformation is not specific to a particular type of data. While an improved preprocessing variant could perhaps correctly address extreme values for a special data type, our transformation is general enough to be directly applied to all types of (metric) predictors including but not limited to proteomic or metabolomic data. Thirdly, it is also a computationally secure choice : it does not involve, say, complicated fitting procedures with potential convergence problems or instable estimation steps. Lastly, it can be easily implemented in any software tool in a few minutes.

3.5 Research perspective : "Top-top-ranking" variables and very sparse models

Omics data often include several thousands of predictors. Out of these thousands of predictors, a few tens usually focus attention. For example, they are included in the "top-list" shown in the paper or they are involved in the final decision function used to perform prediction. As far as the special case of gene expression microarray data is concerned, most published "signatures" include from 20 to 100 gene expression levels as predictors, for example the well-known 70-gene signature for prediction of breast cancer outcome (van't Veer et al., 2002), which belongs to the few microarray-based signatures already commercialized and used in clinical practice.

Considering the total number of candidate predictors included in the data set, a 70-gene signature is the result of a dramatic variable selection. However, 70 is still a large number in the perspective of clinical applications, because it means that chips would have to be used routinely in clinical practice. Chips are expensive and often difficult to use. That is why practitioners would prefer a very sparse signature involving a handful of predictors that can then be measured using more simple low-throughput techniques based, roughly speaking, on "test tube and pipet" experiments. From a statistical point of view, this implies a further reduction of complexity. This problem is surprisingly ignored by most methodological papers on high-dimensional prediction, although it is suspected that very few predictors are actually good for prediction (Haibe-Kains et al., 2008). The lack of stability of the derived models or"top-lists" of predictors also suggests that few predictors are really important. However, it is still unclear how one should reduce the complexity of a signature from, say, 50 predictors (as commonly selected by standard methods like lasso regression) to 5 or 10. Few procedures have been proposed in this context.

This project is connected to many projects presented in the rest of this synthesis. It is related to the problem of stability, because a possible approach is to focus on stable predictors that are selected consistently over resampling iterations. It is also related to the problem of added predicted value, because we do not want to exclude omics predictors that have much added predicted value. It is related to extreme values, because one might for instance exclude predictors that are in the top-list just because of a single extreme value. Finally, it is also related to over-optimism, because the sequential adaptation of the set of predictors to a particular data set can be seen as a form of data dredging. It makes a fair performance evaluation non-trivial. All steps of model selection should be performed using learning data only, implying much computational effort in resampling settings.

Chapitre 4

Concluding remarks

4.1 New methods versus other projects

During my six post-doctoral years I developed several “new methods” but I also performed studies on existing methods to e.g. compare them, assess their bias, determine how they should (or should not) be used. I found both types of projects exciting. New methods are necessary to answer new arising questions or to better answer well-known questions. I think, however, that studies of the second type are also important in biostatistics research to ensure that previously proposed methods work as expected in different situations and that emerging “standard practice rules” are the result of well-designed studies performed by statisticians.

Good comparison studies contain “new results” in the sense that they yield new knowledge, say, that method A is better than method B. In a broad sense, review articles contain new results in the sense that they establish connections between related approaches and bring structure in a previously less structured knowledge field. Studies showing the pitfalls of a methodology (such as the bias of a widely used importance measure) contain new results in the sense that they highlight previously unknown aspects of the method that potentially affect the way it should be used or interpreted. In my opinion, these three types of studies are an important part of methodological biostatistics research.

One might of course argue that we can also do both within the same project, for instance comparing/reviewing existing methods *and* suggesting a new one. Indeed, in practice new methods are usually compared to a few existing methods in order to establish their superiority. In my opinion, such comparison studies are extremely important for illustrative purposes, but should strictly speaking not be considered as comparison studies. In the context of supervised classification with high-dimensional data, hundreds of these comparison studies can be found in the literature, and they almost all suggest the superiority of different methods – namely the considered new method introduced in the paper. This extreme example illustrates the limitation of comparison studies included in papers introducing new methods. This limitation also affects other fields of statistical research, even if not as dramatically. Over-optimism in the reporting of statistical learning research has many sources. Some of them can be addressed through the application of obvious statistical principles and good practice rules. Others are much more difficult to address. For example, if an implementation problem occurs with the competing approaches and slightly worsens their results, researchers often tend to spontaneously accept these inferior results. Conversely, they would probably obstinately look for the error if such problems occur with their new algorithm. All in all, I believe that additional comparison studies focusing on the comparison itself (and not on a new method) may be very useful to the scientific community.

The question "why not both ?" could also be asked with respect to studies showing the pitfall of a method. Why not also simultaneously propose a new method as a solution ? Of course, finding a solution should be the ultimate goal. However, awareness of the problem is an important step towards the solution. In my opinion, showing an important problem without simultaneously suggesting a solution at the same time should not always be considered as a failure. Let us make a keen comparison with the fields of epidemiology and drug research. In epidemiological research, scientists establish links between diseases and risk factors. Even if they often do not directly suggest concrete interventional procedures or drugs to "make the world better" (i.e. to improve the outcome of the patients), epidemiological studies are considered as an extremely important part of biomedical research. Transferring this idea to the statistical research, it would mean that studies that do not suggest any "intervention" (i.e. a new method) could/should also be considered as valuable contributions to "make the world better" (i.e., roughly speaking, to make results of statistical analyses closer to the truth).

4.2 Dirty details ?

My experience as a statistical consultant for medical doctors also convinced me that small "dirty problems" often affect the conclusions of a study at least as much as statistical modelling issues. Statistical research traditionally focuses on modelling issues and not on dirty problems. By "dirty problems", I mean for instance :

- the reporting strategy to adopt when two statistical methods yield different methods (as discussed in Section 3.3),
- the choice of crucial parameters for which there exists no systematic procedure or a highly variable procedure such as cross-validation,
- the influence of the pre-processing of the data including scaling, dichotomization of quantitative predictors, normalization (in the case of high-throughput omics data), handling of missing values, etc,
- the influence of extreme values,
- the strategy to handle a discrepancy between "what the data say" and "what the biomedical community thinks" such that both components benefit from each other,
- the fact that two different persons may obtain different results for the same analysis, because this analysis is highly sensitive to a particular argument that is often not given any attention,
- the discrepancy between different implementations of the same statistical modelling approach,

and many others. These problems are daily faced by most statisticians working with real data in cooperation with non-statisticians. However, there are often considered as "uninteresting details" by methodological researchers and ignored/underconsidered in many publications. In my opinion, these dirty details deserve more attention in methodological research. Statistical consultants, who can roughly be considered as the end users of the new methods developed by methodological researchers, should not be left alone with these dirty problems.

The needed attention can be given at different levels. Dirty problems may be per se important topics of methodological research. For instance, one may investigate whether pre-processing the whole data set before splitting it into training and test data has an impact on the assessment of prediction models and, if yes, develop adequate methods to handle this problem. Dirty details may also be better acknowledged in methodological projects on new methods by giving more details on the choice of parameters, on the considered method variant, on pre-processing, and by specifically pointing to potential dirty details. In particular, this implies that authors partly give up the “brave new world” writing style and that reviewers are disposed to this change.

Last but not least, an important step in this direction is the publication of reproducible codes implementing the analyses presented in a paper. Two important biometric journals (*Biostatistics* and *Biometrical Journal*) have adopted a reproducibility policy and strongly encourage the publication of codes and/or data together with the article. Reproducible research does not address all dirty problems sketched above. But at least it allows readers to better understand which choices have been done by the authors and how the results were produced. It is a step towards more transparent statistical research. Donoho (2010) states “I often cannot really be sure what a student or colleague has done from his/her own presentation, and in fact often his/her description does not agree with my own understanding of what has been done, once I look carefully at the scripts” and consequently recommend to work in a reproducible way and to publish codes. The benefits of reproducibility and concrete strategies to implement it will also be part of my research topics in the next few years.

Bibliographie

Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., Levine, A., 1999. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proceedings of the National Academy of Sciences 96, 6745–6750.

Altman, D. G., Royston, P., 2000. What do we mean by validating a prognostic model ? Statistics in Medicine 19, 453–73.

Ambroise, C., McLachlan, G. J., 2002. Selection bias in gene extraction in tumour classification on basis of microarray gene expression data. Proceedings of the National Academy of Science 99, 6562–6566.

Baggerly, K. A., Coombes, K. R., 2009. Deriving chemosensitivity from cell lines : Forensic bioinformatics and reproducible research in high-throughput biology. Annals of Applied Statistics 3, 1309–1334.

Bair, E., Tibshirani, R., 2004. Semi-supervised methods to predict patient survival from gene expression data. PLoS Biology 2, 0511.

Bengio, Y., Grandvalet, Y., 2004. No unbiased estimator of the variance of k-fold cross-validation. Journal of Machine Learning Research 5, 1089–1105.

Bernau, C., Augustin, T., Boulesteix, A. L., 2011. Correcting the optimally selected resampling-based error rate : A smooth analytical alternative to nested cross-validation. Technical Report 105, Department of Statistics, LMU.

Bernau, C., Boulesteix, A. L., 2010. Variable selection and parameter tuning in high-dimensional prediction. COMPSTAT Proceedings Online.

Binder, H., Schumacher, M., 2008. Adapting prediction error estimates for biased complexity selection in high-dimensional bootstrap samples. Statistical Applications in Genetics and Molecular Biology 7, 12.

Boulesteix, A. L., 2004. PLS dimension reduction for classification with high-dimensional microarray data. Statistical Applications in Genetics and Molecular Biology 3, Issue 3, Article 33.

Boulesteix, A.-L., 2006a. Maximally selected chi-square statistics and binary splits of nominal variables. Biometrical Journal 48, 838–848.

Boulesteix, A. L., 2006b. Maximally selected chi-square statistics for ordinal variables. Biometrical Journal 48, 451–462.

Boulesteix, A. L., 2007. WilcoxCV : An efficient R package for variable selection in cross-validation. Bioinformatics 23, 1702–1704.

Boulesteix, A. L., 2010. Over-optimism in bioinformatics research. Bioinformatics 26, 437–439.

Boulesteix, A. L., Bender, A., Bermejo, J. L., Strobl, C., 2011a. Random forest gini importance favors snps with large minor allele frequency. Briefings in Bioinformatics, doi : 10.1093/bib/bbr053.

Boulesteix, A. L., Guillemot, V., Sauerbrei, W., 2011b. Use of pre-transformation to cope with extreme values in important candidate features. Biometrical Journal 53, 673–688.

Boulesteix, A. L., Hothorn, T., 2010. Testing the additional predictive value of high-dimensional molecular data. BMC Bioinformatics 11, 78.

Boulesteix, A. L., Porzelius, C., Daumer, M., 2008a. Microarray-based classification and clinical predictors : On combined classifiers and additional predictive value. Bioinformatics 24, 1698–1706.

Boulesteix, A. L., Sauerbrei, W., 2011. Added predictive value of high-throughput molecular data to clinical data, and its validation. Briefings in Bioinformatics 12, 215–229.

Boulesteix, A. L., Strimmer, K., 2007. Partial least squares : A versatile tool for the analysis of high-dimensional genomic data. Briefings in Bioinformatics 8, 32–44.

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

Boulesteix, A. L., Strobl, C., Augustin, T., Daumer, M., 2008b. Evaluating microarray-based classifiers : an overview. Cancer Informatics 6, 77–97.

Boulesteix, A. L., Strobl, C., Weidinger, S., Wichmann, H. E., Wagenpfeil, S., 2007. Multiple testing for SNP-SNP interactions. Statistical Applications in Genetics and Molecular Biology 6, 37.

Bovelstad, H. M., Nygard, S., Borgan, O., 2009. Survival prediction from clinico-genomic models – a comparative study. BMC Bioinformatics 10, 413.

Braga-Neto, U. M., Dougherty, E. R., 2004. Is cross-validation valid for small-sample microarray classification. Bioinformatics 20, 374–380.

Breiman, L., 2001. Random forests. Machine Learning 45 (1), 5–32.

Breiman, L., Friedman, J. H., Olshen, R. A., Stone, J. C., 1984. Classification and Regression Trees. Wadsworth, Monterey, CA.

Bühlmann, P., Hothorn, T., 2007. Boosting algorithms : regularization, prediction and model fitting (with discussion). Statistical Science 22, 477–505.

Calle, M., Urrea, V., 2011. Letter to the editor : Stability of random forest importance measures. Briefings in Bioinformatics 12, 86–89.

Daumer, M., Held, U., Ickstadt, K., Heinz, M., Schach, S., Ebers, G., 2008. Reducing the probability of false positive research findings by pre-publication validation : Experience with a large multiple sclerosis database. BMC Medical Research Methodology 8, 18.

Ding, B., Gentleman, R., 2005. Classification using generalized partial least squares. Journal of Computational and Graphical Statistics 14, 280–298.

Donoho, D., 2010. An invitation to reproducible computational research. Biostatistics 11, 385–388.

Edgeworth, F. Y., 1887. On observations relating to several quantities. Hermathena 6, 279–285.

Fort, G., Lambert-Lacroix, S., 2005. Classification using partial least squares with penalized logistic regression. Bioinformatics 21, 1104–1111.

Gerds, T., Schumacher, M., 2006. Consistent estimation of the expected Brier score in general survival models with right-censored event times. Biometrical Journal 48, 698–705.

Goeman, J. J., 2010. penalized : L1 (lasso) and L2 (ridge) penalized estimation in GLMs and in the Cox model. R Package version 0.9-31.

Goeman, J. J., Oosting, J., Cleton-Jansen, A. M., Anninga, J. K., van Houwelingen, H. C., 2005. Testing association of a pathway with survival using gene expression data. Bioinformatics 21, 1950–1957.

Goeman, J. J., van de Geer, S. A., de Kort, F., van Houwelingen, H. C., 2004. A global test for groups of genes : testing association with a clinical outcome. Bioinformatics 20, 93–99.

Grabmeier, J. L., Lambe, L. A., 2007. Decision trees for binary classification variables grow equally with the gini impurity measure and pearson's chi-square test. International Journal of Business Intelligence and Data Mining 2, 213–226.

Graf, E., Schmoor, C., Sauerbrei, W., Schumacher, M., 1999. Assessment and comparison of prognostic classification schemes for survival data. Statistics in Medicine 18, 2529–2545.

Guo, Y., Hastie, T., Tibshirani, R., 2007. Regularized discriminant analysis and its application in microarrays. Biostatistics 8, 86–100.

Haibe-Kains, B., Desmedt, C., Sotiriou, C., Bontempi, G., 2008. A comparative study of survival models for breast cancer prognostication based on microarray data : does a single gene beat them all ? Bioinformatics 24, 2200–2208.

Hanczar, B., Hua, J., Dougherty, E. R., 2007. Decorrelation of the true and estimated classifier errors in high-dimensional settings. EURASIP J. Bioinformatics and Systems Biology 207, 38473.

Hanczar, B., Hua, J., Sima, C., Weinstein, J., Bittner, M. L., Dougherty, E. R., 2010. Small-sample precision of roc-related estimates. Bioinformatics 26, 822–830.

Herold, T., Jurinovic, V., Metzeler, K. H., Boulesteix, A. L., Bergmann, M., Seiler, T., Mulaw, M., Thoene, S., Dufour, A., Pasalic, Z., Schmidberger, M., Schmidt, M., Schneider, S., Kakadia, P. M., Feuring-Buske, M., Braess, J., Spiekermann, K., Mansmann, U., Hiddemann, W., Buske, C., Bohlander, S. K., 2011. An eight-gene expression signature for the prediction of survival and time to treatment in chronic lymphocytic leukemia. Leukemia 25, 1639–1645.

Hothorn, T., Hornik, K., Zeileis, A., 2006. Unbiased recursive partitioning : A conditional inference framework. Journal of Computational and Graphical Statistics 15, 651–674.

Jelizarow, M., Guillemot, V., Tenenhaus, A., Strimmer, K., Boulesteix, A. L., 2010. Over-optimism in bioinformatics : an illustration. Bioinformatics 26, 1990–1998.

Kostka, D., Spang, R., 2008. Microarray based diagnosis profits from better documentation of gene expression signatures. PLoS Computational Biology e22, 4.

Koziol, J. A., 1991. On maximally selected chi-square statistics. Biometrics 47, 1557–1561.

Mehta, T., Murat, T., Allison, D. B., 2004. Towards sound epistemological foundations of statistical methods for high-dimensional biology. Nature Genetics 36, 943–947.

Metzeler, K. H., Hummel, M., Bloomfield, C. D., Spiekermann, K., Braess, J., Sauerland, M. C., Heinecke, A., Radmacher, M., Marcucci, G., Whitman, S. P., Maharry, K., Paschka, P., Larson, R. A., Berdel, W. E., Buchner, T., Wormann, B., Mansmann, U., Hiddemann, W., 2008. An 86 probe set gene expression signature predicts survival in cytogenetically normal acute myeloid leukemia. Blood 112, 4193–4201.

Nicodemus, K., 2011. Letter to the editor : On the stability and ranking of predictors from random forest variable importance measures. Briefings in Bioinformatics 12, DOI :10.1093/bib/bbr016.

Pepe, M. S., Janes, H., et al., G. L., 2004. Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. American Journal of Epidemiology 159, 882–890.

Rocke, D. M., Ideker, T., Troyanskaya, O., Quackenbush, J., Dopazo, J., 2009. Papers on normalization, variable selection, classification or clustering of microarray data. Bioinformatics 25, 701–702.

Rousseeuw, P. J., Leroy, A. M., 2003. Robust regression and outlier detection. Wiley, New York, NY.

Royston, P., Sauerbrei, W., 2007. Improving the robustness of fractional polynomial models by preliminary covariate transformation : A pragmatic approach. Computational Statistics & Data Analysis 51, 4240–4253.

Schäfer, J., Strimmer, K., 2005. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. Statistical Applications in Genetics and Molecular Biology 4, Issue 1, Article 32.

Singh, D., Febbo, P. G., Ross, K., Jackson, D. G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A. A., D'Amico, A. V., Richie, J. P., Lander, E. S., Loda, M., Kantoff, P. W., Golub, T. R., Sellers, W. R., 2002. Gene expression correlates of clinical prostate cancer behavior. Cancer Cell 1, 203–209.

Slawski, M., Daumer, M., Boulesteix, A. L., 2008. CMA – A comprehensive bioconductor package for supervised classification with high dimensional data. BMC Bioinformatics 9, 439.

Strobl, C., Boulesteix, A. L., Augustin, T., 2007a. Unbiased split selection for classification trees based on the Gini Index. Computational Statistics and Data Analysis 52, 483–501.

Strobl, C., Boulesteix, A. L., Kneib, T., Augustin, T., Zeileis, A., 2008. Conditional variable importance for random forests. BMC Bioinformatics 9, 307.

Strobl, C., Boulesteix, A. L., Zeileis, A., Hothorn, T., 2007b. Bias in random forest variable importance measures : Illustrations, sources and a solution. BMC Bioinformatics 8, 25.

Tibshirani, B., Tibshirani, R., 2009. A bias correction for the minimum error rate in cross-validation. Annals of Applied Statistics 3, 822–829.

Tibshirani, R., 1996. Regression shrinkage and selection via the LASSO. Journal of the Royal Statistical Society B 58, 267–288.

Tibshirani, R., Efron, B., 2002. Pre-validation and inference in microarrays. Statistical Applications in Genetics and Molecular Biology 1, 1.

Tibshirani, R., Hastie, T., 2007. Outlier sums for differential gene expression analysis. Biostatistics 8, 2–8.

Tibshirani, R., Hastie, T., Narasimhan, B., Chu, G., 2002. Class prediction by nearest shrunken centroids, with applications to dna microarrays. Statistical Science 18, 104–117.

van Houwelingen, H., 2000. Validation, calibration, revision and combination of prognostic survival models. Statistics in Medicine 19, 3401–3415.

van't Veer, L. J., Dai, H., van de Vijver, M. J., He, Y. D., Hart, A. A. M., Mao, M., Peterse, H. L., van der Kooy, K., Marton, M. J., Witteveen, A. T., Schreiber, G. J., Kerkhoven, R. M., Roberts, C., Linsley, P. S., Bernards, R., Friend, S. H., 2002. Gene expression profiling predicts clinical outcome of breast cancer. Nature 415, 530–536.

Vapnik, V. N., 1995. The nature of statistical learning theory. Springer, New York.

Varma, S., Simon, R., 2006. Bias in error estimation when using cross-validation for model selection. BMC Bioinformatics 7, 91.

Wu, B., 2007. Cancer outlier differential gene expression detection. Biostatistics 8, 566–575.

Yao, M., Huang, Y., Shioi, K., Hattori, K., Murakami, T., Sano, F., Baba, M., Kondo, K., Nakaigawa, N., Kishida, T., Nagashima, Y., Yamada-Okabe, H., Kubota, Y., 2008. A three-gene expression signature model to predict clinical outcome of clear cell renal carcinoma. International Journal of Cancer 123, 1126–1132.

Yousefi, M. R., Hua, J., Sima, C., Dougherty, E. R., 2010. Reporting bias when using real data sets to analyze classification performance. Bioinformatics 26, 68–76.

Zhu, J., Hastie, T., 2004. Classification of gene microarrays by penalized logistic regression. Biostatistics 5, 427–443.

Zuber, V., Strimmer, K., 2009. Gene ranking and biomarker discovery under correlation. Bioinformatics 25, 1700–1707.