Over-optimism in bioinformatics: an illustration

Monika Jelizarow, Vincent Guillemot, Arthur Tenenhaus, K. Strimmer, Anne-Laure Boulesteix

Over-optimism in bioinformatics: an illustration

M. Jelizarow[1] V. Guillemot[1,2] A. Tenenhaus[2] K. Strimmer[3] A.-L. Boulesteix[1*]

[1] Department of Medical Informatics, Biometry and Epidemiology, University of Munich, Marchioninistr. 15, 81377 Munich, Germany

[2] Département SSE, Ecole Supélec, 3 rue Joliot-Curie, 91192 Gif-sur-Yvette, France

[3] Department of Medical Informatics, Statistics and Epidemiology, University of Leipzig, Härtelstr. 16-18, 04107 Leipzig, Germany

Abstract

In statistical bioinformatics research, different optimization mechanisms potentially lead to "over-optimism" in published papers. The present empirical study illustrates these mechanisms through a concrete example from an active research field. The investigated sources of over-optimism include the optimization of the data sets, of the settings, of the competing methods and, most importantly, of the method's characteristics.

We consider a "promising" new classification algorithm that turns out to yield disappointing results in terms of error rate, namely linear discriminant analysis incorporating prior knowledge on gene functional groups through an appropriate shrinkage of the within-group covariance matrix. We quantitatively demonstrate that this disappointing method can artificially seem superior to existing approaches if we "fish for significance". We conclude that, if the improvement of a quantitative criterion such as the error rate is the main contribution of a paper, the superiority of new algorithms should be validated using "fresh" validation data sets.

The R codes and preprocessed versions of the data sets as well as additional files can be downloaded from http://www.ibe.med.uni-muenchen.de/organisation/mitarbeiter/020_professuren/boulesteix/overoptimism/, such that the study is completely reproducible.

Keywords: Validation, fishing for significance, meta-methodology, KEGG, discriminant analysis, shrinkage covariance estimator

*Corresponding author. Email: boulesteix@ibe.med.uni-muenchen.de.

1 Introduction

In statistical bioinformatics research, the reported results on the performance of new algorithms are known to be over-optimistic, as recently discussed in a letter to the editors of Bioinformatics (Boulesteix, 2010). The current paper aims at illustrating the different mechanisms leading to over-optimism through a concrete example from an active methodological research field.

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. 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" in this paper.

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 after they notice that this setting favors the new algorithm compared to existing benchmark approaches. This mechanism is termed as "optimization of the settings" in this paper.

The third source of over-optimism is related to the choice of the existing benchmark methods applied for comparison purposes. Researchers 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. This mechanism is termed as "optimization of the competing methods" in this paper.

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 er-

ror 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" in this paper.

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 methodological bioinformatics research and this issue has long been underconsidered in the literature.

The present paper aims at filling this gap. It reviews and illustrates the problem of validation and false research findings through a concrete example within a hot research field: the incorporation of prior biological knowledge on gene functional groups into high-dimensional microarray-based classification. The "promising idea" we originally pursued was to modify the well-established shrinkage covariance estimator by Schäfer and Strimmer (2005) by incorporating prior knowledge on gene functional groups with the aim to improve the performance of linear discriminant analysis. This new approach can be seen as a combination of a simple and well-established statistical method, namely the shrinkage estimator of the covariance, with a popular concept (the incorporation of prior biological knowledge into classification) that has attracted a lot of attention in the last few years (Tai and Pan, 2007a,b; Rapaport et al., 2007; Li and Li, 2008; Guillemot et al., 2008; Binder and Schumacher, 2009; Jacob et al., 2009; Yousef et al., 2009; Slawski et al., 2010; Hall and Xue, 2010). For these reasons, we considered this new approach as promising. However, it turned out that this interesting method does not yield any improvement in terms of prediction error rate.

Based on this concrete example, we show that over-optimistic results can be obtained through the four mechanisms discussed above. 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. Note that 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 feel that a quantitative demonstration of the optimistic bias affecting methodological research may perhaps increase awareness on such problems and give researchers food

for thoughts.

The remainder of this paper is organized as follows. The promising idea is briefly sketched in Section 2.1 to make our considerations on validation more understandable. The design of the analysis is described in Section 2.2, while Section 3 presents the results of the new and existing methods on four real-life data sets and the different interpretations depending on whether one fishes for significance or not. Further potential sources of biases and possible explanations for the disappointing results of the promising idea are discussed in Section 4.

2 Methods

2.1 A "promising idea"

This section briefly sketches the promising idea we originally pursued to make our considerations and results on over-optimism more understandable. Note, however, that this promising idea is *not* the scientific contribution of our paper, but rather a concrete example serving as an illustration for the four investigated optimization mechanisms. Readers who are not interested in the methodological details of the promising idea but rather in the quantitative evaluation of the optimization mechanisms can skip this section.

2.1.1 Discriminant analysis and its regularized variants

Let us consider a high-dimensional data set with continuous predictors such as microarray gene expression data. The latter are often used to predict a categorical response variable of interest, e.g. the disease status or the long-term disease outcome.

Discriminant analysis (DA) is a widely used classification method. DA is based on the assumption that the random vector $\boldsymbol{x}$ of predictors follows a multivariate normal distribution $\boldsymbol{x}|(Y = r) \sim \mathcal{N}(\mu_r, \boldsymbol{\Sigma}_r)$ within each class r (for $r = 1, \ldots, c$). A new observation $\boldsymbol{x}_{\text{new}}$ is then assigned to the class with maximal posterior probability. This decision rule can be formulated in terms of a simple decision function which is linear in $\boldsymbol{x}_{\text{new}}$ if the covariance matrices $\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_c$ are assumed to be equal, yielding the so-called Linear Discriminant Analysis (LDA). Most importantly, the decision function involves the inverse $\boldsymbol{\Sigma}^{-1}$ of the covariance matrix $\boldsymbol{\Sigma}$. In standard $n > p$ settings, $\boldsymbol{\Sigma}^{-1}$ is simply estimated through the inverse $\tilde{\mathbf{S}}^{-1}$ of the pooled estimator $\tilde{\mathbf{S}}$ of the within-covariance matrix, which is itself defined as a weighted sum of the unbiased estimators of the within-class covariance matrices. More technical details on classical LDA are given in the Additional File 1 available from the companion website.

In the high-dimensional setting considered here the pooled covariance estimator $\tilde{\mathbf{S}}$ mentioned above is singular, thus not invertible. The concept of Regularized Linear Discriminant Analysis (RLDA) aims at solving the singularity problem by modifying $\tilde{\mathbf{S}}$ such that the resulting estimator becomes invertible. See for instance the seminal paper on Regularized (Fisher's) Discriminant Anal-

ysis by Friedman (1989) and the work by Guo et al. (2007) on Shrunken Centroids Regularized Discriminant Analysis (SCRDA) which are both based on the widely employed shrinkage principle (Stein, 1955; Efron and Morris, 1977).

2.1.2 Regularized LDA with KEGG

An increasingly popular approach is to regularize the within-class covariance by incorporating external biological knowledge from databases like the **K**yoto **E**ncyclopedia of **G**enes and **G**enomes (KEGG) (Kanehisa and Goto, 2000). The underlying motivation of this approach is to improve both the prediction accuracy and the results' interpretability.

KEGG is a freely available database of biological systems consisting of multiple sub-databases. KEGG PATHWAY as one of these sub-databases contains a collection of pathway maps representing recent knowledge on molecular interaction and reaction networks for metabolism, various cellular processes and human diseases (Kanehisa and Goto, 2000). More precisely, pathways are represented as graphs in which the edges stand for the chemical reactions or relations and the vertices stand for the genes involved.

In the context of microarray-based classification, Tai and Pan (2007a) assume that a KEGG pathway forms a gene functional group. They postulate that genes from the same functional group tend to be more correlated than genes from different functional groups, and that information from KEGG can thus be used to improve the modelling of between-genes correlation in the context of classification. Starting from these attractive ideas, we propose an alternative simple approach to incorporate prior knowledge from KEGG into the estimation of the correlation, with applications to LDA. The promising idea can be seen as a further variant of RLDA incorporating biological knowledge on gene functional groups extracted from KEGG via a modified shrinkage estimator of the covariance matrix, as outlined in Sections 2.1.3 and 2.1.4.

2.1.3 The shrinkage estimator $\widehat{\boldsymbol{\Sigma}}_{\text{SHIP}}$ incorporating prior knowledge

To address the methodological challenges arising from the $n \ll p$ data situation (the pooled estimate $\tilde{\mathbf{S}}$ of the covariance matrix is not invertible), we now propose a covariance estimation procedure we refer to as **SHIP** standing for **SH**rinking and **I**ncorporating **P**rior knowledge. The resulting covariance estimator $\widehat{\boldsymbol{\Sigma}}_{\text{SHIP}}$ is based on the shrinkage estimator introduced by Ledoit and Wolf (2003, 2004) and applied by Schäfer and Strimmer (2005) in the context of high-dimensional genomic data. Additionally, the new estimator incorporates prior biological knowledge on gene functional groups extracted from the KEGG database.

In a few words, the shrinkage estimator originally proposed by Ledoit and Wolf is the asymptotically optimal convex linear combination $\widehat{\boldsymbol{\Sigma}}^* = \lambda\mathbf{T} + (1-\lambda)\mathbf{S}$, where $\lambda \in [0,1]$ denotes the analytically determined optimal shrinkage intensity, $\mathbf{T}$ stands for a structured covariance target, and $\mathbf{S}$ is the unstructured standard unbiased empirical covariance matrix. The resulting "shrinkage estima-

tor" of the covariance matrix $\mathbf{\Sigma}$ is then invertible (provided $\mathbf{T}$ is chosen adequately) and stabilized. The optimal shrinkage intensity λ is determined with respect to a quadratic loss function which is common and intuitive in statistical decision theory, resulting in a simple analytical formula (Schäfer and Strimmer, 2005). See Additional File 1 for more details on the computation of λ.

The covariance target $\mathbf{T}$ plays an essential role in the computation of the shrinkage estimator by Ledoit and Wolf. Its choice, however, turns out to be very complex. On the one hand, $\mathbf{T}$ is required to be positive definite and to involve only a small number of free parameters. On the other hand, it should reflect important characteristics of the covariance structure between the variables (genes). An overview of commonly used covariance targets A to F is given in Schäfer and Strimmer (2005). In this paper, we consider target D and target F with constant correlation as reference methods (see Table 1, left and middle).

In order to incorporate information from KEGG PATHWAY, we propose a modified version of target F where pairs of connected genes (i.e. genes from the same gene functional group) have non-zero common correlation $\bar{r}$. This correlation is simply given as the mean correlation of all pairs of connected genes. In case a gene does not occur in any gene functional group, we assume this gene forming its own group with group size one as in Tai and Pan (2007a). The resulting target G is displayed in Table 1 and yields the novel estimator $\widehat{\mathbf{\Sigma}}_{\text{SHIP}} = \lambda\mathbf{T} + (1-\lambda)\mathbf{S}$, where $\mathbf{T}$ is defined according to target G and the optimal shrinkage intensity λ can be computed analytically (see Additional File 1). The shrinkage covariance estimator $\widehat{\mathbf{\Sigma}}_{\text{SHIP}}$ is implemented in the `R` package '`SHIP`' which is publicly available from the companion website.

Target D

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

Target F

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

Target G

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

Table 1: Overview of targets D (diagonal, unequal variance), F (constant correlation) and G (where $\bar{r}$ is the average of sample correlations). The notation $i \sim j$ means that genes i and j are connected, i.e. genes i and j are in the same gene functional group. The term s_{ij} denotes the entry of the unbiased covariance matrix in row i, column j.

2.1.4 Linear discriminant analysis using $\widehat{\mathbf{\Sigma}}_{\text{SHIP}}$

The resulting estimator $\widehat{\mathbf{\Sigma}}_{\text{SHIP}}$ of the covariance matrix can then simply be used in the context of LDA. In a nutshell, we compute the shrinkage estimators $\widehat{\mathbf{\Sigma}}^{(r)}_{\text{SHIP}}$ separately for each class $r = 1, ..., c$ and subsequently pool these within-class shrinkage estimators according to the standard procedure known from LDA. See Additional File 1 for more details. Note that the resulting pooled estimator is not necessarily positive definite, because the target is not always positive definite. However, it is typically much better conditioned than $\tilde{\mathbf{S}}$. To cope with this problem, we simply compute the well-known Moore-Penrose pseudoinverse (Penrose, 1955).

Our initial conjecture was that this variant of RLDA borrowing prior knowledge from KEGG may lead to an improvement of prediction accuracy. This conjecture is intentionally formulated rather imprecisely. Of course, one may ask for a more concrete explanation of the term "improvement". This is indeed an important question we address in Section 3.

2.2 Design of the study

Since our quantitative study on the four optimization mechanisms is actually the real contribution of our paper, the design of the study is presented as a part of the Methods section, following the four optimization mechanisms outlined in the introduction.

2.2.1 Data sets

In this study, we successively consider four publicly available microarray data sets to illustrate the potential optimization of the data set and demonstrate the importance of validation on different data sets. Golub's leukemia data set ($n = 72$, $p = 7129$) is part of the `R` package '`golubEsets`' (Golub, 2010), while the CLL data set ($n = 22$, $p = 12625$) is available from the package '`CLL`' (Whalen, 2010). The prostate data set by Singh et al. (2002) ($n = 102$, $p = 12625$) and the breast cancer data set by Wang et al. (2005) ($n = 286$, $p = 22283$) are available from GEO. We normalized them using the GCRMA method. The resulting data matrices are available from the companion website. All data sets include a binary outcome variable which has to be predicted based on gene expression data. A brief overview of the data sets is given in Additional File 1.

2.2.2 Settings

Prediction accuracy is estimated using the well-established $10 \times$ five-fold cross-validation evaluation scheme. Five-fold cross-validation is repeated 10 times in order to achieve more stable results (Braga-Neto and Dougherty, 2004; Boulesteix et al., 2008). We focus on the average misclassification rate as a measure of prediction accuracy, i.e. the average test error obtained over all $10 \times 5 = 50$ test sets.

In order to limit the computational effort and to reduce the influence of noise we do not employ all available genes of a data set, but perform variable selection (for each learning set successively, as commonly recommended). We use three variable selection criteria: the standard t-test, the Limma procedure by Smyth (2004) and the standard rank-based Wilcoxon test, each with four different numbers of selected genes ($p^* = 100, 200, 500, 1000$). Hence, we obtain $3 \times 4 = 12$ combinations of selection procedures and numbers of selected genes.

2.2.3 Competing methods

For comparison purposes, we furthermore apply the Diagonal Linear Discriminant Analysis (DLDA), the Nearest Shrunken Centroids method (NSC) by Tibshirani et al. (2002) that is also called Prediction Analysis with Microarrays (PAM), and Support Vector Machines (SVM) as competing approaches.

We perform variable selection for DLDA with $p^* = 100, 200, 500, 1000$ and three selection methods successively, Following common practice, we skip the variable selection for NSC and SVM where the influence of irrelevant genes is reduced automatically. Tuning parameters for NSC (shrinkage parameter) and SVM (cost) are optimized via internal three-fold cross-validation.

2.2.4 Method's characteristics

When developing a new algorithm, researchers often adapt their method sequentially depending on their experiences with example data sets and preliminary results. Many variants that are tried out at this stage finally turn out to yield bad results or fail for any other reason. In contrast to the aspects of the analysis design discussed above, this aspect often remains unmentioned when writing a paper, except perhaps a few remarks in the discussion. However, the variants that are tried out during the development of the new algorithms are in a broad sense part of the design of the analysis. Indeed, they are often assessed using the same procedures as the final new algorithm that is eventually published.

When assessing the promising idea described in Section 2.1, we also thought of possible variants of the proposed RLDA incorporating prior knowledge. In contrast to standard practice, we publicly mention all these variants in the present paper and demonstrate what happens when one systematically tries to optimize the new algorithm with regard to its characteristics.

Henceforth, the promising idea outlined in Section 2.1 is referred to as rlda.TG unless otherwise emphasized. More precisely, the term rlda.TG specifies the regularized linear discriminant analysis with the shrinkage estimators of the within-class covariance matrices being based on the knowledge-based covariance target G as introduced in Section 2.1.3. During the development phase, we successively considered the ten following variants of rlda.TG termed as rlda.TG$^{(1)}$, ..., rlda.TG$^{(10)}$. These ten variants can be divided into two groups. The first group comprises rlda.TG$^{(1)}$ to rlda.TG$^{(7)}$ which differ in the assignment of ambiguous genes (genes that are in no functional group or genes that are in at least two different functional groups). While rlda.TG$^{(1)}$ excludes genes that are not in any gene functional group (about 50 $\%$ in each data set) from the analysis, rlda.TG$^{(2)}$ eliminates genes occuring in multiple gene functional groups. Both rlda.TG$^{(3)}$ and rlda.TG$^{(4)}$ proceed similarly to Tai and Pan (2007a): if a gene occurs in multiple gene functional groups, the gene is kept in the gene functional group with the smallest (largest) number of genes and ignored in the other ones. In case the smallest (largest) gene functional group is not unique, one of these is chosen by chance. The methods rlda.TG$^{(5)}$ to rlda.TG$^{(7)}$ are obtained by combining rlda.TG$^{(1)}$ with rlda.TG$^{(2)}$, rlda.TG$^{(3)}$ and rlda.TG$^{(4)}$. The second group comprises rlda.TG$^{(8)}$, rlda.TG$^{(9)}$ and rlda.TG$^{(10)}$ which are based on a re-definition of the covariance target G. Variant rlda.TG$^{(8)}$ involves two parameters for the correlation (a positive and a negative one) instead of the single parameter $\bar{r}$, in order to account for negatively correlated genes within the same pathway. The variant rlda.TG$^{(9)}$ completely ignores negative correlations and computes the mean correlation using the positive ones. Finally, rlda.TG$^{(10)}$ tests the correlations (with a level of 0.05) and sets the non-significant correlations to zero before the mean correlation is computed.

3 Results

3.1 General approach

This section presents different interpretations of the results of the new methods rlda.TG, rlda.TG$^{(1)}$, . . . ,rlda.TG$^{(10)}$ and existing methods on four real-life data sets. While Section 3.2 presents the performance of the new algorithm(s) from an over-optimistic point of view (i.e. after fishing for significance), Section 3.3 follows a less biased approach based on validation with independent data sets.

The four optimization mechanisms are introduced sequentially and independently of each other for clarity's sake in Section 1. However, they are in fact tightly embedded in practice, thus making a perfectly realistic study very difficult. In Section 3.2, we consider a simplified optimization process mimicking one of many possible optimization scenarii for illustration purposes. We are aware of the many other potential schemes, but an exhaustive study would go beyond this paper's scope. We feel that the chosen example reflects the influence of the four mechanisms reasonably well. In addition to the results provided in this section, a more extensive report of the results is given in Additional File 2 available from the companion website.

In this study, all four data sets are first analysed independently of each other in Section 3.2 to mimick what would happen if researchers did not try to validate their results on different data sets. It is then shown in Section 3.3 that a proper validation strategy, in which researchers do not use the same data sets to develop and to evaluate their new algorithm, leads to much less favorable results. The whole analysis is completely reproducible using the `R` codes available from the companion website.

3.2 An (over-)optimistic view

3.2.1 Optimization of the settings

We first consider the new promising method rlda.TG while ignoring its variants rlda.TG$^{(1)}$,. . . ,rlda.TG$^{(10)}$. The four data sets are analysed completely independently of each other. For a given data set, someone "fishing for significance" may look for the variable selection scheme and number p^* of selected variables yielding the lowest error rate. In this spirit, Table 2 gives the classification error rates obtained with the 3×4 combinations of variable selection scheme and number p^* of selected variables in each of the four investigated data sets. The bold numbers indicate the minimal error rate for each data set. The standard errors of the error rates over the cross-validation iterations range from 0.005 to 0.024 for the Golub data, from 0.022 to 0.031 for the CLL data, from 0.009 to 0.012 for the Wang data, and from 0.008 to 0.021 for the Singh data. Obviously, the classification error rates strongly depend on the variable selection settings. Moreover, there is no universally better setting performing best for all data sets, although settings with small p^* tend to yield smaller error rates in general.

Selection procedure	p^*	Golub	CLL	Wang	Singh
t-test	100	0.029	0.234	0.382	**0.081**
	200	0.029	0.269	**0.375**	0.133
	500	0.032	0.220	0.383	0.166
	1000	0.049	0.222	0.380	0.211
Limma	100	0.031	0.237	0.383	**0.081**
	200	**0.028**	0.274	**0.375**	0.125
	500	0.039	0.233	0.384	0.182
	1000	0.060	0.225	0.376	0.224
Wilcoxon test	100	0.090	0.192	0.384	0.135
	200	0.170	**0.159**	0.379	0.178
	500	0.168	0.185	0.409	0.158
	1000	0.124	0.221	0.402	0.197

Table 2: Overview of the CV errors obtained for rlda.TG where p^* denotes the number of selected genes. The bold values indicate the minimum values.

A researcher who "fishes for significance" would select the setting yielding the minimal error rate for the data set (s)he considers, thus inducing an optimistic bias through "optimization of the settings".

3.2.2 Optimization of the method's characteristics

Moreover, (s)he would certainly try to further improve the new algorithm's performance by considering the additional variants rlda.TG$^{(1)}$, ..., rlda.TG$^{(10)}$. Figure 1 displays the CV error rates of rlda.TG and its variants in the selected setting(s) for each data set. Especially for the CLL and the Wang data set, it can be clearly seen that some of the variants decrease the error rate substantially compared to rlda.TG. All in all, we achieve the error rates 0.025 for the Golub data (with rlda.TG$^{(5)}$), 0.129 for the CLL data (with rlda.TG$^{(5)}$), 0.342 for the Wang data (with rlda.TG$^{(6)}$), and 0.078 for the Singh data (with rlda.TG$^{(8)}$). This represents an improvement compared to the bold optimal error rates from Table 2, illustrating the mechanism denoted as "optimization of the method's characteristics".

3.2.3 Optimization of the competing approaches

Another mechanism of the optimization process is the choice of the competing approaches that are compared to the new algorithm. For each of the four data sets, Table 3 shows the difference between the error rate of the optimal method in the optimal setting and the error rate of rlda.TD (shrinkage covariance with the diagonal target D), rlda.TF (shrinkage covariance with target F), and DLDA (classical diagonal linear discriminant analysis). These competing approaches are applied after variable selection following the optimal setting identified from Table 2. Further, results are shown for two good standard methods without preliminary variable selection: the Nearest Shrunken Centroids method (NSC) and the Support Vector Machines (SVM). Obviously, these competing approaches

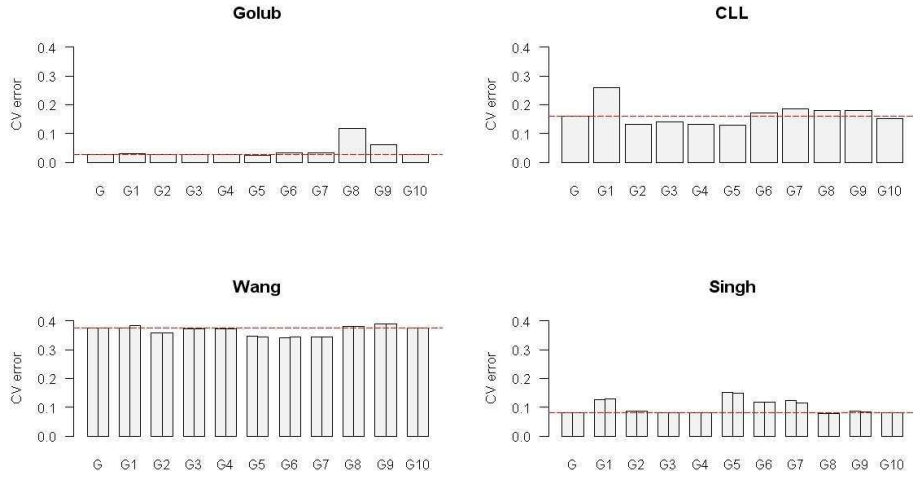

Figure 1: Overview of the CV error rates of the different variants of rlda.TG, obtained for all data sets within the corresponding optimal settings s_{opt}. The dashed line indicates the value obtained for rlda.TG within the data-specific s_{opt}. Note that for both the Wang and the Singh data the optimal setting is not unique. The considered settings are: s_{opt} = (200, Limma) for the Golub data, s_{opt} = (200, Wilcoxon test) for the CLL data, ${s_{opt}}^1$ = (200, t-test) (left bar) and ${s_{opt}}^2$ = (200, Limma) (right bar) for the Wang data, and ${s_{opt}}^1$ = (100, t-test) (left bar) and ${s_{opt}}^2$ = (100, Limma) (right bar) for the Singh data.

perform very differently. Hence, the new algorithm's performance appears more or less impressive depending on the competing methods shown in the comparison study.

A possible (critical) strategy could be to select the competing approaches depending on the tested "research hypothesis". If the hypothesis is that the new algorithm generally improves the performance of state-of-the-art approaches, we would consider as many approaches as possible. If the hypothesis is that it performs better than other LDA approaches, we would consider all LDA-based competitors. If the hypothesis is that the incorporation of correlations is useful, we would consider rlda.TD. If the hypothesis is that the incorporation of correlations becomes better through KEGG-pathways, we would consider rlda.TF. This strategy may seem good at first view, but yields some problems. First, the tested hypothesis should not be chosen a posteriori by the researcher based on the results. Indeed, it can be seen from Table 3 that this also yields a kind of optimization. Second, it may also lead to spurious results. For example, one may conclude from the negative differences $D(M_{opt}$,rlda.TF) that KEGG is useful in this context. Another more realistic explanation is that rlda.TG is better than rlda.TF because the estimated correlation matrix is sparser – and not because of the KEGG pathways.

3.2.4 Optimization of the data set

Some researchers may also "optimize the data set" and choose to show only the results that are more favorable to their method. For an extensive study on this problem including theoretical considerations, see Yousefi et al. (2010). It can be clearly seen from Table 3 that the results on the CLL data are much

	M_{opt}	$D(M_{opt}$,rlda.TD)	$D(M_{opt}$,rlda.TF)	$D(M_{opt}$,DLDA)	$D(M_{opt}$,NSC)	$D(M_{opt}$,SVM)
Golub	rlda.TG$^{(5)}$	- 0.003	- 0.003	- 0.010	0.004	- 0.029
CLL	rlda.TG$^{(5)}$	- 0.017	- 0.083	- 0.055	- 0.204	- 0.269
Wang	rlda.TG$^{(6)}$	- 0.026	- 0.026	- 0.033	- 0.034	0.001
Singh	rlda.TG$^{(8)}$	- 0.008	- 0.003	- 0.048	- 0.052	- 0.022

Table 3: Overview of the differences D between the error rates of the data-specific optimal variant M_{opt} of rlda.TG and the methods rlda.TD, rlda.TF, DLDA, NSC and SVM within the data-specific optimal setting s_{opt}.

more favorable to our new method than the other three data sets. This is probably due to the very small size ($n = 22$) implying a high variability and thus stronger optimization effects. The optimization of the data set and the optimization of the settings may thus be tightly connected.

3.3 On the usefulness of validation with fresh data

Until now, the four data sets were analysed independently of each other. For each data set, we obtained an optimal variant combined with an optimal setting, that seemingly performed better than existing approaches, see Table 3. As previously discussed, these figures are the result of different optimization processes. One of them – the optimization of the method's characteristics – is an inherent component of biostatistics/bioinformatics research and cannot be avoided. Up to a point, the optimization of the settings can also be considered as inherent to data analysis research: for example, nobody expects researchers to focus on settings in which all methods turn out to perform equally bad. So how should we evaluate new methods and report their performance?

In this section, we show the importance of a proper validation using data sets that were not used for the algorithm's development. Table 4 shows the cross-validation error rates of the four combinations of optimal settings and optimal variant when applied on the four data sets. Whereas the error rates in the diagonal are the optimal error rates already mentioned in the previous section, the error rates outside the diagonal can be seen as "validation error rates" computed on independent fresh data sets. They are obviously much higher than the optimal error rates, illustrating the consequences of the optimization processes.

In the same vein, Figure 2 displays the number of variable selection settings (out of $3 \times 4 = 12$) in which each of the variants rlda.TG,rlda.TG$^{(1)}$, . . . ,rlda.TG$^{(10)}$ yields the lowest error rate, for each data set separately. It can be seen that the "optimal variant" strongly depends on the data set (because the four rows are very different) and on the setting (because we have many intermediate values like 2,3,4,5 $<$ 12). There is no clear winner, but readers may have the impression that there is a clear winner if they do not see all the results (i.e. not all data sets or/and not all settings).

In conclusion, validation using fresh independent data that were not used in the development phase would have avoided over-optimistic conclusions on the new algorithm's superiority. This kind of validation automatically corrects the bias induced by the optimization of the settings and the optimization of the method's characteristics.

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

Table 4: Performance of the optimal variants M_{opt} of rlda.TG within the optimal settings s_{opt} selected in each of the four data sets. The figures outside the diagonal can be understood as "validation error rates".

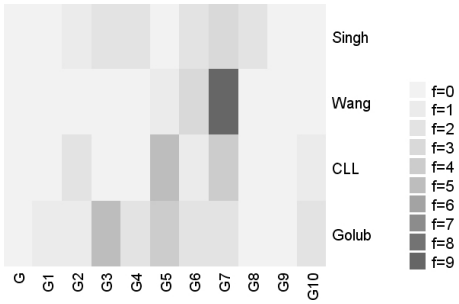

Figure 2: Frequency distribution of the variants of rlda.TG yielding the lowest error rates. The frequencies f are summed over three variable selection methods (t-test, Limma, Wilcoxon test) and four numbers of genes (100, 200, 500, 1000). Note that the "best" variant is not necessarily unique, i.e. the sum of the row-specific frequencies can be > 12.

4 Discussion

As illustrated in Section 3 based on the example of regularized LDA, the four investigated sources of over-optimism may yield substantially over-optimistic results. Beyond the four mechanisms outlined in this paper, 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. Both prediction measures are reported in Additional File 2. 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.

In our study, the optimistic results obtained with the selected variants of RLDA in the selected settings turn out to break down when validated based on "fresh" validation data sets. This indicates that the seemingly favorable results were rather the consequence of intense optimization than the

illustration of a real superiority of the new method. In a nutshell, let us point out possible reasons explaining the disappointing performance of the initially promising idea. A general finding of Bickel and Levina (2004) is that the DLDA highly outperforms the standard LDA in "huge-dimensional" data situations. Assuming independence between the predictor variables hence does not impair the classification performance, but rather yields improvement when $n \ll p$. This phenomenon has often been reported in the literature (Dudoit et al., 2002; Domingos and Pazzani, 1997), and it is shown under broad conditions by Bickel and Levina (2004). Our results confirm this finding in the sense that incorporating between-genes correlations tends to yield higher error rates with increasing p^*.

Another aspect to be considered is whether the assumptions underlying the new approach do apply, i.e. whether these assumptions are consistent (at least not evidently inconsistent) with intrinsic properties of the investigated data. Our own method postulates that genes from the same pathway tend to be more correlated than genes from different pathways. From the current point of view, however, the assumption that the between-genes correlation structure is reflected in KEGG pathways and vice versa is a widespread but dubious assumption on the part of (bio)statisticians. More precisely, this assumption is dubious for the correlation, which is a measure of *linear* association. Hence, considering non-linear association measures might help to uncover the interrelation between KEGG pathways and the between-genes association structure, and might thus lead to a more adequate modelling of the latter.

Taken together, these aspects might explain why RLDA based on $\widehat{\mathbf{\Sigma}}_{\mathbf{SHIP}}$ does not improve the classification accuracy in terms of prediction error rate. Note, however, that this negative finding could merely be made through an appropriate validation of the new algorithm. Without proper validation, we could have obtained a "false positive result". That said, the proposed shrinkage estimator based on target G could lead to interesting applications in other contexts like, e.g. canonical correlation analysis. The "disappointing results" reported in this article refer solely to the combination between target G, KEGG, and linear discriminant analysis – not to the individual components of the combination.

5 Conclusion

In this paper, we demonstrate quantitatively that a combination of various interrelated optimization mechanisms may yield substantially biased results and over-optimistic conclusions on the superiority of a new method. Of course, the content of a methodological article should not be reduced to the effective improvement of accuracy on real data sets. Other aspects of new methods need to be considered, such as their conceptual simplicity, computational efficiency, interpretability, flexibility, ability to generalize or fit in a global framework, the absence of strong assumptions or, most importantly, the originality of the addressed research question. Still we claim that, when improvement of accuracy is presented as the major contribution, it should be validated using independent data sets that were not used during the development of the new method.

Funding This project was supported by grant BO3139 to ALB from the German Science Foundation ("DFG Einzelförderung"), by the LMU-innovativ Project BioMed-S: Analysis and Modelling of Complex Systems in Biology and Medicine, and by the French-Bavarian Cooperation Center for Universities (CCUFB-BFHZ).

References

Bickel, P. J., Levina, E., 2004. Some theory for Fisher's linear discriminant function, "naive bayes", and some alternatives when there are many more variables than observations. Bernoulli 10, 989–1010.

Binder, H., Schumacher, M., 2009. Incorporating pathway information into boosting estimation of high-dimensional risk prediction models. BMC Bioinformatics 10, 18.

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

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., 2008. Evaluating microarray-based classifiers: an overview. Cancer Informatics 6, 77–97.

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

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.

Domingos, P., Pazzani, M., 1997. On the optimality of the simple bayesian classifier under zero-one loss. Machine Learning 29, 103–130.

Dudoit, S., Fridlyand, J., Speed, T. P., 2002. Comparison of discrimination methods for the classification of tumors using gene expression data. Journal of the American Statistical Association 97, 77–87.

Efron, B., Morris, C., 1977. Stein's Paradox in Statistics. Scientific American 236, 119–127.

Friedman, J. H., 1989. Regularized discriminant analysis. Journal of the American Statistical Association 84, 165–175.

Golub, T., 2010. golubEsets. R package version 1.4.7.
URL `http://bioconductor.org/packages/data/experiment/html/golubEsets.html`

Guillemot, V., Brusquet, L. L., Tenenhaus, A., Frouin, V., 2008. Graph-Constrained Discriminant Analysis of functional genomics data. IEEE International Conference on Bioinformatics and Biomedicine Worshops.

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

Hall, P., Xue, J.-H., 2010. Incorporating prior probabilities into high-dimensional classifiers. Biometrika 97, 31–48.

Jacob, L., Obozinski, G., Vert, J.-P., 2009. Group Lasso with Overlap and Graph Lasso. International Conference on Machine Learning (ICML 26).

Kanehisa, M., Goto, S., 2000. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Research 28, 27–30.

Ledoit, O., Wolf, M., 2003. Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. Journal of Empirical Finance 10, 603–621.

Ledoit, O., Wolf, M., 2004. Honey, I shrunk the sample covariance matrix. Journal of Portfolio Management 31, 110–119.

Li, C., Li, H., 2008. Network-constrained regularization and variable selection for analysis of genomic data. Bioinformatics 24, 1175–1182.

Penrose, R., 1955. A generalized inverse for matrices. Proceedings of the Cambridge Philosophical Society 51, 406–413.

Rapaport, F., Zinovyev, A., Dutreix, M., Barillot, E., Vert, J.-P., 2007. Classification of microarray data using gene networks. BMC Bioinformatics 8, 35.

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.

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., 2002. Gene expression correlates of clinical prostate cancer behavior. Cancer Cell 1, 203–209.

Slawski, M., zu Castell, W., Tutz, G., 2010. Feature selection guided by structural information. Annals of Applied Statistics 4, (in press).

Smyth, G., 2004. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. Statistical Applications in Genetics and Molecular Biology 3, 3.

Stein, C., 1955. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability.

Tai, F., Pan, W., 2007a. Incorporating prior knowledge of gene functional groups into regularized discriminant analysis of microarray data. Bioinformatics 23, 3170–3177.

Tai, F., Pan, W., 2007b. Incorporating prior knowledge of predictors into penalized classifiers with multiple penalty terms. Bioinformatics 23, 1775–1782.

Tibshirani, R., Hastie, T., Narasimhan, B., Chu, G., 2002. Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proceedings of the National Academy of Sciences 99, 6567–6572.

Wang, Y., Klijn, J., Zhang, Y., 2005. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. Lancet 365, 671–679.

Whalen, E., 2010. CLL. R package version 1.2.8.
URL `http://www.bioconductor.org/packages/2.5/data/experiment/html/CLL.html`

Yousef, M., Ketany, M., Manevitz, L., Showe, L. C., Showe, M. K., 2009. Classification and biomarker identification using gene network modules and support vector machines. BMC Bioinformatics 10, 337.

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.