BIOINFORMATICS ORIGINAL PAPER

Vol. 26 no. 16 2010, pages 1990–1998
doi:10.1093/bioinformatics/btq323

Gene expression

Advance Access publication June 26, 2010

Over-optimism in bioinformatics: an illustration

Monika Jelizarow[1], Vincent Guillemot[1,2], Arthur Tenenhaus[2], Korbinian Strimmer[3] and Anne-Laure Boulesteix[1,*]

[1]Department of Medical Informatics, Biometry and Epidemiology, University of Munich, Marchioninistr. 15, 81377 Munich, Germany, [2]SUPELEC Sciences des Systèmes (E3S)-Department of Signal Processing and Electronics Systems - 3, rue Joliot Curie, Plateau de Moulon, 91192 Gif-sur-Yvette Cedex, France and [3]Department of Medical Informatics, Statistics and Epidemiology, University of Leipzig, Härtelstr. 16-18, 04107 Leipzig, Germany

Associate Editor: John Quackenbush

ABSTRACT

Motivation: In statistical bioinformatics research, different optimization mechanisms potentially lead to 'over-optimism' in published papers. So far, however, a systematic critical study concerning the various sources underlying this over-optimism is lacking.

Results: We present an empirical study on over-optimism using high-dimensional classification as example. Specifically, we consider a 'promising' new classification algorithm, namely linear discriminant analysis incorporating prior knowledge on gene functional groups through an appropriate shrinkage of the within-group covariance matrix. While this approach yields poor results in terms of error rate, we quantitatively demonstrate that it can artificially seem superior to existing approaches if we 'fish for significance'. The investigated sources of over-optimism include the optimization of datasets, of settings, of competing methods and, most importantly, of the method's characteristics. 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 always be demonstrated on independent validation data.

Availability: The R codes and relevant data can be downloaded from http://www.ibe.med.uni-muenchen.de/organisation/mitarbeiter/020_professuren/boulesteix/overoptimism/, such that the study is completely reproducible.

Contact: boulesteix@ibe.med.uni-muenchen.de

Received on April 30, 2010; revised on June 8, 2010; accepted on June 10, 2010

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 article 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 dataset such that their new method works better than existing approaches, yielding a so-called 'dataset bias'. While a method cannot reasonably be expected to yield 'universally better' results in all datasets, it would be wrong to report only favorable datasets 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 dataset' in this article.

The second source of over-optimism, which is related to the optimal choice of the dataset 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 with existing benchmark approaches. This mechanism, which is strongly related to data overfitting, is termed as 'optimization of the settings' in this article.

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' in this article.

Finally, researchers often tend to optimize their new algorithms to the datasets 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 datasets. Hence, the superiority of the novel method over an existing method [as measured, e.g. through the difference between

[*]To whom correspondence should be addressed.

the cross-validation (CV) 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 dataset yields a downwardly biased error estimate commonly termed as apparent error. In the same spirit, computing CV 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 dataset may lead to optimistically biased results in the sense that the new algorithm's characteristics overfit the used dataset. This source of over-optimism is termed as 'optimization of the method's characteristics' in this article.

The four mechanisms discussed above may lead to over-optimistic conclusions regarding the superiority of the new method compared with 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 article aims at filling this gap. 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 (LDA). This approach combines a simple and well-established statistical method, regularized discriminant analysis (DA), 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 (Binder and Schumacher, 2009; Guillemot *et al.*, 2008; Hall and Xue, 2010; Jacob *et al.*, 2009; Li and Li, 2008; Rapaport *et al.*, 2007; Slawski *et al.*, 2010; Tai and Pan, 2007a, b; Yousef *et al.*, 2009).

Intriguingly, 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. Note that we could have used virtually any method to illustrate these mechanisms of over-optimism. However, classification with prior knowledge addresses a non-trivial and still unanswered research question within an evolving bioinformatics field. Based on this example, we demonstrate quantitatively that optimization of the dataset, 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 are convinced that most biased results are presented by mistake and that the involved researchers are disposed to make efforts toward better practice. It would be naive to believe that over-optimism in published research can be completely avoided, but we feel that a quantitative demonstration of the optimistic bias affecting methodological research may perhaps increase awareness on such problems and give many researchers food for thoughts.

The remainder of this article 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 datasets and the different interpretations depending on whether one fishes for significance or not. Other potential sources of biases, possible explanations for the disappointing error rates of the promising idea and further perspectives are presented in Section 4.

2 METHODS

2.1 A 'promising idea'

In this section, we describe the technical details of the classification method which we later use to illustrate the diverse sources and pitfalls of over-optimism. Readers who are not interested in the methodological part may skip this section.

2.1.1 DA and regularization We consider a high-dimensional dataset with continuous predictors such as microarray gene expression data, with the aim to predict a categorical response variable of interest, e.g. the disease status or the long-term disease outcome.

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 it is assumed that $\boldsymbol{\Sigma}_1=\cdots=\boldsymbol{\Sigma}_c$, yielding the so-called 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 from our web site.

In the high-dimensional setting considered here the pooled covariance estimator $\tilde{\mathbf{S}}$ is singular and hence not invertible. In regularized LDA (RLDA), the singularity problem is resolved by employing a shrinkage (Efron and Morris, 1977; Stein, 1955) rather than an empirical estimator of the covariance—see the seminal paper by Friedman (1989). More recently, variants of RLDA are considered, e.g., in Guo *et al.* (2007) and Ahdesmäki and Strimmer (2010).

2.1.2 RLDA with KEGG An increasingly popular approach is to regularize the within-class covariance by incorporating external biological knowledge from databases such as the Kyoto Encyclopedia of Genes and Genomes (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 subdatabases. KEGG PATHWAY as one of these subdatabases 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

Table 1. Overview of targets D (diagonal, unequal variance), F (constant correlation) and G (where $\bar{r}$ is the average of sample correlations)

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

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.

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 modeling 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 $\widehat{\mathbf{\Sigma}}_{\text{SHIP}}$ *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 which we refer to as SHIP standing for SHrinking and Incorporating Prior knowledge. The resulting covariance estimator $\widehat{\mathbf{\Sigma}}_{\text{SHIP}}$ is based on the Stein-type shrinkage estimator discussed by Ledoit and Wolf (2003, 2004) and applied to correlation 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 $\widehat{\mathbf{\Sigma}}^*=\lambda\mathbf{T}+(1-\lambda)\mathbf{S}$, 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 estimator' 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 from our web site 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 article, we consider targets D and 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 $\bar{r}$, as in Tai and Pan (2007a). 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 from our web site). The shrinkage covariance estimator $\widehat{\mathbf{\Sigma}}_{\text{SHIP}}$ is implemented in the R package 'SHIP' which is publicly available from the companion web site and from the CRAN.

2.1.4 LDA 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 $r=1,\dots,c$ and subsequently pool these within-class shrinkage estimators according to the standard procedure known from LDA. See Additional File 1 from our web site 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).

2.2 Design of the study

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

2.2.2 Settings Prediction accuracy is estimated using the well-established 10 times 5-fold CV evaluation scheme. The 5-fold CV is repeated 10 times in order to achieve more stable results (Boulesteix *et al.*, 2008; Braga-Neto and Dougherty, 2004). We focus on the average misclassification rate as a measure of prediction accuracy, i.e. the average test error obtained over all $10\times5=50$ 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 dataset, 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 $3\times4=12$ combinations of selection procedures and numbers of selected genes.

2.2.3 Competing methods For comparison purposes, we furthermore apply the diagonal LDA (DLDA), the nearest shrunken centroids (NSC) method 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 3-fold CV.

2.2.4 Method's characteristics When developing a new algorithm, researchers often adapt their method sequentially depending on their experiences with example datasets 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 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 article 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 RLDA 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 10 following variants of rlda.TG termed as rlda.TG$^{(1)}$, ..., rlda.TG$^{(10)}$.

These 10 variants can be divided into two groups. The first group comprises rlda.TG$^{(1)}$ to rlda.TG$^{(7)}$ which differ in the assignment of 'problematic' genes. By problematic genes, we mean either genes that are in no functional group or genes that are in at least two different functional groups, thus making their assignment to a functional group impossible or arbitrary, respectively. In contrast to the original variant rlda.TG, variant rlda.TG$^{(1)}$ simply excludes genes that are not in any gene functional group (∼50% in each dataset) from the analysis. Variant rlda.TG$^{(2)}$ differs from rlda.TG in the treatment of genes occurring in multiple gene functional groups: they are simply eliminated from the dataset. In contrast, both rlda.TG$^{(3)}$ and rlda.TG$^{(4)}$ handle these genes similarly to Tai and Pan (2007a): if a gene occurs in multiple gene functional groups, it is considered as belonging to the gene functional group with the smallest or largest number of genes, respectively. If the smallest (respectively largest) gene functional group is not unique, rlda.TG$^{(3)}$ and rlda.TG$^{(4)}$ choose one of them by chance and consider it as the smallest (respectively largest). Note that, while these two variants may seem arbitrary, they have been applied in a previous relevant publication (Tai and Pan, 2007a). Trying them out thus appears to be natural. The methods rlda.TG$^{(5)}$ to rlda.TG$^{(7)}$ are obtained by combining rlda.TG$^{(1)}$ (i.e. eliminating genes that are in no functional group) with rlda.TG$^{(2)}$, rlda.TG$^{(3)}$ and rlda.TG$^{(4)}$ (to handle genes occurring in more than one functional group). The second group comprises rlda.TG$^{(8)}$, rlda.TG$^{(9)}$ and rlda.TG$^{(10)}$ which are based on a redefinition of the covariance target G. Variant rlda.TG$^{(8)}$ involves two parameters for the correlation (the average $\bar{r}_+$ of the positive correlations and the average $\bar{r}_-$ of the negative correlations) instead of the single parameter $\bar{r}$, to take into account that genes from the same pathway might be negatively correlated. The variant rlda.TG$^{(9)}$ completely ignores negative correlations and computes the average correlation $\bar{r}$ based on positive correlations only. Finally, rlda.TG$^{(10)}$ tests the correlations (at the significance level 0.05) and sets the non-significant correlations to zero before the mean correlation $\bar{r}$ is computed.

Table 2. Overview of the CV errors obtained for rlda.TG where p^* denotes the number of selected genes

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

The bold values indicate the minimum values.

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 datasets. 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 datasets.

The four optimization mechanisms are introduced sequentially and independently of each other in Section 1. However, they are in fact tightly linked 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 scenarios for illustration purposes. We are aware of the many other potential schemes, but an exhaustive study would go beyond the scope of this article. 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 from our web site.

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

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$^{(10)}$. For a given dataset, 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 datasets. The bold numbers indicate the minimal error rate for each dataset. The standard errors of the error rates over the CV iterations range

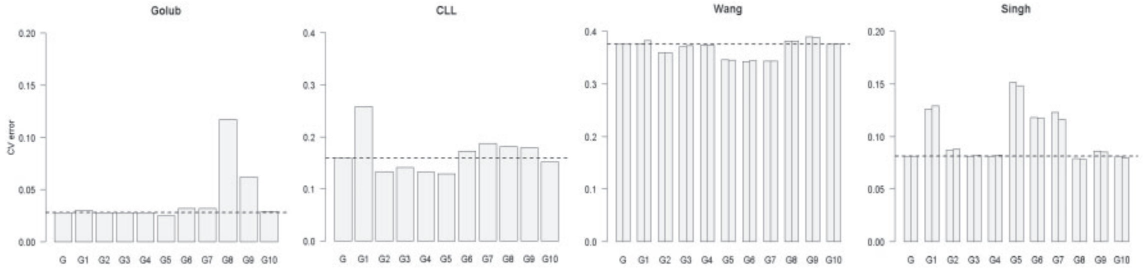

Fig. 1. Overview of the CV error rates of the different variants of rlda.TG, obtained for all datasets within the corresponding optimal settings s_{opt}. The dashed line indicates the value obtained for rlda.TG within the data-specific s_{opt}. Note that for both the Wang and the Singh data the optimal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

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}

	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

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 datasets, although settings with small p^* tend to yield smaller error rates in general.

Researchers who 'fish for significance' would select the setting yielding the minimal error rate for the dataset they consider, thus inducing an optimistic bias through 'optimization of the settings'.

3.2.2 Optimization of the method's characteristics Moreover, they 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 dataset. Especially for the CLL and the Wang dataset, 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 with the new algorithm. For each of the four datasets, 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 DLDA). 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 NSC method and the SVM. Obviously, these competing approaches 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 was 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 was that it performs better than other LDA approaches, we would consider all LDA-based competitors. If the hypothesis was that the incorporation of correlations is useful, we would consider rlda.TD. If the hypothesis was 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 dataset Some researchers may also 'optimize the dataset' 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 more favorable to our new method than the other three datasets. This is probably due to the very small size ($n = 22$) implying a high variability and thus

Table 4. Performance of the optimal variants M_{opt} of rlda.TG within the optimal settings s_{opt} selected in each of the four datasets

	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**

The figures outside the diagonal can be understood as 'validation error rates'.

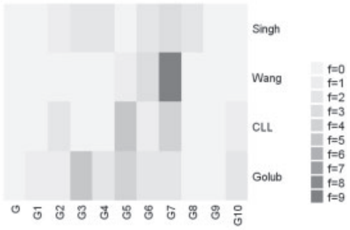

Fig. 2. Frequency of selection of the 11 investigated variants of rlda.TG over the three variable selection methods (*t*-test, Limma, Wilcoxon test) and four numbers of genes (100, 200, 500, 1000), i.e. over 3 × 4 = 12 settings. By 'selection' we mean that the variant yields the smallest error rate over the 11 variants. For example, in the Wang dataset, the lowest error rate is reached by rlda.TG7 in 9 of the 12 considered settings and by rlda.TG6 in only three settings. Note that the 'best' variant is not necessarily unique, i.e. for a specific dataset (row) the frequencies' sum may be >12.

stronger optimization effects. The optimization of the dataset 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 datasets were analyzed independently of each other. For each dataset, 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 datasets that were not used for the algorithm's development. Table 4 shows the CV error rates of the four combinations of optimal settings and optimal variant when applied on the four datasets. Whereas the error rates on 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 datasets. 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 × 4 = 12) in which each of the variants rlda.TG, rlda.TG$^{(1)}$,..., rlda.TG$^{(10)}$ yields the lowest error rate, for each dataset separately. It can be seen that the 'optimal variant' strongly depends on the dataset (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 datasets 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.

4 DISCUSSION

4.1 Other sources of bias

As illustrated in Section 3 based on the example of RLDA, the four investigated sources of over-optimism may yield substantially over-optimistic results. Beyond the four mechanisms outlined in this article, 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 from our web site. The applied normalization technique may also affect the results (Lim *et al.*, 2007) 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 in this article would not help in this case, since the error in the competing methods would also affect the validation phase. To conclude this enumeration of sources of bias, we point out that the occasional publication of negative results in methodological journals may in the long run encourage researchers to be less biased.

4.2 On CV as a potential solution

Section 3 demonstrates that validation based on independent datasets 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 variants rlda.TG, rlda.TG$^{(1)}$, ..., rlda.TG$^{(10)}$). In this case, however, the application of a CV procedure is much more problematic in practice because the different variants of rlda.TG (rlda.TG$^{(1)}$, ..., rlda.TG$^{(10)}$ in our example) are usually not investigated simultaneously. 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. Moreover, 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 10 variants are considered simultaneously in the current article, 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 CV is that it ensures a more stringent separation between data used for development and data used for evaluation. CV 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 dataset after having developed a method. Moreover, validation based on other independent datasets has the considerable advantage that it takes the variability between datasets into account, a very important aspect discussed in Section 4.3.

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; and (iii) select the best variant of the method (among rlda.TG, rlda.TG$^{(1)}$, ..., rlda.TG$^{(10)}$) internally rapidly becomes computationally intractable and, in general, cannot be recommended in practice.

4.3 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 (Hanczar *et al.*, 2010; Yousefi *et al.*, 2010). Optimization biases occur because CV error estimates have large unknown variance (Braga-Neto and Dougherty, 2004), and 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 dataset 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 datasets. 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 dataset. 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 dataset because the internal variance (i.e. the variance within a single dataset) can be estimated but not the external variance (i.e. the variance between datasets). 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 datasets. However, the estimation of external variability based on a small number of datasets is also a non-trivial issue.

4.4 On simulations as a potential solution

Another way to take this 'between-datasets' 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 datasets. Our example with KEGG-based RLDA 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: the incorporation of cluster structure through target G and the usefulness of KEGG in this context. 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.

4.5 Potential pitfalls of the promising idea

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 datasets. 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 error rates 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. 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 (Domingos and Pazzani, 1997; Dudoit *et al.*, 2002), 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

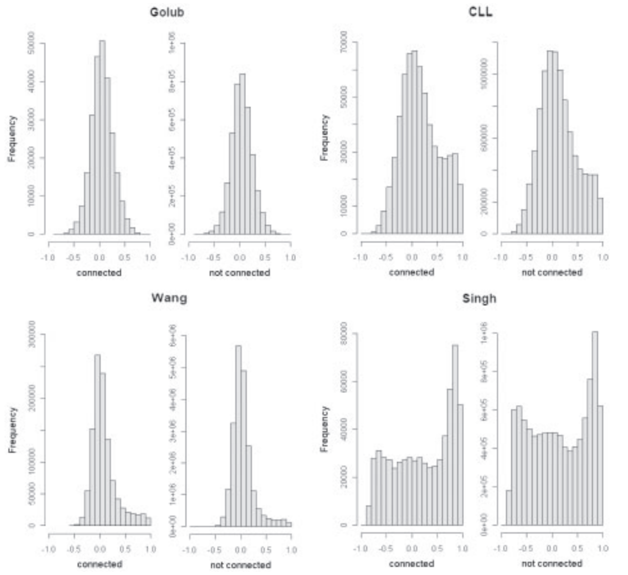

Fig. 3. Illustration of (i) correlations between connected genes and (ii) correlations between not connected genes by means of histograms. The illustration is given for the datasets Golub ($n^{(i)} = 255\,441$ versus $n^{(ii)} = 4\,115\,005$), CLL ($n^{(i)} = 606\,903$ versus $n^{(ii)} = 9\,901\,917$, Wang ($n^{(i)} = 1\,096\,193$ versus $n^{(ii)} = 20\,985\,142$) and Singh ($n^{(i)} = 606\,903$ versus $n^{(ii)} = 9\,901\,917$), where $n^{(i)}$ and $n^{(ii)}$ denote the number of available correlations between connected genes and not connected genes, respectively.

a widespread but vague assumption on the part of (bio)statisticians. Histograms of the Pearson's correlations between connected genes (i.e. genes sharing at least one pathway) and not connected genes (i.e. genes that do not have any pathway in common) are depicted in Figure 3, separately for each dataset. In this rather limited study based on only four datasets, genes belonging to the same pathway do not seem to be noticeably more correlated. To some extent, the vague assumption made by biostatisticians might be inappropriate because Pearson's correlation is a measure of *linear* association. For example, genes from the same pathway might correlate only in case the pathway is activated, hence inducing a complex dependence pattern that cannot be captured by linear correlation measures. Considering more complex association structures beyond Pearson's correlation might provide more insight into the interrelation between KEGG pathways and the between-genes association (Hausser and Strimmer, 2009).

Finally, let us point out that the 'disappointing results' reported in this article refer solely to the investigated combination between target G, KEGG and (R)LDA—neither to the individual components of the combination nor to further more sophisticated combinations.

4.6 Epistemological considerations

More and more applications in bioinformatics and systems biology require systematic integration of data from different sources. Consequently, statistical approaches for incorporating prior biological knowledge are becoming more important. However, as yet it is unclear how to properly take account of this information and how to best handle the data from multiple sources. Furthermore, this raises questions on the consistency and also on the relevance of the available information. Facing the lack of *complete* biological knowledge, one must be aware that the employed information might not be specific enough for the phenotype considered.

More generally, the need for a sound epistemological basis of statistical methods for high-dimensional biological data has to be addressed. Even though the relevance of this problem has been pointed out in the literature (Braga-Neto, 2007; Dougherty, 2008; Keller, 2005; Mehta *et al.*, 2004), it is not adequately considered yet in biostatistical and bioinformatical research. A key aspect of method development is that the apparent usefulness (as measured, e.g. by the error rate) of a method cannot be equated with the method's validity. In particular, the question of validity has to be resolved first, requiring a detailed exposition of the new method's characteristics and properties including the biological relevance of these characteristics and properties. The substantive context is indeed crucial in the context of methodological research (Keiding, 2010). This is outlined by Mehta *et al.* (2004) who propose a framework in which the epistemological foundation of statistical methods for high-dimensional data can be evaluated. Only once this has been done, the new method can be assessed objectively.

5 CONCLUSION

In this article, 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. Over-optimism is widespread in statistical methods development (Hand, 2006) and also in bioinformatics and systems biology. Therefore, to properly evaluate a method other aspects need to be considered in addition to improvement of accuracy on real datasets, such as their conceptual simplicity, computational efficiency, interpretability, flexibility, ability to generalize or fit in a global framework, the absence of strong assumptions, the originality of the addressed research question or, most importantly, the validity of the underlying model. As noted by Mehta *et al.* (2004), 'illustration with single datasets of unknown nature, though interesting, is not a sound epistemological foundation for method development'. Hence, when improvement of accuracy is presented as the major contribution, it should be validated using independent datasets that were not used during the development of the new method. There is, of course, no *uniformly* best method which can be shown to perform best on every real dataset. However, precisely because of that an adequate validation including a report of both positive and negative results is essential to substantiate and objectify the statements on a method's strength.

ACKNOWLEDGEMENTS

We thank the reviewers for the constructive comments, leading to a considerable improvement of the manuscript. We further thank Christoph Bernau for helpful remarks.

Funding: German Science Foundation ("DFG Einzelförderung" grant BO3139/2-1 to A.L.B.); the LMU-innovativ Project BioMed-S; French-Bavarian Cooperation Center for Universities (CCUFB-BFHZ).

Conflict of Interest: none declared.

REFERENCES

Ahdesmäki,M. and Strimmer,K. (2010) Feature selection in omics prediction problems using cat scores and false non-discovery rate control. *Ann. Appl. Stat.*, **4**, 503–519.

Ambroise,C. and McLachlan,G.J. (2002) Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc. Natl Acad. Sci. USA*, **99**, 6562–6566.

Bickel,P.J. and 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. and Schumacher,M. (2009) Incorporating pathway information into boosting estimation of high-dimensional risk prediction models. *BMC Bioinformatics*, **10**, 18.

Boulesteix,A.L. and Strobl,C. (2009) Optimal classifier selection and negative bias in error rate estimation: An empirical study on high-dimensional prediction. *BMC Med. Res. Methodol.*, **9**, 85.

Boulesteix,A.L. *et al.* (2008) Evaluating microarray-based classifiers: an overview. *Cancer Informat.*, **6**, 77–97.

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

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

Braga-Neto,U.M. (2007) Fads and fallacies in the name of small-sample microarray classification. *IEEE Sign. Process. Mag.*, **24**, 91–99.

Daumer,M. *et al.* (2008) Reducing the probability of false positive research findings by pre-publication validation: Experience with a large multiple sclerosis database. *BMC Med. Res. Methodol.*, **8**, 18.

Domingos,P. and Pazzani,M. (1997) On the optimality of the simple Bayesian classifier under zero-one loss. *Mach. Learn.*, **29**, 103–130.

Dougherty,E.R. (2008) On the epistemological crisis in genomics. *Curr. Genomics*, **9**, 69–79.

Dudoit,S. *et al.* (2002) Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Am. Stat. Assoc.*, **97**, 77–87.

Efron,B. and Morris,C. (1977) Stein's paradox in statistics. *Sci. Am.*, **236**, 119–127.

Friedman,J.H. (1989) Regularized discriminant analysis. *J. Am. Stat. Assoc.*, **84**, 165–175.

Golub,T. (2010) *golubEsets*. R package version 1.4.7. Available at http://bioconductor.org/packages/2.6/data/experiment/html/golubEsets.html.

Guillemot,V. *et al.* (2008) Graph-Constrained Discriminant Analysis of functional genomics data. In *IEEE International Conference on Bioinformatics and Biomedicine Worshops*. Philadelphia, USA, pp. 207–210.

Guo,Y. *et al.* (2007) Regularized linear discriminant analysis and its application in microarrays. *Biostatistics*, **8**, 86–100.

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

Hanczar,B. *et al.* (2007) Decorrelation of the true and estimated classifier errors in high-dimensional settings. *EURASIP J. Bioinformatics Syst. Biol.*, **207**, 38473.

Hanczar,B. *et al.* (2010) Small-sample precision of roc-related estimates. *Bioinformatics*, **26**, 822–830.

Hand,D.J. (2006) Classifier technology and the illusion of progress. *Stat. Sci.*, **21**, 1–14.

Hausser,J. and Strimmer,K. (2009) Entropy inference and the James-Stein estimator, with application to nonlinear gene association networks. *J. Mach. Learn. Res.*, **10**, 1469–1484.

Jacob,L. *et al.* (2009) Group Lasso with Overlap and Graph Lasso. In *International Conference on Machine Learning (ICML 26)*. Montreal, Canada, pp. 433–440.

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

Keiding,N. (2010) Reproducible research and the substantive context. *Biostatistics*, **11**, 376–378.

Keller,E.F. (2005) Revisiting scale-free networks. *BioEssays*, **27**, 1060–1068.

Ledoit,O. and Wolf,M. (2003) Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *J. Empir. Finan.*, **10**, 603–621.

Ledoit,O. and Wolf,M. (2004) Honey, I shrunk the sample covariance matrix. *J. Portf. Manag.*, **31**, 110–119.

Lim,W.K. *et al.* (2007) Comparative analysis of microarray normalization procedures: effects on reverse engineering gene networks. *Bioinformatics*, **23**, 282–288.

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

Mehta,T. *et al.* (2004) Towards sound epistemological foundations of statistical methods for high-dimensional biology. *Nat. Genet.*, **36**, 943–947.

Penrose,R. (1955) A generalized inverse for matrices. *Proc. Camb. Philo. Soc.*, **51**, 406–413.

Rapaport,F. *et al.* (2007) Classification of microarray data using gene networks. *BMC Bioinformatics*, **8**, 35.

Rocke,D.M. *et al.* (2009) Papers on normalization, variable selection, classification or clustering of microarray data. *Bioinformatics*, **25**, 701–702.

Schäfer,J. and Strimmer,K. (2005) A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Stat. Appl. Genet. Mol. Biol.*, **4**, Article 32.

Singh,D. *et al.* (2002) Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, **1**, 203–209.

Slawski,M. *et al.* (2010) Feature selection guided by structural information. *Ann. Appl. Stat.*, **4**, in press.

Smyth,G. (2004) Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.*, **3**, 3.

Stein,C. (1955) Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*. University of California Press, Statistical Laboratory of the University of California, Berkeley, USA, pp. 197–206.

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

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

Tibshirani,R. *et al.* (2002) Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Natl Acad. Sci. USA*, **99**, 6567–6572.

Wang,Y. *et al.* (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. Available at http://bioconductor.org/packages/2.6/data/experiment/html/CLL.html.

Yousefi,M.R. *et al.* (2010) Reporting bias when using real datasets to analyze classification performance. *Bioinformatics*, **26**, 68–76.

Yousef,M. *et al.* (2009) Classification and biomarker identification using gene network modules and support vector machines. *BMC Bioinformatics*, **10**, 337.