Additional File 1

(Over-optimism in bioinformatics: an illustration)

1 Overview of linear discriminant analysis

Suppose there are c different classes being indicated by $Y \in \{1, ..., c\}$ and a vector $\boldsymbol{x} \in \mathbb{R}^p$ of predictors. Decision theory for classification tells us that for an optimal classification based on $\boldsymbol{x}$, we need to know the class posteriors $\mathbb{P}(Y|\boldsymbol{x})$. Suppose f_r is the conditional density of $\boldsymbol{x}$ in class $Y = r$, and let us denote π_r as the prior probability of class r, with

$$\sum_{r=1}^{c} \pi_r = 1.$$

The application of the Bayes theorem leads to

$$\mathbb{P}(Y = r|\boldsymbol{x}) = \frac{f_r(\boldsymbol{x})\pi_r}{\sum_{j=1}^{c} \pi_j f_j(\boldsymbol{x})}, \tag{1}$$

for a particular vector $\boldsymbol{x} = (x_1, ..., x_p)^\top$.
The quantities f_r and π_r are needed to obtain $\mathbb{P}(Y = r|\boldsymbol{x})$. In linear and quadratic discriminant analysis f_r is modeled as a multivariate normal density (for $r = 1, \ldots, c$), i.e.

$$\begin{aligned} \boldsymbol{x}|(Y = r) &\sim \mathcal{N}(\boldsymbol{\mu}_r, \boldsymbol{\Sigma}_r) && (2) \\ \text{i.e. } f_r(\boldsymbol{x}) &= \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}_r|^{1/2}} e^{\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_r)^\top \boldsymbol{\Sigma}_r^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_r)}, && (3) \end{aligned}$$

with $\boldsymbol{\Sigma}_r$ denoting the covariance matrix and $\boldsymbol{\mu}_r$ the mean in class r. Linear discriminant analysis (LDA) arises in the special case when we assume that the classes have a common covariance matrix $\boldsymbol{\Sigma}_r = \boldsymbol{\Sigma}$, for $r = 1, \ldots, c$. To compare the posterior probability of two classes r and l, we can look at their log-ratio and obtain after a short calculation

$$\log\left(\frac{\mathbb{P}(Y = r|\boldsymbol{x})}{\mathbb{P}(Y = l|\boldsymbol{x})}\right) = \log\left(\frac{\pi_r}{\pi_l}\right) - \frac{1}{2}(\boldsymbol{\mu}_r + \boldsymbol{\mu}_l)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_r - \boldsymbol{\mu}_l) + \boldsymbol{x}^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_r - \boldsymbol{\mu}_l), \tag{4}$$

which is linear in $\boldsymbol{x}$, hence the term "linear discriminant analysis". As can be seen from the previous equation, the discriminant function can be formulated as

$$d_r(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_r - \frac{1}{2}\boldsymbol{\mu}_r^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_r + \log(\pi_r), \tag{5}$$

where a new observation $\boldsymbol{x}^*$ is assigned to class $\tilde{r}$ if

$$\tilde{r} = \underset{r=1,\ldots,c}{\operatorname{argmax}}\, d_r(\boldsymbol{x}^*). \tag{6}$$

In practice, to build discriminant functions we need to estimate the parameters of the multivariate distributions from a finite sample $(\mathbf{x}_{i.}, y_i)_{i=1,\ldots,n}$. The parameters are usually estimated as follows.

- $\hat{\pi}_r = \frac{n_r}{n}$ where n_r is the number of observations with $y_i = r$
- $\hat{\boldsymbol{\mu}}_r = \frac{1}{n_r} \sum_{i:y_i=r} \mathbf{x}_{i.}$
- $\tilde{\mathbf{S}} = \frac{1}{n-c} \sum_{r=1}^{c} \sum_{i:y_i=r} (\mathbf{x}_{i.} - \hat{\boldsymbol{\mu}}_r)(\mathbf{x}_{i.} - \hat{\boldsymbol{\mu}}_r)^\top = \frac{1}{n-c} \sum_{r=1}^{c} (n_r - 1)\mathbf{S}_r.$

The $p \times p$ matrix $\tilde{\mathbf{S}}$ is usually referred to as the pooled empirical covariance matrix that can be written as a weighted sum of the $p \times p$ standard unbiased empirical within-class covariance matrices $\mathbf{S}_r$, $r = 1, ..., c$. The LDA now can be applied in a straightforward way in the $p \ll n$ case (i.e. the number p of predictor variables does not exceed the number n of observations). However, in the high-dimensional setting the covariance estimator $\tilde{\mathbf{S}}$ from 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 well-conditioned. For details we recommend Friedman's seminal work on regularized (Fisher's) discriminant analysis [1] and the shrunken centroids regularized discriminant analysis (SCRDA) by Guo et al. [2] which both are based on the widely employed shrinkage principle.

Furthermore, an increasingly popular approach is to regularize the within-class covariance by incorporating external biological knowledge on gene functions from databases like the **K**yoto **E**ncyclopedia of **G**enes and **G**enomes (KEGG) [3]. The motivation behind is to improve both the prediction accuracy and the results' interpretability. Within the scope of current scientific focus, we propose a further variant of RLDA incorporating biological knowledge on gene functional groups. An outline of our idea elaborated in [4] is given in 2 and 5, respectively.

2 The shrinkage estimator $\widehat{\boldsymbol{\Sigma}}_{\text{SHIP}}$

Starting from the methodological challenges arising from the $n \ll p$ data situation, we now propose a covariance estimation procedure we refer to as **SHIP**: **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 [5, 6, 7] and applied by Schäfer and Strimmer in the context of genomic data [8, 9]. Additionally, the new estimator incorporates prior biological knowledge on gene functional groups extracted from the database KEGG. Note that we first refer to a standard framework. The generalization to the special case of LDA requiring a pooled covariance estimator is discussed in 5.

In a nutshell, 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 with which the structured (i.e. low-dimensional)

covariance target $\mathbf{T} = (t_{ij})$ is shrunken towards the unstructured (i.e. high-dimensional) standard unbiased empirical covariance matrix $\mathbf{S} = (s_{ij})$, $i, j = 1, ..., p$. In this way, both the singularity problem is resolved and the covariance estimator is stabilized. Moreover, optimality is meant with respect to a quadratic loss function which is common and intuitive in statistical decision theory [10]. For details on the less intuitive asymptotic result we refer to Ledoit and Wolf.

Apparently, the concrete form of both the covariance target $\mathbf{T}$ and the optimal shrinkage intensity λ is unclear yet. We discuss these aspects below. Once $\mathbf{T}$ is chosen and λ is computed, some of the shrinkage estimator's nice properties are: (i) It is more efficient and more accurate than the empirical covariance matrix. (ii) It is positive definite and invertible which are crucial properties with regard to the estimation of the inverse of the true covariance matrix. (iii) It has guaranteed minimum mean squared error (MSE) resulting from the quadratic loss function [10]. (iv) It does not assume any fully specified distribution since merely second moments are required. These properties similarly hold for $\widehat{\mathbf{\Sigma}}_{\text{SHIP}}$ since $\widehat{\mathbf{\Sigma}}^*$ and $\widehat{\mathbf{\Sigma}}_{\text{SHIP}}$ only differ in terms of a covariance target $\mathbf{T}$ which - if suitably chosen - does not affect the properties from above.

3 Choice of the covariance target T

The covariance target $\mathbf{T}$ plays an essential role in the computation of the shrinkage estimator $\widehat{\mathbf{\Sigma}}_{\text{SHIP}}$. 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 [8]. In this paper, we deal with targets D and F (see Table 1).

Note that biological knowledge on gene functional groups has not been considered so far. To incorporate the latter from KEGG PATHWAY, we propose a modified version of target F where genes that are biologically connected have constant correlation $\bar{r}$. Hence, in order to obtain $\bar{r}$ we just account for the correlations between genes having at least one gene functional group in common. The resulting target G is defined in Table 1 (see below). In case a gene does not occur in any gene functional group, we assume this gene forming its own group with group size one which corresponds to Tai and Pan [11].

Unlike the diagonal target D both target F and target G do not necessarily fulfill the positive definiteness requirement. Hence, the resulting shrinkage estimator is not automatically positive definite unless target D is employed. A strategy to overcome this problem is computing the inverse by means of the "Moore-Penrose pseudoinverse" ; which can be applied to singular matrices and is based on the singular value decomposition [12]. The computation can be done using the function `pseudoinverse()` implemented in the open source `R` package `corpcor`.

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

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 occur in the same gene functional group.

4 The optimal shrinkage intensity λ

Having studied the choice of the covariance target **T**, we now address the selection of the optimal shrinkage intensity $\lambda \in [0, 1]$. In contrast to common approaches that are based on cross-validation [1], Markov Chain Monte Carlo (MCMC) or the bootstrap [13], Ledoit and Wolf propose an analytical determination of λ with its distinct advantage being the considerably lower computational effort. A detailed analytical derivation of λ including the theoretical background is given in [5, 8]. Nevertheless, the calculation of λ is not straightforward since it depends on the unobservable true covariance matrix. For the sake of convenience, let us continue with the formula for the estimated λ given in Schäfer and Strimmer [8]. It holds

$$\hat{\lambda} = \frac{\sum_{i=1}^{p}\sum_{j=1}^{p}\widehat{Var}(s_{ij}) - \widehat{Cov}(t_{ij}, s_{ij})}{\sum_{i=1}^{p}\sum_{j=1}^{p}(t_{ij} - s_{ij})^2}.$$

Since in finite samples $\hat{\lambda} \notin [0, 1]$ may occur, we truncate the estimated shrinkage intensity as $\hat{\lambda} \leftarrow \max(0, \min(1, \hat{\lambda}))$. In order to compute the estimator $\hat{\lambda}$ of the optimal shrinkage intensity, it is necessary to estimate the components of the given formula which in particular are $\widehat{Var}(s_{ij})$ and $\widehat{Cov}(t_{ij}, s_{ij})$. Let $\mathbf{x}_{ki}$ be the k-th observation of the variable (gene) x_i and $\bar{\mathbf{x}}_i = \frac{1}{n}\sum_{k=1}^{n}\mathbf{x}_{ki}$ its empirical mean. Now set $w_{kij} = (\mathbf{x}_{ki} - \bar{\mathbf{x}}_i)(\mathbf{x}_{kj} - \bar{\mathbf{x}}_j)$ and $\bar{w}_{ij} = \frac{1}{n}\sum_{k=1}^{n} w_{kij}$. Tedious calculations then yield the target-specific formulae for $\hat{\lambda}$ given below in Table 2.

$$\hat{\lambda}_D = \frac{\sum_{i\neq j}\widehat{Var}(s_{ij})}{\sum_{i\neq j} s_{ij}^2} \quad \hat{\lambda}_F = \frac{\sum_{i\neq j}\widehat{Var}(s_{ij}) - \bar{r}f_{ij}}{\sum_{i\neq j}(s_{ij} - \bar{r}\sqrt{s_{ii}s_{jj}})^2} \quad \hat{\lambda}_G = \frac{\sum_{i\neq j}\widehat{Var}(s_{ij}) - \sum_{i\sim j}\bar{r}f_{ij}}{\sum_{i\neq j}(s_{ij} - I(i\sim j)\bar{r}\sqrt{s_{ii}s_{jj}})^2}$$

Table 2: Overview of the target-specific estimators of the optimal shrinkage intensity, where $\widehat{Var}(s_{ij}) = \frac{n}{(n-1)^3}\sum_{k=1}^{n}(w_{kij} - \bar{w}_{ij})^2$, $\widehat{Cov}(s_{ij}, s_{lm}) = \frac{n}{(n-1)^3}\sum_{k=1}^{n}(w_{kij} - \bar{w}_{ij})(w_{klm} - \bar{w}_{lm})$ and $f_{ij} = \frac{1}{2}\{\sqrt{\frac{s_{jj}}{s_{ii}}}\widehat{Cov}(s_{ii}, s_{ij}) + \sqrt{\frac{s_{ii}}{s_{jj}}}\widehat{Cov}(s_{jj}, s_{ij})\}$. $I(\cdot)$ denotes the indicator function.

In case a covariance target only shrinks the off-diagonal elements of **S** and leaves the diagonal elements intact (e.g. targets D, F and G) we follow the suggestions of Schäfer and Strimmer and parameterize the covariance matrix in terms of variances and correlations rather than in variances and covariances, i.e. $\sigma_{ij} = r_{ij}\sqrt{\sigma_{ii}\sigma_{jj}}$. Then shrinkage is applied to the correlations. This procedure has the advantage that the off-diagonal elements determining the shrinkage intensity

are all at the same scale. The modified estimator of λ can thus be formulated by replacing all covariances by their corresponding correlations. For the purpose of computation, $\widehat{Var}(r_{ij})$ can be obtained by applying the formula for $\widehat{Var}(s_{ij})$ to the standardized data matrix. Analogously, applying the formula for $\widehat{Cov}(s_{ij}, s_{lm})$ to the standardized data matrix yields the desired estimator $\widehat{Cov}(r_{ij}, r_{lm})$.

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

So far, we have dealt with $\widehat{\mathbf{\Sigma}}_{\text{SHIP}}$ under the assumption that the observations come from one homogeneous population. Within the scope of LDA, however, where the predictor vectors fall into groups or classes, the previous procedure cannot be directly applied. Here, we briefly sketch how the idea of the shrinkage estimator from 2 can technically be included into the framework 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. We obtain

$$\hat{\mathbf{\Sigma}}^{*}_{\text{SHIP}} = \frac{1}{n-c} \sum_{r=1}^{c} (n_r - 1) \hat{\mathbf{\Sigma}}^{(r)}_{\text{SHIP}},$$

following the classical definition of the pooled covariance matrix.

6 Data sets

We analyze the four following real-life microarray data sets: two leukemia data sets CLL and Golub included in the packages 'CLL' [14] and 'golubEsets' [15], respectively, a breast cancer data set Wang by [16] and the prostate data set Singh by [17]. We preprocessed the two latter data sets ourselves with the GCRMA method using the raw data available from GEO. All data sets include a binary outcome variable which has to predicted based on gene expression data. A brief overview of the data sets' characteristics is given in Table 3.

Data set	n	$n_1 : n_2$	Classes	p	Raw data avail.	Collection	Normalization
CLL	22	14:8	proges. vs. stable	12 625	No	`hgu95av2`	GCRMA
Golub	72	47:25	ALL vs. AML	7 129	No	`hu6800`	MAS5
Singh	102	52:50	normal vs. tumor	12 625	Yes	`hgu95av2`	GCRMA
Wang	286	179:107	rel. vs. no rel.	22 283	Yes	`hgu133a`	GCRMA

Table 3: Overview of the four data sets' characteristics.

References

[1] J. H. Friedman. Regularized discriminant analysis. *Journal of the American Statistical Association*, 84:165–175, 1989.

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

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

[4] M. Jelizarow. Regularized Discriminant Analysis Incorporating Prior Knowledge on Gene Functional Groups. Master's thesis, Fakultät für Mathematik, Informatik und Statistik der Ludwig-Maximilians Universität München, 2009.

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

[6] O. Ledoit and M. Wolf. Honey, I shrunk the sample covariance matrix. *Journal of Portfolio Management*, 31:110–119, 2004.

[7] O. Ledoit and M. Wolf. A Well-Conditioned Estimator for Large-Dimensional Covariance Matrices. *Journal of Multivariate Analysis*, 88: 365–411, 2004.

[8] J. Schäfer and K. Strimmer. 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, 2005.

[9] J. Schäfer. *Small-Sample Analysis and Inference of Networked Dependency Structured from Complex Genomic Data*. PhD thesis, Fakultät für Mathematik, Informatik und Statistik der Ludwig-Maximilians Universität München, 2005.

[10] T. Augustin. Entscheidungstheorie. Vorlesungsskript, 2007.

[11] F. Tai and W. Pan. Incorporating prior knowledge of gene functional groups into regularized discriminant analysis of microarray data. *Bioinformatics*, 23:3170–3177, 2007.

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

[13] B. Efron. Estimating the Error Rate of a Prediction Rule: Improvement on Cross-Validation. *Journal of the American Statistical Association*, 78:316–331, 1983.

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

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

[16] Yixin Wang, Jan G M Klijn, Yi Zhang, Anieta M Sieuwerts, Maxime P Look, Fei Yang, Dmitri Talantov, Mieke Timmermans, Marion E Meijer van Gelder, Jack Yu, Tim Jatkoe, Els M J J Berns, David Atkins, and John A Foekens. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet*, 365:671–679, 2005. doi: 10.1016/S0140-6736(05)17947-1. URL `http://dx.doi.org/10.1016/S0140-6736(05)17947-1`.

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