SHrinkage Covariance Estimation Incorporating Prior Biological Knowledge with Applications to High-Dimensional Data

Guillemot, Vincent
Department of Medical Informatics, Biometry and Epidemiology, University of Munich
Marchioninistr. 15, 81377 München, Germany
E-mail: vincent.guillemot@ibe.med.uni-muenchen.de

Jelizarow, Monika
Department of Medical Informatics, Biometry and Epidemiology, University of Munich
Marchioninistr. 15, 81377 München, Germany
E-mail: monika.jelizarow@ibe.med.uni-muenchen.de

Tenenhaus, Arthur
Supelec, Department of Signal Processing & Electronic Systems
3, rue Joliot-Curie, 91190 Gif-sur-Yvette, France
E-mail: arthur.tenenhaus@supelec.fr

Boulesteix, Anne-Laure
Department of Medical Informatics, Biometry and Epidemiology, University of Munich
Marchioninistr. 15, 81377 München, Germany
E-mail: boulesteix@ibe.med.uni-muenchen.de

1 Introduction

Methods integrating prior knowledge on the structure of covariates into prediction models have become very popular in the last few years in the context of statistical bioinformatics. This knowledge may be given, e.g. as a set of clusters of covariates that are involved in the same biological process or have a similar function, or as a set of pairwise connections between covariates in the form of a graph. The methods integrating prior knowledge into prediction models – or more generally into multivariate statistical methods – implicitly postulate that, say, connected covariates should have a similar regression coefficient, are more correlated than non-connected covariates, or should be selected following a none-or-all principle. These methods are generally denoted as "integration methods" in the rest of this paper, where the term "integration" refers to the integration of prior biological knowledge on the structure of the covariates into multivariate statistical analyses.

While some integration methods primarily aim at providing more interpretable results, others are presented as a means of improving an objective criterion, for example the prediction error. The methods proposed in the statistical bioinformatics literature are as diverse as the backgrounds of their authors, ranging from statistical model-based approaches to machine learning procedures. In this paper, we demonstrate new applications of the covariance estimator $\widehat{\mathbf{\Sigma}}_{\mathrm{SHIP}}$ (standing for SHrinkage covariance estimator Incorporating Prior biological knowledge) [6] to various multivariate methods.

In Section 2 we first outline the theory behind the shrinkage estimator $\widehat{\mathbf{\Sigma}}_{\mathrm{SHIP}}$ and further give a short introduction to the scope of the corresponding `R` package `SHIP`. In section 3 we present a wide range of multivariate statistical methods which can benefit from the incorporation of biological knowledge through the covariance estimator $\widehat{\mathbf{\Sigma}}_{\mathrm{SHIP}}$ and critically discuss its usefulness for the considered problems. In this section we also introduce a new variant of $\widehat{\mathbf{\Sigma}}_{\mathrm{SHIP}}$ specifically designed to address the particularities of one of the applications.

2 The SHIP covariance estimator

2.1 Method

The covariance estimator $\widehat{\mathbf{\Sigma}}_{\text{SHIP}}$ is based on the shrinkage estimator introduced by [8, 9] and applied by [11] in the context of high-dimensional genomic data for which the number of variables p exceeds the sample size n. It addresses the methodological challenges arising from the $n \ll p$ data setting, where the empirical covariance matrix of rank at most $n-1$ and dimensions $p \times p$ is not invertible. In addition, it can incorporate additional assumptions, for instance based on prior biological knowledge on gene functional groups extracted from the database KEGG (Kyoto encyclopedia of genes and genomes [7]). In a few words, the shrinkage estimator [11] 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 parameter, $\mathbf{T}$ stands for a structured covariance target, and $\mathbf{S}$ is the unstructured 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, resulting in a target-specific analytical formula [11]. For statistical details on the computation of λ see [11].

The choice of the covariance target $\mathbf{T}$ is essential in the computation of the shrinkage estimator, but 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 suspected true covariance structure between the variables. An overview of commonly used covariance targets is given in [11]. One of these targets, denoted as target F in [11] (see Table 1), is the starting point for the development of new targets incorporating biological information, e.g. from KEGG. A modified version of target F where pairs of connected variables (i.e. genes from the same gene functional group) have non-zero common correlation ($\bar{r}$) has been suggested for this purpose [6]. The resulting target G defines a new matrix $\mathbf{T}$ whose elements are given on the right of Table 1. The estimator $\widehat{\mathbf{\Sigma}}^*$ obtained by plugging this new matrix $\mathbf{T}$ into the formula $\lambda\mathbf{T} + (1-\lambda)\mathbf{S}$ is denoted as $\widehat{\mathbf{\Sigma}}_{\text{SHIP}}$. Note that a multiplicity of other F-type targets incorporating prior biological knowledge are conceivable [6]. In Section 2.2 we introduce the special cases considered in this paper. For details on the choice of the parameter λ, see Additional File 1 of [6].

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

Table 1: Overview of targets D, F and G (where $\bar{r}$ is the average of the sample correlations between connected variables). The notation $i \sim j$ means that variables i and j are connected. The term s_{ij} denotes the entry of the unbiased estimator of the covariance matrix in row i, column j.

2.2 The R Package SHIP

Different variants of the covariance estimator $\widehat{\mathbf{\Sigma}}_{\text{SHIP}}$ [6] are implemented in the publicly available `R` package `SHIP` [5]. They differ by the user-chosen type of covariance target. The target is the highly structured matrix used to shrink the unbiased empirical covariance matrix, and optionally incorporates prior knowledge from the database KEGG. The function `build.target()` is a wrapper function to build the various targets by specifying the argument `type`. In particular, the settings `type="D"`, `type="F"` and `type="G"` create the targets displayed in Table 1. Several variants of target G are implemented: `type="cor"` is a modified version of target G testing the correlations (with a significance level of 0.05) and setting the non-significant ones to zero before the mean correlation $\bar{r}$ is computed, while `type="Gpos"` completely ignores negative correlations and computes the mean correlation using the positive ones only. Prior knowledge is incorporated through the argument `genegroups`, a list with as many elements as variables (genes) in the dataset. Each element of the list `genegroups` gives the groups (pathways) to which the corresponding variable (gene) belongs. For more details we refer to the package manual [5]. The function `shrink.estim()` then computes the chosen variant of the covariance estimator $\widehat{\mathbf{\Sigma}}_{\text{SHIP}}$. For example, the command `shrink.estim(x,genegroups,build.target(x,type="G"))` yields $\widehat{\mathbf{\Sigma}}_{\text{SHIP}}$ based on target G, where `x` is the $n \times p$ data matrix. Depending on the target, the positive definiteness of the resulting estimate of the covariance matrix may not be ensured, for example with targets G or G*. Therefore, in the following applications we use the function `make.positive.definite` (from library `corpcor`) to make the new estimate positive definite if necessary.

3 Using $\widehat{\mathbf{\Sigma}}_{\text{SHIP}}$ in various multivariate statistical methods

The estimation of the covariance matrix, or of its inverse, is essential in many multivariate analysis methods, and becomes critical when the number of variables p exceeds the number of individuals n. In this part, we show how the estimator $\widehat{\mathbf{\Sigma}}_{\text{SHIP}}$ can be integrated into three multivariate methods as diverse as linear discriminant analysis (LDA), regularized generalized canonical correlation analysis (RGCCA) and global analysis of covariance (GlobalANCOVA). Each of these methods uses in a different way the information contained in the estimated covariance matrix. In the rest of this section, these methods are modified by replacing a standard covariance estimator by the $\widehat{\mathbf{\Sigma}}_{\text{SHIP}}$ estimator.

3.1 Simulation settings

The use of the covariance estimator $\widehat{\mathbf{\Sigma}}_{\text{SHIP}}$ in the three considered methods (LDA, RGCCA, Global-ANCOVA) and the resulting performance is demonstrated using simulated data. All simulations are based on the assumption of a multivariate normal distribution with block-diagonal covariance matrix of size $p \times p$ of the form

$$(1) \quad \mathbf{\Sigma} = \begin{bmatrix} \mathbf{A}_1 & 0 & \dots & 0 \\ 0 & \mathbf{A}_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \mathbf{A}_K \end{bmatrix},$$

where each submatrix $\mathbf{A}_k$ is of size $p_k \times p_k$, with $\sum_{k=1}^{K} p_k = p$. The submatrices $\mathbf{A}_k$ have the form $\mathbf{A}_k = (1 - a_k)\mathbf{I}_{p_k} + a_k\mathbf{J}_{p_k}$ where a_k is a scalar in $]0, 1[$, $\mathbf{I}_{p_k}$ stands for the identity matrix and $\mathbf{J}_{p_k}$ for the matrix of ones, both of them of size $p_k \times p_k$. Each $\mathbf{A}_k$ thus corresponds to a "group" of correlated variables. For the LDA and RGCCA applications, each variable may correspond to a gene, and groups represent gene functional groups that are supposed to be more correlated than genes from different functional groups. Groups may be quite large (typically $p_k > 10$) but the within-group correlation a_k tends to be moderate. In the application to GlobalANCOVA, each variable corresponds to a probe of the microarray, and a group corresponds to different probesets targeting *the same gene*. In contrast to the LDA and RGCCA settings, the groups are very small (usually $1 \leq p_k \leq 5$) and the within-group correlation a_k is typically very high.

3.2 Application to LDA

Linear Discriminant Analysis (LDA) is a widely used classification method based on the assumption that the random vector of explanatory variables follows a multivariate normal distribution $\mathcal{N}(\boldsymbol{\mu}_r, \boldsymbol{\Sigma})$ within each class r (for $r = 1, \ldots, c$, where c denotes the total number of classes). A new observation is assigned to the class with maximal posterior probability. Note that the linearity of the decision function results from the assumption of equal within-class covariance matrices (i.e. $\boldsymbol{\Sigma}_1 = \cdots = \boldsymbol{\Sigma}_c$). This decision function involves the inverse $\boldsymbol{\Sigma}^{-1}$ of the covariance matrix $\boldsymbol{\Sigma}$ which in standard $n > p$ settings is estimated through the inverse $\widetilde{\mathbf{S}}^{-1}$ of the pooled empirical covariance matrix $\widetilde{\mathbf{S}}$. In high-dimensional settings, however, $\widetilde{\mathbf{S}}$ is singular and thus not invertible. Regularized linear discriminant analysis aims at solving the singularity problem by modifying $\widetilde{\mathbf{S}}$ such that the resulting estimator becomes invertible [2, 3]. Here, we estimate $\boldsymbol{\Sigma}^{-1}$ through the inverse of $\widehat{\boldsymbol{\Sigma}}_{\mathrm{SHIP}}$ introduced in Section 2.1 and follow the formulation of multiclass LDA from [1], shrinking the correlations only according to [11]. The resulting classification method is denoted as SHIP-LDA in the rest of this paper.

For this application, the groups of variables represent functional groups of genes sharing a similar function. In our simulations the genes belonging to such "functional groups" are assumed to be correlated ($a_k > 0$), while the correlation between two genes belonging to different groups is set to 0. Figure 1 shows the error rates of SHIP-LDA with target D, target G, and target G with randomly permuted groups of variables.

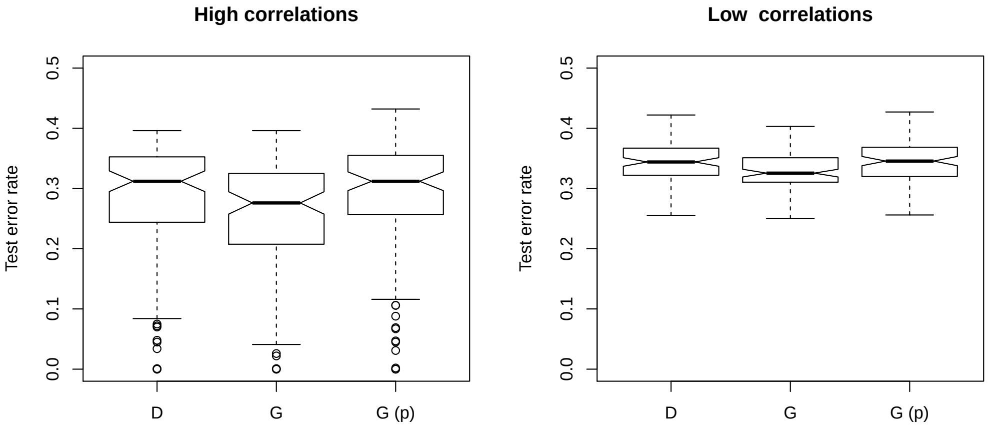

Figure 1: Performance (test error rate) of the modified versions of LDA. Within the two classes ($Y = 0, 1$) of size 25 each ($n = 50$), the data are generated as *i.i.d.* realizations of a multivariate normal distribution with mean 0 in class $Y = 0$ and $\boldsymbol{\mu}$ in class $Y = 1$, where the vector $\boldsymbol{\mu}$ contains independent realizations of the distribution $\mathcal{N}(0, \sigma^2)$. The parameter σ controls the level of separation of the two classes and is set to $\sigma = 0.15$ in this example to obtain a neither too easy nor too difficult classification problem. The variables are split into 3 groups of variables of sizes 200, 500 and 300 with a null correlation between groups and correlation a_k within each group $k = 1, 2, 3$. The boxplots display the test error rates obtained from 100 simulated datasets by applying SHIP-LDA with targets D, G, and G with randomly permuted groups of variables. The test error rates are estimated based on 1000 independent test observations generated from the same distribution. A selection of 30 variables is performed on the training sets based on the p-value of the t-test before any classification method is applied. **LEFT**: the within-group correlations a_k are high ($a_1 = 0.97$, $a_2 = 0.9$ and $a_3 = 0.95$). **RIGHT**: the within-group correlations a_k are low ($a_1 = 0.2$, $a_2 = 0.3$ and $a_3 = 0.1$).

These results show that the performance of linear discriminant analysis is impacted by the estimation of the covariance matrix. When the SHIP estimator with target G (middle boxplot) is used and the correct groups of variables are specified, the performance is slightly improved when compared to the use of target D (left boxplot) or target G with permuted groups of variables (right boxplot). Further, we observe that the difference between the performance of the three SHIP-LDA versions tends to decrease when the number of individuals increases, for instance to $n = 100$ or $n = 400$. This is probably because the estimation of the covariance matrix becomes less critical as the sample size grows, even without prior information on the group structure. On the whole, our simulation shows a positive influence of the incorporation of group information in the form of target G on the test error rate of LDA, thus endorsing the concept of SHIP-LDA with target G. This improvement, however, is very moderate in the investigated settings.

3.3 Application to RGCCA

Regularized Generalized Canonical Correlation Analysis (RGCCA) [13] is a generalization of regularized canonical correlation analysis to three or more blocks of variables. It constitutes a very general framework for studying relationships between several blocks of variables observed on the same set of individuals. Let us denote by $\mathbf{X}^{(1)}, ..., \mathbf{X}^{(B)}$ the B (centered) data matrices corresponding to B blocks of variables, each of them measured on the same n individuals. The objective of RGCCA is to find for each block linear combinations of variables (denoted as *latent components*) such that i) these components explain their own block well and/or ii) components related to blocks that are assumed to be connected are highly correlated. The RGCCA algorithm requires to compute for each block the inverse of the shrinkage covariance matrix $\widehat{\mathbf{\Sigma}}^{*(b)} = \lambda_b \mathbf{I} + (1-\lambda_b)\frac{1}{n}\mathbf{X}^{(b)\top}\mathbf{X}^{(b)}$, for $b = 1, ..., B$, where each shrinkage parameter λ_b is derived from an analytical formula [11]. RGCCA is implemented in the `R` package `RGCCA` [12]. We suggest to plug the covariance estimator $\widehat{\mathbf{\Sigma}}^{(b)}_{\text{SHIP}}$ into the RGCCA algorithm in place of $\widehat{\mathbf{\Sigma}}^{(b)}, b = 1, ..., B$. The performance of the combination of RGCCA with $\widehat{\mathbf{\Sigma}}^{(b)}_{\text{SHIP}}$ is evaluated on simulated data.

In our simulations, we consider $B = 3$ blocks, where the $n \times m^{(b)}$ data matrices $\mathbf{X}^{(b)}$ $(b = 1, 2, 3)$ have the form:

$$\mathbf{X}^{(b)} = \alpha \boldsymbol{\eta}^{(b)} \mathbb{1}^\top_{m^{(b)}} + \mathbf{Z}^{(b)}.$$

In the above formula, the n-vector $\boldsymbol{\eta}^{(b)}$ corresponds to the first latent component of the block b, α is a scalar reflecting the importance of the latent component $\boldsymbol{\eta}^{(b)}$, $\mathbb{1}_{m^{(b)}}$ is the $m^{(b)}$-vector of ones, and the $n \times m^{(b)}$ matrix $\mathbf{Z}^{(b)}$ is an additional term. The rows of $\mathbf{Z}^{(b)}$ are *i.i.d.* realizations of a multivariate normal variable with mean 0 and $m^{(b)} \times m^{(b)}$ covariance matrix of the form of Eq.(1):

$$\mathbf{\Sigma}^{(b)}_{\mathbf{Z}} = \begin{bmatrix} (1-a_1^{(b)})\mathbf{I} + a_1^{(b)}\mathbf{J} & 0 & \dots & 0 \\ 0 & (1-a_2^{(b)})\mathbf{I} + a_2^{(b)}\mathbf{J} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & (1-a_K^{(b)})\mathbf{I} + a_K^{(b)}\mathbf{J} \end{bmatrix},$$

where $a_k^{(b)}$ (for $k = 1, \dots, K$ and $b = 1, \dots, K$) are scalars in $]0, 1[$. The elements of the three vectors $\boldsymbol{\eta}^{(1)}, \boldsymbol{\eta}^{(2)}, \boldsymbol{\eta}^{(3)}$ are *i.i.d.* realizations of a multivariate normal variable with mean 0 and covariance matrix

$$\mathbf{\Sigma}_{\boldsymbol{\eta}} = \begin{bmatrix} 1 & \rho_{1,2} & \rho_{1,3} \\ \rho_{2,1} & 1 & \rho_{2,3} \\ \rho_{3,1} & \rho_{3,2} & 1 \end{bmatrix}.$$

It follows that the n rows of the matrices $\mathbf{X}^{(b)}$ are themselves the realizations of multivariate normal variables structured into groups. The resulting covariance matrix of the whole random vector obtained by concatenating the three blocks can be written as

$$\boldsymbol{\Sigma} = \begin{bmatrix} \mathbf{M}^{(1)} & \rho_{1,2}\alpha^2\mathbf{J} & \rho_{1,3}\alpha^2\mathbf{J} \\ \rho_{2,1}\alpha^2\mathbf{J} & \mathbf{M}^{(2)} & \rho_{2,3}\alpha^2\mathbf{J} \\ \rho_{3,1}\alpha^2\mathbf{J} & \rho_{3,2}\alpha^2\mathbf{J} & \mathbf{M}^{(3)} \end{bmatrix}, \text{ where } \mathbf{M}^{(b)} = \begin{bmatrix} \mathbf{A}_1^{(b)} & \alpha^2 & \dots & \alpha^2 \\ \alpha^2 & \mathbf{A}_2^{(b)} & \dots & \alpha^2 \\ \vdots & \vdots & \ddots & \vdots \\ \alpha^2 & \alpha^2 & \dots & \mathbf{A}_K^{(b)} \end{bmatrix}$$

is the covariance matrix of block b, each block being itself structured into groups of variables with covariance matrix $\mathbf{A}_k^{(b)} = (1 - a_k^{(b)})\mathbf{I} + (\alpha^2 + a_k^{(b)})\mathbf{J}$.

Since $\alpha^2 \neq 0$, the use of target G is not appropriate. To address this issue, we slightly modify target G into a new target, called target H, in order to adapt it to the case where all variables of a block are expected to be correlated even if they are not in a common group (i.e. not connected). The new target is defined in Table 2.

Target H

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

Table 2: The new target H adapted to the application of SHIP to RGCCA. $\bar{r}_C$ (resp. $\bar{r}_{NC}$) is the average of the sample correlations between **C**onnected (resp. **N**on-**C**onnected) variables.

Figure 2 shows the usefulness of the combination of RGCCA with the SHIP estimator (target H or H with randomly permuted groups) in terms of MSE for $n = 50$ and $n = 200$ when $\alpha = 0.1$ and $\alpha = \sqrt{2}$. When considering target H, the estimated shrinkage parameters λ_b are close to 1 for each block, which means that target H is taken into account for the estimation of the covariance matrices. Conversely, for targets D and H(p), the shrinkage parameters are close to 0 for each block.

Furthermore, when the number of individuals is set to $n = 50$ and $\alpha = \sqrt{2}$, the performance with target H is significantly worse than the performance with target H(p), as can been observed from the right-bottom panel of Figure 2. We conjecture that this unexpected result is due to a better conditioning of the covariance matrix obtained from target H with permuted groups.

The peculiar settings of RGCCA led us to design the new target H which takes into account the within and between group correlation separately. In our simulation this new target yields a significant improvement of the MSE both over target G (data not shown) and target D. Indeed, when using target G, the shrinkage parameters are close to 0 for each block, which means that target G is not taken into account for the construction of $\widehat{\boldsymbol{\Sigma}}$. With target G the MSE is worse than with target D, probably because the resulting estimated covariance matrix is then ill-conditioned.

Finally, let us point out that the MSE is not computable in real data applications, where the true latent components and their correlation are unknown. The evaluation of the performance of the combination of RGCCA with the SHIP estimator beyond simulations is thus not straightforward, and it is difficult to evaluate whether or not the integration of group structure information in form of the SHIP estimator could benefit real applications.

3.4 Application to GlobalANCOVA

In the last few years, global testing methods have been proposed as a useful tool for the analysis of high-dimensional genomic data. Single variables are not always the primary focus. For example, one may be more interested in sets of genes from a common pathway rather than in single genes. The GlobalANCOVA approach [4] implemented in the `R` package `GlobalANCOVA` [10] is one the testing methods proposed in the literature to globally test groups of variables. It tests the global null-hypothesis that all variables have the same mean in the considered groups. The estimation of the covariance matrix estimation is necessary to compute asymptotical p-values (as opposed to permutation-based p-values). The current version of GlobalANCOVA uses the shrinkage covariance estimator [11] with target D. We propose to incorporate priori knowledge on the group structure of the variables into the computation of asymptotical p-values by using target G instead of target D.

Our simulation design mimics the realistic case of a group of genes that are represented by several probesets in a microarray. In this setting, variables are probesets and groups are genes, as opposed to the previous examples where variables were genes and groups were pathways. Thus, the groups of variables are now very small (including 2 to 4 variables), because each gene is targeted by a very small number of probesets, but highly correlated ($\rho = 0.8$ to 0.95), because probesets targeting the same gene measure the same quantity. The correlation between genes is considered to be null, following the general covariance structure given in Eq.(1).

The empirical distribution of the p-values – obtained from 1000 simulated data sets under the null-hypothesis – is showed for three different shrinkage estimators of the covariance matrix on Figure 3. Under the null-hypothesis, these p-values are expected to be uniformly distributed. However, we see that the empirical distribution of p-values is noticeably non-uniform when the covariance matrix is estimated with target D or with target G after random permutation of the groups of variables. In contrast, when knowledge on the group structure of the variable is integrated into the procedure through the SHIP estimator, the distribution is approximately uniform under the null-hypothesis – as required from a statistical test.

4 Conclusion

In all three applications of the SHIP estimator – with target G or with the new target H – we showed a quantitative improvement compared to targets ignoring the group structure: in terms of prediction error in LDA, in terms of MSE in RGCCA, and in terms of uniformity of the distribution of p-values under the null-hypothesis in GlobalANCOVA. Based on simulated data with known and strong group structure, we thus demonstrated the advantage of integrating prior knowledge on this structure into the estimation of the covariance matrix for use in various multivariate methods.

This study, however, is intended as a proof of concept, and does not aim to definitely establish the superiority of the suggested SHIP-based variants of LDA, RGCCA and GlobalANCOVA. Firstly, more simulations in different settings would be needed for each of these three methods to obtain more general results. Secondly, the SHIP covariance estimator is sometimes ill-conditioned, depending on the group structure and of the strength of the correlations. This problem could be addressed in future research, e.g. by adding an additional diagonal matrix with its own shrinkage factor. Thirdly, the group structures considered in our simulations are intentionally more simplistic than in real data settings. Fourthly, our approach is limited to situations where variables within a group have higher correlations than variables from different groups. This idea might seem natural from the point of view of a statistician, but in real life not all group structures can be translated in terms of higher correlations. For example, genes from common KEGG pathways do not necessarily have higher correlations than genes from different pathways [6]. And even if such groups exist, they may be (partially) unknown to the biomedical experts.

In conclusion, we consider the integration of information on the group structure into the shrinkage-based estimation of the covariance matrix as promising, but believe that caution and careful consideration of the substantive context are necessary in practice.

REFERENCES (RÉFERENCES)

[1] M. Ahdesmäki and K. Strimmer. Feature selection in omics prediction problems using cat scores and false non-discovery rate control. Annals of Applied Statistics, 4:503–509, 2010.

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

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

[4] M. Hummel, R. Meister, and U. Mansmann. GlobalANCOVA: exploration and assessment of gene group effects. Bioinformatics, 24(1):78–85, 2008.

[5] M. Jelizarow and V. Guillemot. SHIP (SHrinkage covariance Incorporating Prior knowledge). R package, version 1.0.1, 2010.

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

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

[8] 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.

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

[10] U. Mansmann, R. Meister, M. Hummel, R. Scheufele, and with contributions from S. Knueppel. GlobalAncova: Calculates a global test for differential gene expression between groups, 2010. R package version 3.18.0.

[11] J. Schäfer and K. Strimmer. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. Stat Appl Genet Mol Biol, 4:Issue 1, Article 32, 2005.

[12] A. Tenenhaus. RGCCA: Regularized Generalized Canonical Correlation Analysis, 2010. R package version 1.0.

[13] A. Tenenhaus and M. Tenenhaus. Regularized generalized canonical correlation analysis. Psychometrika, 76(2):257–284, 2011.

RÉSUMÉ (ABSTRACT)

In "-omic data" analysis, information on the structure of covariates are broadly available either from public databases describing gene regulation processes and functional groups such as the Kyoto encyclopedia of genes and genomes (KEGG), or from statistical analyses – for example in form of partial correlation estimators. The analysis of transcriptomic data might benefit from the incorporation of such prior knowledge. In this paper we focus on the integration of structured information into statistical analyses in which at least one major step involves the estimation of a (high-dimensional) covariance matrix. More precisely, we revisit the recently proposed "SHrinkage Incorporating Prior" (SHIP) covariance estimation method which takes into account the group structure of the covariates, and suggest to integrate the SHIP covariance estimator into various multivariate methods such as linear discriminant analysis (LDA), global analysis of covariance (GlobalANCOVA), and regularized generalized canonical correlation analysis (RGCCA). We demonstrate the use of the resulting new methods based on simulations and discuss the benefit of the integration of prior information through the SHIP estimator. Reproducible R codes are available at `http://www.ibe.med.uni-muenchen.de/organisation/mitarbeiter/020_professuren/boulesteix/shipproject/index.html`.

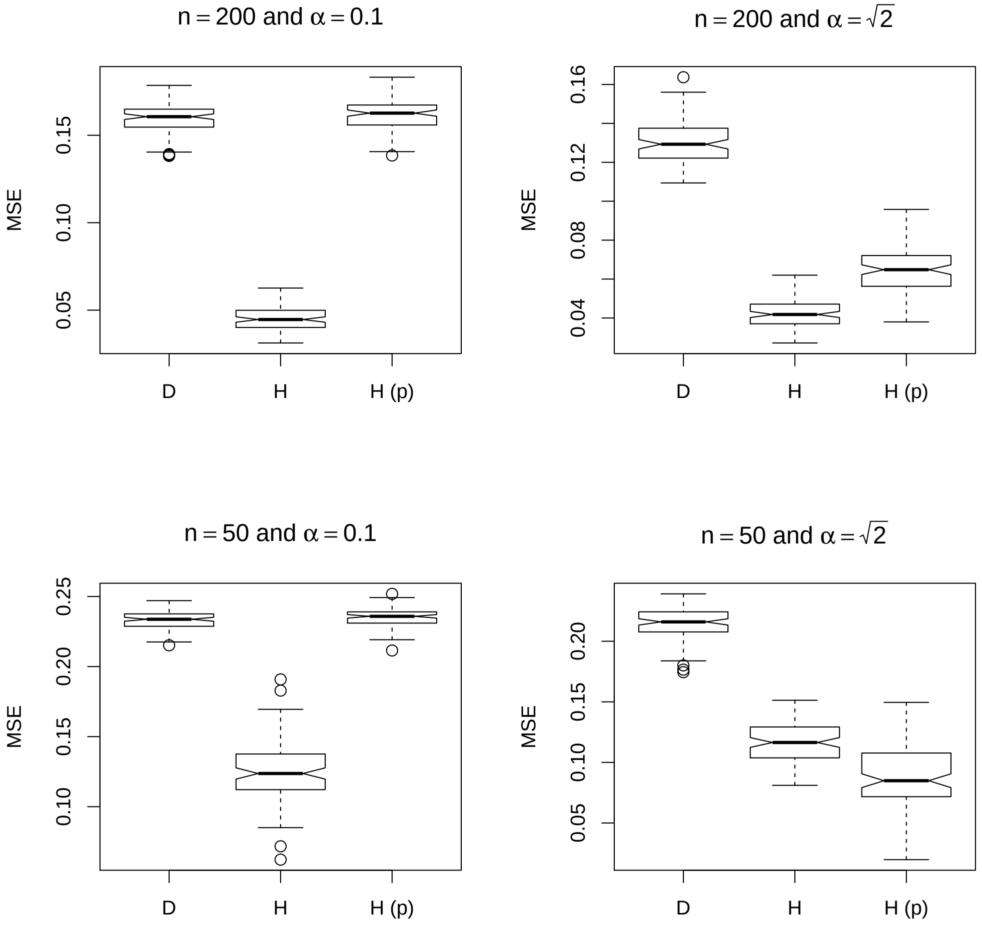

Figure 2: Mean Square Error of the estimated correlation matrix $\widehat{\mathbf{\Sigma}}_{\boldsymbol{\eta}}$ between the components. To obtain boxplots, we generated 100 independent datasets. Three blocks are simulated, each one containing respectively 100, 200 and 50 variables. The correlations of the latent components are set to $\rho_{1,2} = \rho_{1,3} = 0.7$ and $\rho_{2,3} = 0$. The first block contains 5 groups of 20 variables (with correlations $a_1^{(1)} = 0.8$, $a_2^{(1)} = 0.9$, $a_3^{(1)} = 0.7$, $a_4^{(1)} = 0.85$ and $a_5^{(1)} = 0.96$), the second block contains 2 groups of 100 variables (with correlations $a_1^{(2)} = 0.8$ and $a_2^{(2)} = 0.9$) and the last block contains 3 groups of respectively 20, 20 and 10 variables (with correlations and $a_1^{(3)} = 0.8$, $a_2^{(3)} = 0.9$ and $a_3^{(3)} = 0.85$). In the **UPPER FIGURES** the sample size is set to $n = 200$, whereas in the **LOWER FIGURES** $n = 50$. For the **LEFT FIGURES**, α is set to $\alpha = 0.1$, which means that the group-structured component $\mathbf{Z}^{(b)}$ dominates the latent component $\boldsymbol{\eta}^{(b)}$ of the block. For the **RIGHT FIGURES**, $\alpha = \sqrt{2}$, which means that the latent component $\boldsymbol{\eta}^{(b)}$ dominates the group-structured component $\mathbf{Z}^{(b)}$.

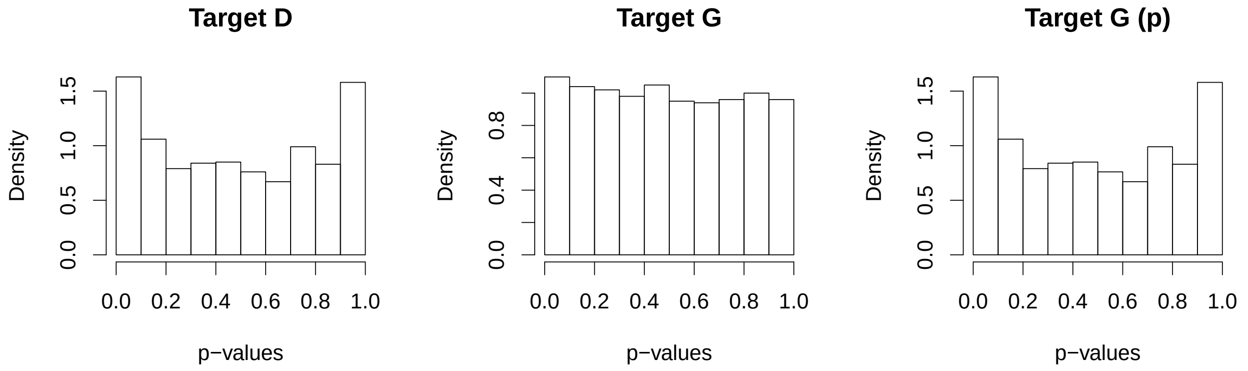

Figure 3: Histograms of the p-values under the null hypothesis. The displayed p-values are obtained from 1000 datasets with $p = 200$ variables (probesets) and $n = 100$ observations. The $p = 200$ variables are divided into 34 groups of 2, 30 groups of 3 and 18 groups of 4. Each of these groups are assigned a correlation (a_k) chosen randomly from the set $\{0.8, 0.85, 0.9, 0.95\}$. **LEFT FIGURE**: Target D is used for the estimation of the covariance matrix. **MIDDLE FIGURE**: Target G is used. **RIGHT FIGURE**: Target G is used with permuted groups.