Statistical Applications in Genetics and Molecular Biology

Volume 4, *Issue* 1 2005 *Article* 32

A Shrinkage Approach to Large-Scale Covariance Matrix Estimation and Implications for Functional Genomics

Juliane Schäfer* Korbinian Strimmer†

*Department of Statistics, University of Munich, Germany, juliane.schaefer@stat.uni-muenchen.de

†Department of Statistics, University of Munich, Germany, korbinian.strimmer@lmu.de

A Shrinkage Approach to Large-Scale Covariance Matrix Estimation and Implications for Functional Genomics*

Juliane Schäfer and Korbinian Strimmer

Abstract

Inferring large-scale covariance matrices from sparse genomic data is an ubiquitous problem in bioinformatics. Clearly, the widely used standard covariance and correlation estimators are ill-suited for this purpose. As statistically efficient and computationally fast alternative we propose a novel shrinkage covariance estimator that exploits the Ledoit-Wolf (2003) lemma for analytic calculation of the optimal shrinkage intensity.

Subsequently, we apply this improved covariance estimator (which has guaranteed minimum mean squared error, is well-conditioned, and is always positive definite even for small sample sizes) to the problem of inferring large-scale gene association networks. We show that it performs very favorably compared to competing approaches both in simulations as well as in application to real expression data.

KEYWORDS: Shrinkage, covariance estimation, "small n, large p" problem, graphical Gaussian model (GGM), genetic network, gene expression.

---

*We thank W. Schmidt-Heck, R. Guthke and K. Bayer for kindly providing us with the E. coli data and for discussion. This work was supported by the Deutsche Forschungsgemeinschaft (DFG) via an Emmy-Noether research grant to K.S.

1. Introduction

Estimation of large-scale covariance matrices is a common though often implicit task in functional genomics and transcriptome analysis:

- For instance, consider the clustering of genes using data from a microarray experiment (e.g. Eisen et al., 1998). In order to construct a hierarchical tree describing the functional grouping of genes an estimate of the similarities between all pairs of expression profiles is needed. This is typically based on a distance measure related to the empirical correlation. Thus, if p genes are being analyzed (with p perhaps in the order of 1,000 to 10,000), a covariance matrix of size $p \times p$ has to be calculated.

- Another example is the construction of so-called relevance networks (Butte et al., 2000). These visually represent the marginal (in)dependence structure among the p genes. The networks are built by drawing edges between those pairs of genes whose absolute pairwise correlation coefficients exceed a pre-specified threshold (say, 0.8).

- Related to gene relevance networks (though conceptually quite different) are gene association networks. These are graphical models that have recently been suggested as a means of displaying the conditional dependencies among the considered genes (e.g., Toh and Horimoto, 2002; Dobra et al., 2004; Schäfer and Strimmer, 2005a). An essential input to inferring such a network is the $p \times p$ covariance matrix.

- Furthermore, the covariance matrix evidently plays an important role in the classification of genes.

- In addition, there are numerous bioinformatics algorithms that rely on the pairwise correlation coefficient as part of an (often rather adhoc) optimality score.

Thus, a common key problem in all of these examples is as follows: How should one obtain an accurate and reliable estimate of the population covariance matrix $\mathbf{\Sigma}$ if presented with a data set that describes a large number of variables but only contains comparatively few samples ($n \ll p$)?

In the vast majority of analysis problems in bioinformatics (specifically excluding classification) the simple solution is to rely either on the maximum likelihood estimate $\boldsymbol{S}^{\mathrm{ML}}$ or on the related unbiased empirical covariance matrix $\boldsymbol{S} = \frac{n}{n-1}\boldsymbol{S}^{\mathrm{ML}}$,

with entries defined as

$$s_{ij} = \frac{1}{n-1} \sum_{k=1}^{n} (x_{ki} - \bar{x_i})(x_{kj} - \bar{x_j}), \tag{1}$$

where $\bar{x_i} = \frac{1}{n} \sum_{k=1}^{n} x_{ki}$ and x_{ki} is the k-th observation of the variable X_i. However, unfortunately both $\boldsymbol{S}$ and $\boldsymbol{S}^{\mathrm{ML}}$ exhibit serious defects in the "small n, large p" data setting commonly encountered in functional genomics. Specifically, in this case the empirical covariance matrix can *not* anymore be considered a good approximation of the true covariance matrix (this is true also for moderately sized data with $n \approx p$).

For illustration consider Fig. 1 where the sample covariance estimator $\boldsymbol{S}$ is compared with an alternative estimator $\boldsymbol{S}^{\star}$ developed in Section 2 of this paper and summarized in Tab. 1. This figure shows the sorted eigenvalues of the estimated matrices in comparison with the true eigenvalues for fixed $p = 100$ and various ratios $\frac{p}{n}$. It is evident by inspection that for small n the eigenvalues of $\boldsymbol{S}$ differ greatly from the true eigenvalues of $\boldsymbol{\Sigma}$. In addition, for n smaller than p (bottom row in Fig. 1) $\boldsymbol{S}$ looses its full rank as a growing number of eigenvalues become zero. This has several undesirable consequences. First, $\boldsymbol{S}$ is not positive definite any more, and second, it can not be inverted as it becomes singular. For comparison contrast the poor performance of $\boldsymbol{S}$ with that of $\boldsymbol{S}^{\star}$ (fat green line in Fig. 1). This improved estimator exhibits none of the defects of $\boldsymbol{S}$, in particular it is more accurate, well conditioned, and always positive definite. Nevertheless, $\boldsymbol{S}^{\star}$ can be computed in only about twice the time required to calculate $\boldsymbol{S}$.

With this paper we pursue three aims. First, we argue against the blind use of the empirical covariance matrix $\boldsymbol{S}$ in data situations where it is not appropriate – noting that this affects many current application areas in bioinformatics. Second, we describe a route to obtain improved estimates of the covariance matrix via shrinkage combined with analytic determination of the shrinkage intensity according to the Ledoit-Wolf theorem (Ledoit and Wolf, 2003). Third, we show that this new regularized estimator greatly enhances inferences of gene association networks.

The remainder of the paper is organized as follows. In the next section we provide an overview over shrinkage, the Ledoit-Wolf lemma and its application to shrinkage of covariance matrices. We discuss several potentially useful lower-dimensional targets (cf. Tab. 2), with special focus on the "diagonal, unequal variance" model. In the second part of the paper, we review methodology for inferring large-scale genetic networks (relevance and association networks). We conduct computer simulations to show that using $\boldsymbol{S}^{\star}$ in genetic network model selection is highly advantageous in terms of power and other performance criteria. Finally, we illustrate the described approach by analyzing a real gene expression data set from *E. coli*.

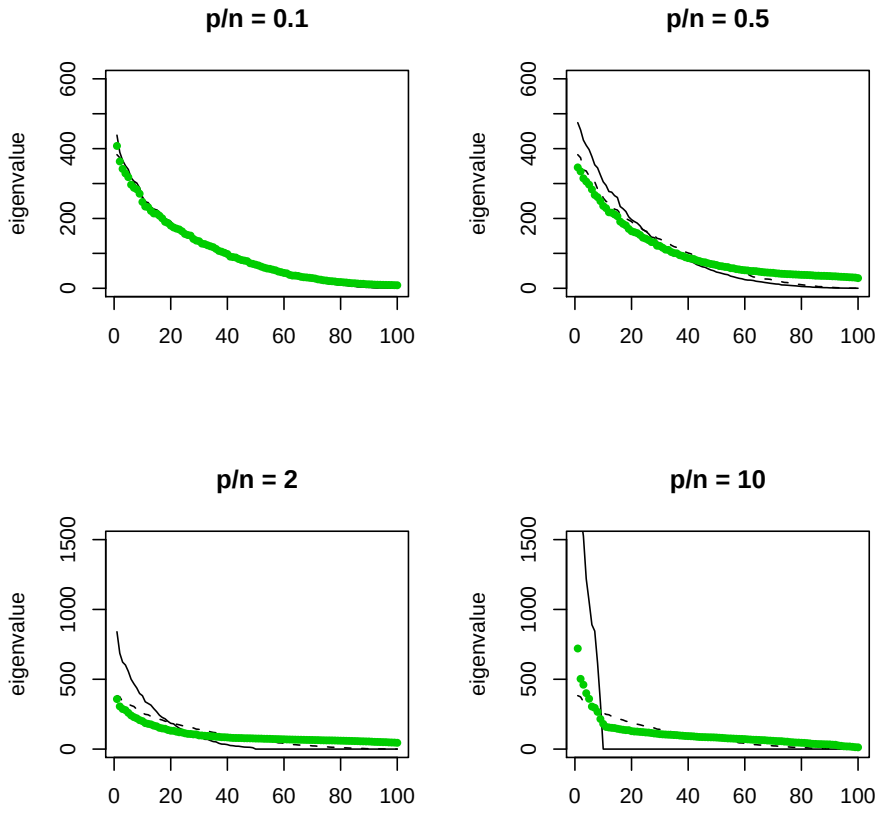


Figure 1: Ordered eigenvalues of the sample covariance matrix $\boldsymbol{S}$ (thin black line) and that of an alternative estimator $\boldsymbol{S}^{\star}$ (fat green line, for definition see Tab. 1), calculated from simulated data with underlying p-variate normal distribution, for $p = 100$ and various ratios p/n. The true eigenvalues are indicated by a thin black dashed line.

"Small *n*, Large *p*" Covariance and Correlation Estimators $S^\star$ and $R^\star$:

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

and

$$r^\star_{ij} = \begin{cases} 1 & \text{if } i = j \\ r_{ij}\min(1, \max(0, 1 - \hat{\lambda}^\star)) & \text{if } i \neq j \end{cases}$$

with

$$\hat{\lambda}^\star = \frac{\sum_{i \neq j} \widehat{\mathrm{Var}}(r_{ij})}{\sum_{i \neq j} r^2_{ij}}$$

Table 1: Small sample shrinkage estimators of the unrestricted covariance and correlation matrix suggested in this paper (Section 2.4). The coefficients s_{ii} and r_{ij} denote the empirical variance (unbiased) and correlation, respectively. For details of the computation of $\widehat{\mathrm{Var}}(r_{ij})$ see Appendix A. Further variants of these estimators are discussed in Section 2.4.

2. Shrinkage estimation of covariance matrices in a "small n, large p" setting

2.1. Strategies for obtaining a more efficient covariance estimator

It has long been known that the two widely-employed estimators of the covariance matrix, i.e. the unbiased ($\boldsymbol{S}$) and the related maximum likelihood ($\boldsymbol{S}^{\mathrm{ML}}$) estimator, are both statistically inefficient. In a nutshell, this can be explained as a consequence of the so-called "Stein phenomenon" discovered by Stein (1956) in the context of estimating the mean vector of a multivariate normal distribution. Stein demonstrated that in high-dimensional inference problems it is often possible to improve (sometimes dramatically!) upon the maximum likelihood estimator. This result is at first counterintuitive, as maximum likelihood can be proven to be *asymptotically* optimal, and as such it seems not unreasonable to expect that these favorable properties of maximum likelihood also extend to the case of finite data. However, further insight into the Stein effect is provided by Efron (1982) who points out that one needs to distinguish between two different aspects of maximum likelihood inference. First, maximum likelihood as a means of summarizing the observed data and producing a *maximum likelihood summary* (MLS). Second, maximum likelihood as a procedure to obtain a *maximum likelihood estimate* (MLE). The conclusion is that maximum likelihood is unassailable as a data summarizer but that it has some clear defects as an estimating procedure.

This applies directly to the estimation of covariance matrices: $\boldsymbol{S}^{\mathrm{ML}}$ constitutes the best estimator in terms of actual fit to the data but for medium to small data sizes it is far from being the optimal estimator for recovering the population covariance – as is well illustrated by Fig. 1. Fortunately, the Stein theorem also demonstrates that it is possible to construct a procedure for improved covariance estimation. In addition to increased efficiency and accuracy, it is desirable for such a method to exhibit the following characteristics *not* found in $\boldsymbol{S}$ and $\boldsymbol{S}^{\mathrm{ML}}$:

1. The estimate should always be positive definite, i.e. all eigenvalues should be distinct from zero.

2. The estimated covariance matrix should be well-conditioned.

The positive-definiteness requirement is an intrinsic property of the true covariance matrix that is satisfied as long as the considered random variables have non-zero variance. If a matrix is well-conditioned, i.e. if the ratio of its maximum and minimum singular value is not too large, it has full-rank and can be easily inverted.

Thus, by producing a well-conditioned covariance estimate one automatically also obtains an equally well-conditioned estimate of the *inverse covariance* – a quantity of crucial importance, e.g., in interval estimation, classification, and graphical models.

A (naive) strategy to obtain a positive definite estimator of the covariance runs as follows: Take the sample covariance $\boldsymbol{S}$ and apply, e.g., the algorithm by Higham (1988). This will adjust all eigenvalues to be larger than some prespecified threshold ϵ and thus guarantee positive definiteness. However, the resulting matrix will not be well conditioned.

Another more general procedure to obtain an improved covariance estimator is variance reduction. Consider the well-known bias-variance decomposition of the mean squared error (MSE) for the sample covariance, i.e.

$$\mathrm{MSE}(\boldsymbol{S}) = \mathrm{Bias}(\boldsymbol{S})^2 + \mathrm{Var}(\boldsymbol{S}). \tag{2}$$

As $\mathrm{Bias}(\boldsymbol{S}) = 0$ by construction, the only way to decrease the overall accuracy of $\boldsymbol{S}$ is by reducing its variance. A simple non-parametric approach to variance reduction is offered, e.g., by bootstrap aggregation ("bagging") of the empirical covariance matrix. This can be done by explicitly approximating the expectation $E(\boldsymbol{S})$ via the bootstrap. In previous work (Schäfer and Strimmer, 2005a) we have resorted to this strategy to produce improved estimates of the correlation matrix and its inverse. However, especially for the very large dimensions commonly encountered in genomics problems (often with $p > 1,000$) this approach is computationally by far too demanding.

Instead, in this paper we investigate "shrinking" or more general "biased estimation" (e.g., Hoerl and Kennard, 1970a,b; Efron, 1975; Efron and Morris, 1975, 1977) as a means of variance reduction of $\boldsymbol{S}$. In particular, we consider a recent analytic result from Ledoit and Wolf (2003) that allows to construct an improved covariance estimator that is not only suitable for small sample size n and large numbers of variables p but at the same time is also completely inexpensive to compute.

2.2. Shrinkage estimation and the lemma of Ledoit-Wolf

In this section we briefly review the general principles behind shrinkage estimation and discuss an analytic approach by Ledoit and Wolf (2003) for determining the optimal shrinkage level. We note that the theory outlined here is not restricted to covariance estimation but applies generally to large-dimensional estimation problems.

Let $\boldsymbol{\Psi} = (\psi_1, \ldots, \psi_p)$ denote the parameters of the unrestricted high-dimensional model of interest, and $\boldsymbol{\Theta} = (\theta_i)$ the matching parameters of a lower dimensional

restricted submodel. For instance, $\mathbf{\Psi}$ could be the mean vector of a p-dimensional multivariate normal, and $\mathbf{\Theta}$ the vector of a corresponding constrained submodel where the means are all assumed to be equal (i.e. $\theta_1 = \theta_2 = \cdots = \theta_p$). By fitting each of the two different models to the observed data associated estimates $\boldsymbol{U} = \hat{\mathbf{\Psi}}$ and $\boldsymbol{T} = \hat{\mathbf{\Theta}}$ are obtained. Clearly, the unconstrained estimate $\boldsymbol{U}$ will exhibit a comparatively high variance due to the larger number of parameters that need to be fitted, whereas its low-dimensional counterpart $\boldsymbol{T}$ will have lower variance but potentially also considerable bias as an estimator of the true $\mathbf{\Psi}$.

Instead of choosing between one of these two extremes, the linear shrinkage approach suggests to *combine* both estimators in a weighted average

$$\boldsymbol{U}^{\star} = \lambda \boldsymbol{T} + (1 - \lambda)\boldsymbol{U}, \tag{3}$$

where $\lambda \in [0, 1]$ denotes the shrinkage intensity. Note that for $\lambda = 1$ the shrinkage estimate equals the shrinkage target $\boldsymbol{T}$ whereas for $\lambda = 0$ the unrestricted estimate $\boldsymbol{U}$ is recovered. The key advantage of this construction is that it offers a systematic way to obtain a regularized estimate $\boldsymbol{U}^{\star}$ that outperforms the individual estimators $\boldsymbol{U}$ and $\boldsymbol{T}$ both in terms of accuracy and and statistical efficiency.

A key question in this procedure is how to select an optimal value for the shrinkage parameter. In some instances, it may suffice to fix the intensity λ at some given value, or to make it depend on the sample size according to a simple function. Often more appropriate, however, is to choose the parameter λ in a data-driven fashion by explicitly minimizing a risk function

$$R(\lambda) = E(L(\lambda)) = E(\sum_{i=1}^{p}(u_i^{\star} - \psi_i)^2), \tag{4}$$

here for example the mean squared error (MSE).

One common but also computationally very intensive approach to estimate the minimizing λ is by using cross-validation - for an example see Friedman (1989) where shrinkage is applied in the context of regularized classification. Another widely applied route to inferring λ views the shrinkage problem in an empirical Bayes context. In this case the quantity $E(\boldsymbol{T})$ is interpreted as prior mean and λ as a hyper-parameter that may be estimated from the data by optimizing the marginal likelihood (e.g., Morris, 1983; Greenland, 2000).

It is less well known that the optimal regularization parameter λ may often also be determined *analytically*. Specifically, Ledoit and Wolf (2003) recently derived a simple theorem for choosing λ that guarantees minimal MSE without the need of having to specify any underlying distributions, and without requiring computationally expensive procedures such as MCMC, the bootstrap, or cross-validation.

This lemma is obtained in a straightforward fashion. Assuming that the first two moments of the distributions of $\boldsymbol{U}$ and $\boldsymbol{T}$ exist, the squared error loss risk function from Eq. 4 may be expanded as follows:

$$\begin{aligned} R(\lambda) &= \sum_{i=1}^{p} \mathrm{Var}\,(u_i^\star) + \left[E(u_i^\star) - \psi_i\right]^2 \\ &= \sum_{i=1}^{p} \mathrm{Var}\,(\lambda t_i + (1-\lambda)u_i) + \left[E(\lambda t_i + (1-\lambda)u_i) - \psi_i\right]^2 \\ &= \sum_{i=1}^{p} \lambda^2 \,\mathrm{Var}(t_i) + (1-\lambda)^2 \,\mathrm{Var}(u_i) + 2\lambda(1-\lambda)\,\mathrm{Cov}(u_i, t_i) \\ &\quad + \left[\lambda E(t_i - u_i) + \mathrm{Bias}(u_i)\right]^2 . \end{aligned} \tag{5}$$

Analytically minimizing this function with respect to λ gives, after some tedious algebraic calculations, the following expression for the optimal value

$$\lambda^\star = \frac{\sum_{i=1}^{p} \mathrm{Var}(u_i) - \mathrm{Cov}(t_i, u_i) - \mathrm{Bias}(u_i)\,E(t_i - u_i)}{\sum_{i=1}^{p} E[(t_i - u_i)^2]}, \tag{6}$$

for which minimum MSE $R(\lambda^\star)$ is achieved. It can be shown that $\lambda^\star$ always exists and that it is unique. If $\boldsymbol{U}$ is an *unbiased* estimator of $\boldsymbol{\Psi}$ with $E(\boldsymbol{U}) = \boldsymbol{\Psi}$ this equation reduces to

$$\lambda^\star = \frac{\sum_{i=1}^{p} \mathrm{Var}(u_i) - \mathrm{Cov}(t_i, u_i)}{\sum_{i=1}^{p} E[(t_i - u_i)^2]}, \tag{7}$$

which is – apart from some further algebraic simplification – the expression given in Ledoit and Wolf (2003).

Closer inspection of Eq. 6 yields a number of insights into how the optimal shrinkage intensity is chosen:

- First, the smaller the variance of the high-dimensional estimate $\boldsymbol{U}$, the smaller becomes $\lambda^\star$. Therefore, with increasing sample size the influence of the target $\boldsymbol{T}$ diminishes.

- Second, $\lambda^\star$ also depends on the correlation between estimation error of $\boldsymbol{U}$ and of $\boldsymbol{T}$. If both are positively correlated then the weight put on the shrinkage target decreases. Hence, the inclusion of the second term in the numerator of Eq. 6 adjusts for the fact that the two estimators $\boldsymbol{U}$ and $\boldsymbol{T}$ are both inferred from the same data set. It also takes into account that the "prior" information associated with $\boldsymbol{T}$ is not independent of the given data.

- Third, with increasing mean squared difference between $\boldsymbol{U}$ and $\boldsymbol{T}$ (in the denominator of Eq. 6) the weight $\lambda^\star$ also decreases. Note that this automatically protects the shrinkage estimate $\boldsymbol{U}^\star$ against a misspecified target $\boldsymbol{T}$.

- Fourth, if the unconstrained estimator is biased, and the bias points already towards the target, the shrinkage intensity is correspondingly reduced.

Furthermore, it is noteworthy that variables that by design are kept identical in the constrained and unconstrained estimators (i.e. $t_i = u_i$ for some i) play no role in determining the intensity $\lambda^\star$, as their contributions to the various terms in Eq. 6 cancel out.

This can be generalized further by allowing *multiple targets*, each with its own different optimal shrinkage intensity. This is especially appropriate if there exists a natural grouping of parameters in the investigated high-dimensional model. In this case one simply computes the individual targets and applies Eq. 6 to each group of variables separately. As one referee suggests, it may be helpful to cluster variables according to their variances $\mathrm{Var}(u_i)$ – typically the predominant term to determine the shrinkage level $\lambda^\star$.

Finally, it is important to consider the transformation properties of the shrinkage procedure. From Eq. 6 it is clear that $\lambda^\star$ is invariant against translations. For instance, the underlying data may be centered without affecting the estimation of the optimal shrinkage intensity. However, $\lambda^\star$ is *not* generally invariant against scale transformations. This dependence on the absolute scales of the considered variables is a general property that shrinkage shares with other approaches to biased estimation, such as ridge regression and partial least squares (e.g. Hastie et al., 2001).

2.3. Estimation of the optimal shrinkage intensity

For practical application of Eq. 6 one needs to obtain an estimate $\hat{\lambda}^\star$ of the optimal shrinkage intensity. In their paper Ledoit and Wolf (2003) emphasize that the parameters of Eq. 6 should be estimated consistently. However, this is only a very weak requirement, as consistency is an asymptotic property and a basic requirement of any sensible estimator. Furthermore, we are interested in small sample inference. Thus, instead we suggest to compute $\hat{\lambda}^\star$ by replacing all expectations, variances, and covariances in Eq. 6 by their *unbiased* sample counterparts. This leads to

$$\hat{\lambda}^\star = \frac{\sum_{i=1}^{p} \widehat{\mathrm{Var}}(u_i) - \widehat{\mathrm{Cov}}(t_i, u_i) - \widehat{\mathrm{Bias}}(u_i)\,(t_i - u_i)}{\sum_{i=1}^{p} (t_i - u_i)^2}. \qquad (8)$$

In finite samples $\hat{\lambda}^\star$ may exceed the value one, and in some cases it may even become negative. In order to avoid overshrinkage or negative shrinkage we truncate

the estimated intensity correspondingly, using $\hat{\lambda}^{\star\star} = \max(0, \min(1, \hat{\lambda}^{\star}))$ when constructing the shrinkage estimator of Eq. 3.

It is also worth noting that Eq. 8 is valid regardless of the sample size n at hand. In particular, n may be substantially smaller than p, a fact we will exploit in our suggested approach to inferring gene association networks.

2.4. Shrinkage estimation of the covariance matrix

Estimation of the unrestricted covariance matrix requires the determination of $(p^2 + p)/2$ free parameters, and thus constitutes a high-dimensional inference problem. Consequently, application of shrinkage offers a promising approach to obtain improved estimates.

Daniels and Kass (2001) provide a fairly extensive review of empirical Bayes shrinkage estimators proposed in recent years. Unfortunately, most of the suggested estimators appear to suffer from at least one of the following drawbacks, which renders them unsuitable for the analysis of genomic data:

1. Typically, the application is restricted to data with $p < n$, in order to ensure that the empirical covariance $\boldsymbol{S}$ can be inverted. However, most current genomic data sets contain vastly more features than samples ($p \gg n$).

2. Many of the suggested estimators are computationally expensive (e.g. those based on MCMC sampling), or assume specific underlying distributions.

These difficulties are elegantly avoided by resorting to the (almost) distribution-free Ledoit-Wolf approach to shrinkage.

In a matrix setting the equivalent to the squared error loss function is the squared Frobenius norm. Thus,

$$\begin{aligned} L(\lambda) &= \|\boldsymbol{S}^{\star} - \boldsymbol{\Sigma}\|_{\mathrm{F}}^2 \\ &= \|\lambda \boldsymbol{T} + (1-\lambda)\boldsymbol{S} - \boldsymbol{\Sigma}\|_{\mathrm{F}}^2 \\ &= \sum_{i=1}^{p}\sum_{j=1}^{p}\left(\lambda t_{ij} + (1-\lambda)s_{ij} - \sigma_{ij}\right)^2 \end{aligned} \tag{9}$$

is a natural quadratic measure of distance between the true ($\boldsymbol{\Sigma}$) and inferred covariance matrix ($\boldsymbol{S}^{\star}$). In this formula the unconstrained unbiased empirical covariance matrix $\boldsymbol{S}$ replaces the unconstrained estimate $\boldsymbol{U}$ of Eq. 3.

Selecting a suitable estimated covariance target $\boldsymbol{T} = (t_{ij})$ requires some diligence. In general, the choice of a target should be guided by the presumed lower-dimensional structure in the data set as this determines the increase of efficiency

Target A: "diagonal, unit variance"
0 estimated parameters

$$t_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$

$$\hat{\lambda}^{\star} = \frac{\sum_{i \neq j} \widehat{\text{Var}}(s_{ij}) + \sum_i \widehat{\text{Var}}(s_{ii})}{\sum_{i \neq j} s_{ij}^2 + \sum_i (s_{ii} - 1)^2}$$

Target B: "diagonal, common variance"
1 estimated parameter: v

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

$$\hat{\lambda}^{\star} = \frac{\sum_{i \neq j} \widehat{\text{Var}}(s_{ij}) + \sum_i \widehat{\text{Var}}(s_{ii})}{\sum_{i \neq j} s_{ij}^2 + \sum_i (s_{ii} - v)^2}$$

Target C: "common (co)variance"
2 estimated parameters: v, c

$$t_{ij} = \begin{cases} v = \text{avg}(s_{ii}) & \text{if } i = j \\ c = \text{avg}(s_{ij}) & \text{if } i \neq j \end{cases}$$

$$\hat{\lambda}^{\star} = \frac{\sum_{i \neq j} \widehat{\text{Var}}(s_{ij}) + \sum_i \widehat{\text{Var}}(s_{ii})}{\sum_{i \neq j} (s_{ij} - c)^2 + \sum_i (s_{ii} - v)^2}$$

Target D: "diagonal, unequal variance"
p estimated parameters: s_{ii}

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

$$\hat{\lambda}^{\star} = \frac{\sum_{i \neq j} \widehat{\text{Var}}(s_{ij})}{\sum_{i \neq j} s_{ij}^2}$$

Target E: "perfect positive correlation"
p estimated parameters: s_{ii}

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

$$f_{ij} = \frac{1}{2} \{ \sqrt{\frac{s_{jj}}{s_{ii}}} \widehat{\text{Cov}}(s_{ii}, s_{ij}) + \sqrt{\frac{s_{ii}}{s_{jj}}} \widehat{\text{Cov}}(s_{jj}, s_{ij}) \}$$

$$\hat{\lambda}^{\star} = \frac{\sum_{i \neq j} \widehat{\text{Var}}(s_{ij}) - f_{ij}}{\sum_{i \neq j} (s_{ij} - \sqrt{s_{ii} s_{jj}})^2}$$

Target F: "constant correlation"
$p + 1$ estimated parameters: s_{ii}, $\bar{r}$

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

$$\hat{\lambda}^{\star} = \frac{\sum_{i \neq j} \widehat{\text{Var}}(s_{ij}) - \bar{r} f_{ij}}{\sum_{i \neq j} (s_{ij} - \bar{r} \sqrt{s_{ii} s_{jj}})^2}$$

Table 2: Six commonly used shrinkage targets for the covariance matrix and associated estimators of the optimal shrinkage intensity – see main text for discussion. *Abbreviations:* v, average of sample variances; c, average of sample covariances; $\bar{r}$, average of sample correlations.

over that of the empirical covariance. However, it is also a remarkable consequence of Eq. 6 that in fact *any* target will lead to a reduction in MSE, albeit only a minor one in case of a strongly misspecified target (then $\boldsymbol{S}^\star$ will simply reduce to the unconstrained estimate $\boldsymbol{S}$).

Six commonly used covariance targets are compiled in Tab. 2, along with a brief description, the dimension of the target, and the resulting estimate $\hat{\lambda}^\star$. In order to compute the optimal shrinkage intensity it is necessary to estimate the variances of the individual entries of $\boldsymbol{S}$ – see Appendix A for the technical details. Note that the resulting shrinkage estimators $\boldsymbol{S}^\star$ all exhibit the same order of algorithmic complexity as the standard estimate $\boldsymbol{S}$.

Probably the most commonly employed shrinking targets are the identity matrix and its scalar multiple. These are denoted in Tab. 2 "diagonal, unit variance" (target A) and "diagonal, common variance" (target B). A further extension is provided by the two parameter covariance model that in addition to the common variance (as in target B) also maintains a common covariance ("common (co)variance", target C). The three targets share several properties. First, they are all extremely low-dimensional (0 to 2 free parameters), thus they impose a rather strong structure which in turn requires only little data to fit. Second, the resulting estimators *shrink all components* of the empirical covariance matrix, i.e. both diagonal and off-diagonal entries. In the literature it is easy to find examples where one of the above targets is employed – albeit *not* in combination with analytic estimation of the shrinkage level. For instance, the unit diagonal target A is typically used in ridge regression and the related Tikhonov regularization (e.g. Hastie et al., 2001). The target B is utilized, e.g., by Friedman (1989) who estimates λ by means of cross-validation, by Leung and Chan (1998) who use a fixed $\lambda = \frac{2}{n+2}$, by Dobra et al. (2004) as a parameter in an inverse Wishart prior for the covariance matrix, and finally also by Ledoit and Wolf (2004b). The two-parameter target C appears not to be widely used.

Another class of covariance targets is given by the "diagonal, unequal variance" model (target D), the "perfect positive correlation" model (target E) and the "constant correlation" model (target F) of Tab. 2. A shared feature of these three targets is that they are comparatively parameter-rich, and that they only lead to *shrinkage of the off-diagonal elements* of $\boldsymbol{S}$. The last two shrinkage targets were introduced with the purpose of modeling stock returns. These tend – on average – to be strongly positively correlated (Ledoit and Wolf, 2003, 2004a).

In this paper, we focus on the shrinkage target D for the estimation of covariance and correlation matrices arising in genomics problems. This "diagonal, unequal variance" model represents a compromise between the low-dimensional targets A, B, and C and the correlation models E and F. Like the simpler targets A and B it shrinks the off-diagonal entries to zero. However, unlike shrinkage targets A

and B, target D leaves diagonal entries intact, i.e. it does *not* shrink the variances. Thus, this model assumes that the parameters of the covariance matrix fall into two classes, and both are treated differently in the shrinkage process.

This clear separation also suggests that for shrinking purposes it may be useful to parameterize the covariance matrix in terms of variances and correlations (rather than variances and covariances) so that $s^\star_{ij} = r^\star_{ij} \sqrt{s_{ii} s_{jj}}$. In this formulation, shrinkage is applied to the correlations rather than the covariances. This has two distinct advantages. First, the off-diagonal elements determining the shrinkage intensity are all on the same scale. Second, the (partial) correlations derived from the resulting covariance estimator $\boldsymbol{S}^\star$ are independent of scale and location transformations of the underlying data matrix, just as is the case for those computed from $\boldsymbol{S}$.

It is this form of target D that we propose for estimating correlation and covariance matrices. For reference, the corresponding formulae are collected in Tab. 1. Note the remarkably simple expression for the shrinkage intensity

$$\hat{\lambda}^\star = \frac{\sum_{i \neq j} \widehat{\mathrm{Var}}(r_{ij})}{\sum_{i \neq j} r_{ij}^2} \tag{10}$$

– see also Tab. 2 (Target D). For technical details such as the calculation of $\widehat{\mathrm{Var}}(r_{ij})$ we refer to Appendix A. In this formula a concern may be the use of the empirical correlation coefficients r_{ij} – after all, these are the ones that we aim to improve. Thus, it seems we face a circularity problem, namely that for an accurate estimate of the shrinkage intensity reliable estimates of correlation are needed, and vice versa. However, it is a remarkable feature of target D that it completely resolves this "chicken-egg" issue: regardless whether standard or shrinkage estimates of correlation are substituted into Eq. 10 the resulting $\hat{\lambda}^\star$ remains all the same.

Using the target D has another important advantage: the resulting shrinkage covariance estimate will automatically be positive definite. The target D itself is always positive definite, and the convex combination of a positive definite matrix ($\boldsymbol{T}$) with another matrix that is positive semidefinite ($\boldsymbol{S}$) always yields a positive definite matrix. Note that this is also true for targets A and B but *not* for the targets C, E, and F (consider as counterexample the target E with all variances set equal to one).

Further variants of the proposed estimator (Tab. 1) are easily constructed. One possible extension is to shrink the diagonal elements as well, using a different intensity for variances and correlations. Shrinking the variances to a common mean is standard practice in in genomic case-control studies (e.g. Cui et al., 2005). It is particular helpful if there are so few samples that the gene-specific variances are difficult to obtain. In such as case, however, it may make no sense at all to consider estimating the full covariance matrix.

3. Inference of gene networks from small sample genomic data

3.1. Methodological background

We consider here two simple approaches for modeling net-like dependency structures in genome expression data, both of which require as input an estimated large-scale covariance matrix. The first and conceptually simpler model is that of a "gene relevance network". This was introduced by Butte et al. (2000) and is built in the following simple fashion. First, the $p \times p$ correlation matrix $\boldsymbol{P} = (\rho_{ij})$ is inferred from the data. Second, for estimated correlation coefficients exceeding a prespecified threshold (say $r > 0.8$) an edge is drawn between the two respective genes. Thus, relevance networks represent the *marginal* (in)dependence structure among the p genes. In statistical terminology this type of network model is also known as "covariance graph".

Despite the popularity of relevance networks (which stems from the relative ease of construction) there are many problems connected with their proper interpretation. For instance, the cut-off value that determines the "significant" edges is typically chosen in a rather arbitrary fashion – often simply a large value is selected with the vague aim to exclude "spurious" edges. However, this misses the statistical interpretation of the marginal correlation which takes account of both direct as well as indirect associations. As a direct consequence, in a reasonably well-connected genetic network most genes will by construction be correlated with each other (e.g. see the analysis of the *E. coli* data below). Thus, in this case even a large observed degree of correlation will provide only weak evidence for the direct dependency of any two considered genes. Instead, the *absence of correlation* (i.e. $r \approx 0$) will be a *strong measure of their independence*. Therefore, even ignoring the difficulties of obtaining accurate measures of correlation from small sample data, gene relevance networks are suitable tools *not* for elucidating the dependence network among genes but rather for uncovering independence.

In contrast, with the class of graphical Gaussian models (GGMs), also called "covariance selection" or "concentration graph" models, a simple statistical approach exists that allows to detect direct dependence between genes. This "gene association network" approach is based on investigating the estimated *partial* correlations $\tilde{r}$ for all pairs of considered genes. The traditionally developed theory of GGMs (e.g. Whittaker, 1990) is only applicable for $n \gg p$. However, with the increasing interest in "small n, large p" inference a number of refinements to the GGM theory have recently been proposed that allow its application also to genomic data – see Schäfer and Strimmer (2005a,b) for a discussion and a comprehensive list of

references. In essence, in a small sample setting both the estimation of the partial correlations as well as the subsequent model selection need to be suitably modified. In the following we discuss several such approaches, including one based on the suggested covariance shrinkage estimator.

3.2. Small sample GGM selection using false discovery rate multiple testing

Standard graphical modeling theory (e.g. Whittaker, 1990) shows that the matrix of partial correlations $\tilde{\boldsymbol{P}} = (\tilde{\rho}_{ij})$ is related to the inverse of the covariance matrix $\boldsymbol{\Sigma}$. This relationship leads to the straightforward estimator

$$\tilde{r}_{ij} = \hat{\tilde{\rho}}_{ij} = -\hat{\omega}_{ij} / \sqrt{\hat{\omega}_{ii}\hat{\omega}_{jj}}, \tag{11}$$

where

$$\hat{\boldsymbol{\Omega}} = (\hat{\omega}_{ij}) = \hat{\boldsymbol{\Sigma}}^{-1}. \tag{12}$$

We note that in the last equation, it is absolutely crucial that the covariance is estimated accurately, and that $\hat{\boldsymbol{\Sigma}}$ is well-conditioned – otherwise the above formulae will result in a rather poor estimate of partial correlation (cf. Schäfer and Strimmer, 2005a). *Here, we adopt the shrinkage estimator $\boldsymbol{S}^{\star}$ developed in the first part of this paper (Tab. 1).* As we show below this leads to (sometimes dramatic) improvement in accuracy over alternative procedures. In this context it is also interesting to note that the difficulty of obtaining reliable estimates of $\tilde{\boldsymbol{P}}$ has led some researchers to instead consider partial correlations of limited order (e.g., de la Fuente et al., 2004; Wille et al., 2004; Magwene and Kim, 2004). However, using partial correlations of first or second order as a measure of dependence amounts to employing a network model that is much more similar to relevance than to association networks, and hence also inherits their interpretation difficulties.

The second critical part of inferring GGMs is model selection. In Schäfer and Strimmer (2005a) we have suggested a simple yet quite effective search heuristic based on large-scale multiple testing of edges. This approach is based on two rationales. First, it exploits the fact that genetic networks are typically sparse, i.e. that most of the $p(p-1)/2$ partial correlation coefficients $\tilde{\rho}$ vanish. In turn, this allows to estimate the null distribution from the data, and thus to decide which edges are present or absent. Second, GGM search by multiple testing implicitly assumes that for all cliques (i.e. fully connected subset of nodes) of size three and more the underlying joint distribution is well approximated by the product of the bivariate densities associated with the respective undirected edges (Cox and Reid, 2004).

Specifically, in the approach of Schäfer and Strimmer (2005a) the distribution of

the observed partial correlations $\tilde{r}$ *across edges* is taken as the mixture

$$f(\tilde{r}) = \eta_0 f_0(\tilde{r};\kappa) + (1-\eta_0) f_{\mathrm{A}}(\tilde{r}), \tag{13}$$

where f_0 is the null distribution, η_0 is the (unknown) proportion of "null edges", and f_{A} the distribution of observed partial correlations assigned to actually existing edges. The null density f_0 is given in Hotelling (1953) as

$$\begin{aligned} f_0(\tilde{r};\kappa) &= (1-\tilde{r}^2)^{(\kappa-3)/2} \frac{\Gamma(\frac{\kappa}{2})}{\pi^{\frac{1}{2}}\Gamma(\frac{\kappa-1}{2})} \\ &= |\tilde{r}|\,\mathrm{Be}(\tilde{r}^2; \frac{1}{2}, \frac{\kappa-1}{2}), \end{aligned} \tag{14}$$

where $\mathrm{Be}(x;a,b)$ is the Beta distribution and κ is the degree of freedom, equal to the reciprocal variance of the null $\tilde{r}$. Fitting this mixture density allows κ, η_0 and even f_{A} to be determined – for an algorithm to infer the latter see Efron (2004, 2005b). Subsequently, it is straightforward to compute the edge-specific "local false discovery rate" (fdr) via

$$\mathrm{Prob}(\text{null edge}|\tilde{r}) = \mathrm{fdr}(\tilde{r}) = \frac{\hat{\eta}_0 f_0(\tilde{r};\hat{\kappa})}{\hat{f}(\tilde{r})}, \tag{15}$$

i.e. the posterior probability that an edge is null given $\tilde{r}$. Finally, an edge is considered "present" or "significant" if its local fdr is smaller than 0.2 (Efron, 2005b).

Closely related to the empirical Bayes local "fdr" statistic is the frequentist "Fdr" (also called q-value) approach advocated by Storey (2002), and the Benjamini and Hochberg (1995) "FDR" rule. In our original GGM model selection proposal (Schäfer and Strimmer, 2005a) we have relied on the FDR method to identify edges in the network. However, we now suggest to employ the local fdr, as this fits more naturally with the mixture model setup, and because it takes account of the dependencies among the estimated partial correlation coefficients (Efron, 2005a).

3.3. Small sample GGM selection using lasso regression

Partial correlations may not only be estimated by inversion of the covariance or correlation matrix (Eq. 11 and Eq. 12). An alternative route is offered by regressing each gene $i \in \{1,\ldots,p\}$ in turn against the remaining set of $p-1$ variables. The partial correlations are then simply

$$\tilde{r}_{ij} = \mathrm{sign}(\hat{\beta}_i^{(j)}) \sqrt{\hat{\beta}_i^{(j)} \hat{\beta}_j^{(i)}}, \tag{16}$$

where $\hat{\beta}_j^{(i)}$ denotes the estimated regression coefficient of predictor variable X_j for the response X_i. Note that while in general $\hat{\beta}_i^{(j)} \neq \hat{\beta}_j^{(i)}$ the signs of these two regression coefficients are identical.

This opens the way for obtaining small sample estimates of partial correlation and GGM inference by means of regularized regression. This avenue is pursued, e.g., by Dobra et al. (2004) who employ Bayesian variable selection. Another possibility to determine the regression coefficients is by penalized regression, for instance ridge regression (Hoerl and Kennard, 1970a,b) or the the lasso (Tibshirani, 1996). The latter approach has the distinct advantage that it will set many of the regression coefficients (and hence also partial correlations) exactly equal to zero. Thus, for covariance selection no additional testing is required, and an edge is recovered in the GGM network if both $\hat{\beta}_i^{(j)}$ and $\hat{\beta}_j^{(i)}$ differ from zero.

GGM inference using the lasso is investigated in Meinshausen and Bühlmann (2005) where it is suggested to choose the lasso penalty λ_i for regression against variable X_i according to

$$\hat{\lambda}_i = 2\sqrt{\frac{s_{ii}^{\mathrm{ML}}}{n}}\,\boldsymbol{\Phi}^{-1}\left(1 - \frac{\alpha}{2p^2}\right), \tag{17}$$

where $\boldsymbol{\Phi}(z)$ is the cumulative distribution function of the standard normal, α is a constant (set to 0.05 in our computations below) that controls the probability of falsely connecting two distinct connectivity components (Meinshausen and Bühlmann, 2005), and s_{ii}^{ML} is the maximum-likelihood estimate of the variance of X_i. Note that this adaptive choice of penalty ensures that for small sample variance $\hat{\lambda}_i$ vanishes and hence in this case no penalization takes place.

3.4. Performance for synthetic data

In an extensive simulation study we compared the shrinkage and lasso approach to GGM selection in terms of accuracy, power, and positive predictive accuracy. In addition to those two methods we also investigated two further estimators of partial correlation denoted $\hat{\mathbf{\Pi}}^1$ and $\hat{\mathbf{\Pi}}^2$. These are discussed in Schäfer and Strimmer (2005a). $\hat{\mathbf{\Pi}}^1$ employs the pseudoinverse instead of the matrix inverse in Eq. 12, thus for $n > p$ it reduces to the classical estimate of partial correlation. $\hat{\mathbf{\Pi}}^2$ uses the bootstrap to obtain a variance-reduced positive definite estimate of the correlation matrix. Note that in our previous study we found that $\hat{\mathbf{\Pi}}^2$ exhibited the overall best performance.

Specifically, the simulation setup was as follows:

1. We controlled parameters of interest such as the number of features p, the

fraction of non-zero edges $\eta_A = 1 - \eta_0$ and the sample size n of the simulated data. Specifically, we fixed at $p = 100$ and $\eta_A = 0.04$, and varied $n = 10, 20, \ldots, 200$.

2. We generated $R = 200$ random networks (i.e. partial correlation matrices) and simulated data of size n from the corresponding multivariate normal distribution.

3. From each of the R data sets we estimated the partial correlation coefficients with the four methods "shrinkage", "lasso", $\hat{\mathbf{\Pi}}^1$, and $\hat{\mathbf{\Pi}}^2$. The number of bootstrap replications required for $\hat{\mathbf{\Pi}}^2$ was set to $B = 500$.

4. Subsequently, we computed the mean squared error by comparison with the known true values.

5. Similarly, we dtermined the average number of edges detected as significant, the power, and the "positive predictive value" (PPV), i.e. the fraction of correct edges among all significant edges. The PPV is sometimes also called the "true discovery rate" (TDR). Note that it is only defined if there is at least one significant edge. The fdr cut-off was set to 0.2 as suggested in Efron (2005b).

In order to simulate random "true" partial correlation matrices we relied on an algorithm producing diagonally dominant matrices – see Schäfer and Strimmer (2005a) for details. This method allows to generate positive definite random correlation matrices of arbitrary size $p \times p$ with an a priori fixed proportion η_A of non-null entries. Unfortunately, further structural and distributional properties are not easily specified – see for instance Hirschberger et al. (2004). This would be desirable as the present simulation algorithm produces networks with edges that represent mostly weak links. Note that this renders their inference disproportionally hard.

In Fig. 2 we compare the accuracy of the four investigated estimators of partial correlation. Both the shrinkage and the lasso GGM estimator outperform the two others regardless of sample size. The previously recommended estimator $\hat{\mathbf{\Pi}}^2$ is nearly as accurate for small sample size, however, it is much more computer expensive than the shrinkage estimator. The peak at $n = 100$ associated with the estimator $\hat{\mathbf{\Pi}}^1$ is a dimension resonance effect due to the use of the pseudoinverse (recall that $p = 100$) – see Schäfer and Strimmer (2005a) for a discussion and references.

Fig. 3a and Fig. 3b summarize the results with regard to GGM selection. Fig. 3a shows the number of edges that were detected as significant using each of the four methods. For $\eta_A = 0.04$ and $p = 100$ there exist exactly 198 edges in any of the simulated networks. The number of edges detected as significant for the shrinkage estimator remains well below this threshold, however in comparison with $\hat{\mathbf{\Pi}}^1$ and

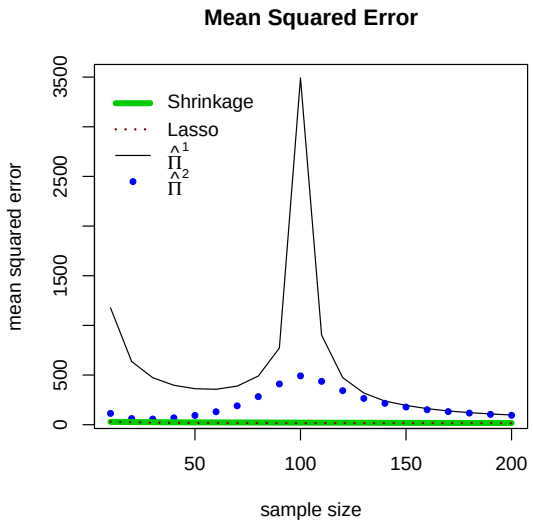
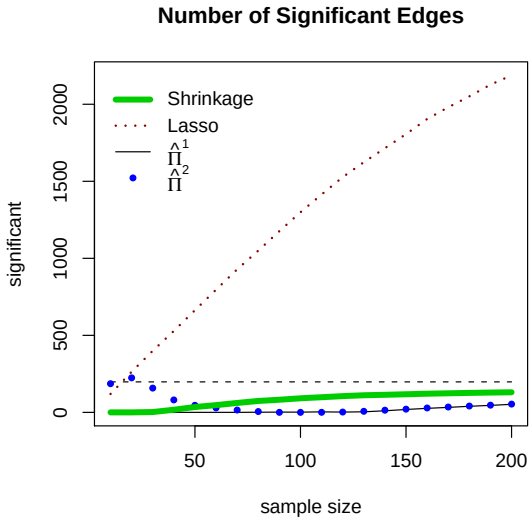


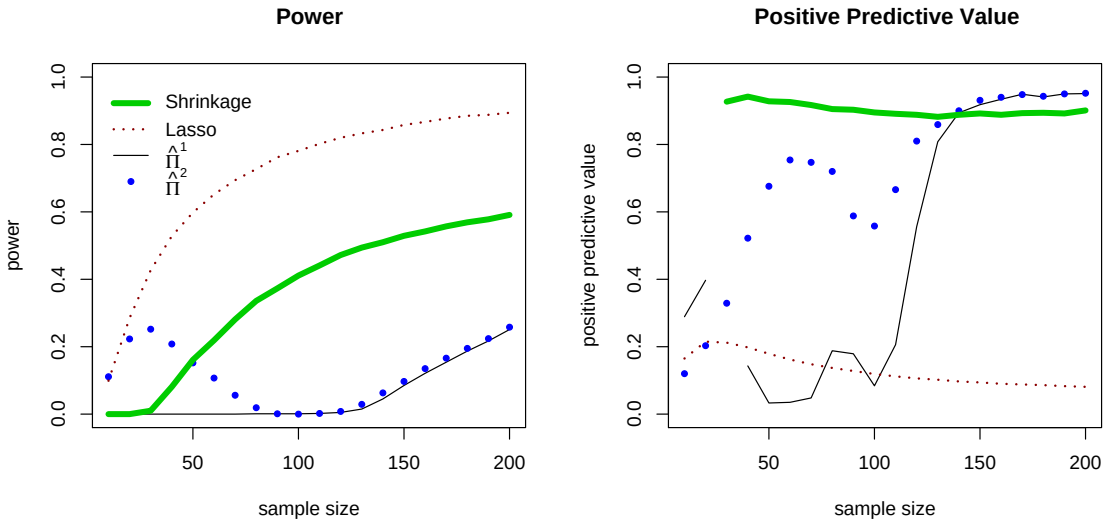
Figure 2: Mean squared error of the four investigated small-sample estimators of partial correlation ("shrinkage", "lasso", $\hat{\mathbf{\Pi}}^1$, and $\hat{\mathbf{\Pi}}^2$) in dependence of the sample size for $p = 100$ genes. Note that the curves for "shrinkage" and "lasso" completely overlap.

$\hat{\mathbf{\Pi}}^2$ it typically finds the largest number of edges. In contrast, for simulated data the lasso GGM network approach recovers even for small sample size many more edges than are actually present. This indicates that the choice of penalization according to Eq. 17 may still be too permissive. The large number of significant edges for $\hat{\mathbf{\Pi}}^2$ for very small sample sizes is a systematic bias related to the improper fit of the null model (Eq. 14).

Fig. 3b illustrates the corresponding power (i.e. the proportion of correctly identified edges) and PPV. The latter quantity is of key practical importance as it is an estimate of the proportion of true edges among the list of edges returned as significant by the algorithm. For the shrinkage estimator the PPV is constant across the whole range of samples sizes and close to the desired level near $1 - \text{Fdr} \approx 0.9$ (Efron, 2005b). The lasso GGM estimator exhibits a very low PPV of about 0.2 only. The other two estimators reach the appropriate level of PPV, but only for $n > p$. In terms of power the shrinkage and the lasso GGM approach outperform the other two investigated estimators $\hat{\mathbf{\Pi}}^1$ and $\hat{\mathbf{\Pi}}^2$ which exhibit reasonable power only for $n > p$. The power of the lasso regression approach is distinctly higher than that of the shrinkage estimator. However, this is due to the fact that the former liberally includes many edges in the resulting network without controlling the rate of false positives. In our simulations the shrinkage estimator has non-zero power only



(a)



(b)

Figure 3: Performance of methods for GGM network inference: (a) Average number of edges detected as significant. Note that there are 198 true edges in the simulated network (horizontal dashed line). (b) Power and positive predictive value (PPV) for reconstructing the GGM network topology. Gaps in the curves for the PPV indicate situations in which the PPV could not be computed (no significant edges).

from $n \geq 30$ (for $p = 100$). As discussed above this is very likely a consequence of our simulation setup which produces partial correlation networks that are hard to infer. Thus, it is deciding to note the high PPV of this estimator: this indicates that if there is a significant edge then the probability is very high that it actually corresponds to a true edge.

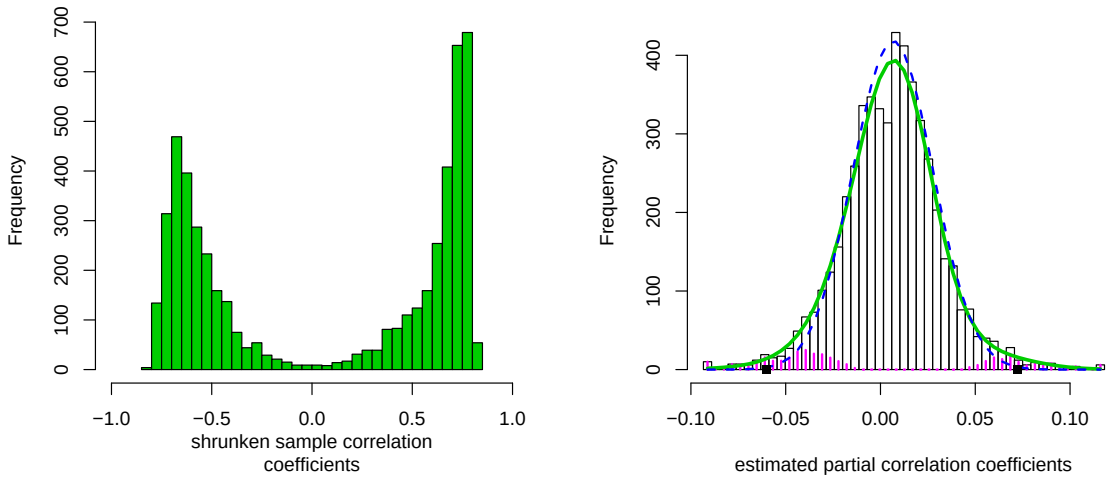
3.5. Analysis of expression profiles from an *E. coli* experiment

For illustration we now apply the above methods for inferring gene networks to a real data set from a microarray experiment conducted at the Institute of Applied Microbiology, University of Agricultural Sciences of Vienna (Schmidt-Heck et al., 2004). This was set up to measure the stress response of the microorganism *Escherichia coli* during expression of a recombinant protein. The resulting data monitors all 4,289 protein coding genes of *E. coli* 8, 15, 22, 45, 68, 90, 150, and 180 minutes after induction of the recombinant protein SOD (human superoxide dismutase). In a comparison with pooled samples before induction 102 genes were identified by Schmidt-Heck et al. (2004) as differentially expressed in one or more samples after induction. In the following we try to establish the gene network among these 102 preselected genes.

A first impression of the dependency structure can be obtained by investigating the estimated correlation coefficients. For the shrinkage approach we obtain $\hat{\lambda}^{\star} = 0.18$. The resulting correlation matrix has full rank (102) with condition number equal to 386.6. In contrast, the standard correlation matrix has rank 8 only and is ill-conditioned (infinite condition number). Thus, already for calculating the correlation coefficients the benefits of using the shrinkage estimator are apparent.

Fig. 4a shows the distribution of the estimated correlation coefficients, most of which *differ* from zero. This is indicates that essentially all genes are are either directly or indirectly associated with each other. Thus, constructing a traditional relevance network (Butte et al., 2000) will – at least for this data – *not* lead to uncovering of the dependency structure. This is compared with the corresponding *partial* correlation matrix. Fig. 4b shows the distribution of the Fisher-transformed coefficients (cf. Hotelling, 1953). The contrast with the previous figure is apparent, as the distribution of partial correlations is unimodal and centered around zero. This means that most partial correlations vanish, that the number of direct interactions is small, and hence that the resulting gene association network is sparse.

Fig. 5 shows the corresponding gene association and relevance networks. The shrinkage GGM network is depicted in Fig. 5a and was derived by fitting the mixture distribution defined in Eq. 13 to the estimated partial correlations with a cut-off

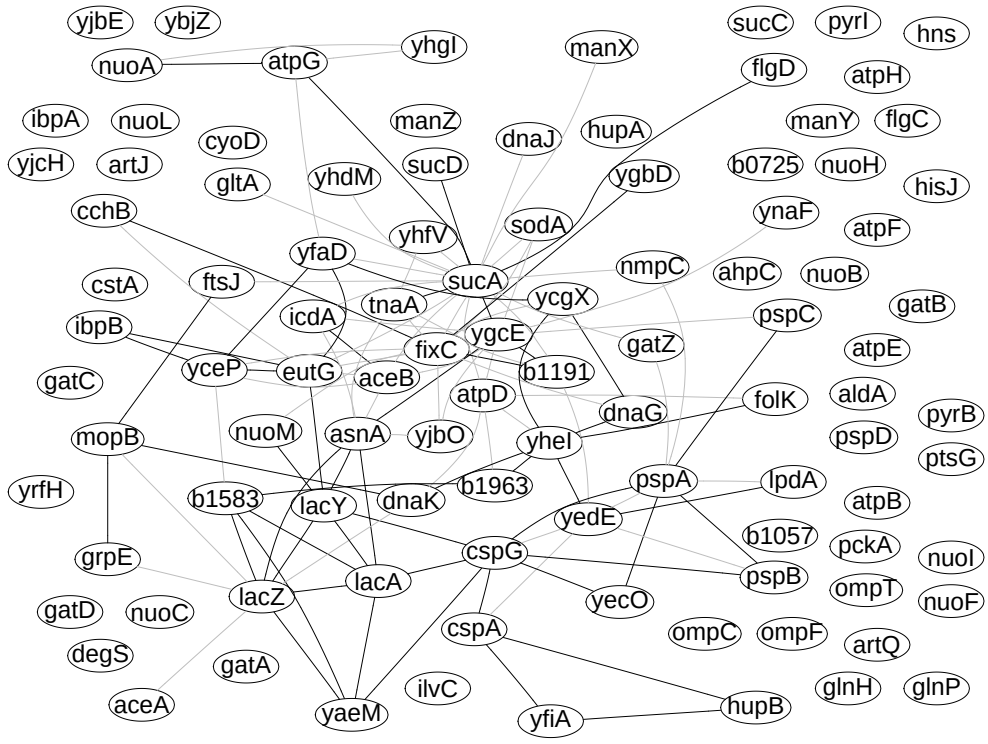


(a) (b)

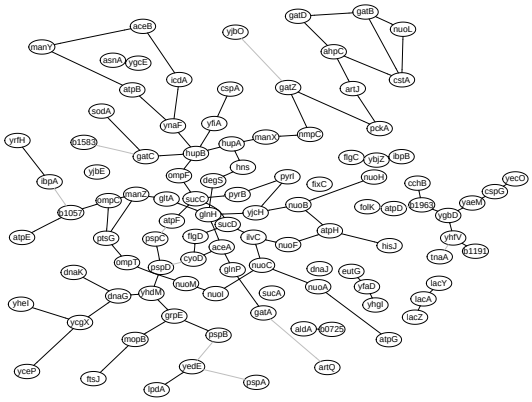
Figure 4: (a) Histogram of the estimated shrunken correlation coefficients computed for all $102 \times 101/2 = 5,151$ pairs of genes. (b) Distribution of estimated *partial* correlation coefficients (green line) after Fisher's normalizing z-transformation (atanh) was applied for normalization purposes. Also shown are the fitted null distribution (dashed blue line) and the alternative distribution (pink) as inferred by the `locfdr` algorithm (Efron, 2004, 2005b). The black squares indicate the 0.2 local fdr cut-off values for the partial correlations.

fdr < 0.2. The network comprises 116 significant edges which amount to about 2% of the 5,151 possible edges for 102 genes. This shows that for real data – in sharp contrast to our comparable simulations – the shrinkage estimator is powerful for small sample size.

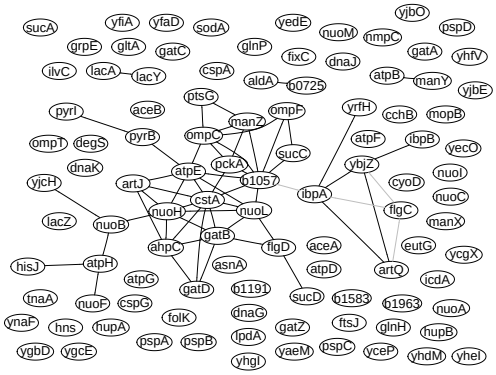
Several aspects of the inferred network are interesting. First, we recover the "hub" connectivity structure for the gene sucA. This gene is involved in the citric acid cycle. The existence of these hubs is a well-known property of biomolecular networks (e.g. Barabási and Oltvai, 2004). It is a strength of the present method that these nodes can be identified without any particular additional effort. Second, the edges connecting the genes lacA, lacZ and lacY are the strongest in the network, with the largest absolute values of partial correlation, and correspondingly also with the smallest local fdr values. Interestingly, these are exactly the genes on which the experiment was based: lacA, lacY and lacZ are induced by IPTG (isopropyl-beta-D-thiogalactopyranoside) dosage and initiate recombinant protein synthesis (Schmidt-Heck et al., 2004).



(a) Shrinkage GGM network



(b) Lasso GGM network



(c) Relevance network

Figure 5: Gene networks inferred from the *E. coli* data by (a) the shrinkage GGM approach presented in this paper (Tab. 1), (b) the lasso GGM approach by Meinshausen and Bühlmann (2005), and (c) the relevance network with abs(r) > 0.8. Black and grey edges indicate positive and negative (partial) correlation, respectively.

For comparison, the lasso GGM network is shown in Fig. 5b. It was computed from the standardized *E. coli* data and contains 100 edges. Closer inspection of this network reveals an interesting structural bias introduced by the lasso regression for GGM inference. As can clearly be seen in Fig. 5b the lasso limits the number of edges going in and out of a node. The reason for this is that the lasso imposes sparsity on the regression coefficients *per node* so that in each regression only a few non-zero coefficients exist. As a consequence, the degree distribution of the *E. coli* lasso GGM network has an implicit upper bound. Thus, the lasso prevents the identification of hubs and also excludes power-law-type connectivity patterns. Note that in contrast in the shrinkage GGM approach sparsity is imposed on the network level rather than locally at node level.

Finally, Fig. 5c shows the relevance network obtained by applying the conventional 0.8 cut-off on the absolute values of the shrunken correlation coefficients. The resulting network contains 58 edges and bears no resemblance to the GGM networks. As is clear from inspecting Fig. 4a there are many more genes that are strongly correlated, so from this network the direct dependencies among genes cannot be deduced. Instead, we argue here that correlations should rather be employed for detecting *independence* among genes. The corresponding null hypothesis is that the two gene are dependent. Fur this purpose the mixture model of Eq. 13 is still applicable, except that the roles of f_0 and f_A are interchanged. Thus any edge with fdr > 0.8 (defined as in Eq. 15!) would be considered significant.

As a last comment we remark that in our analysis we have plainly ignored the fact that the *E. coli* data derive from a time series experiment. This appears not to be too harmful for the GGM selection process – at least part of the longitudinal correlation will be accounted for by empirically fitting the null distribution (see also Efron (2005a)).

4. Discussion and summary

In this paper we draw attention to the problem of the widespread and largely uncritical use of the standard covariance estimator in the analysis of functional genomics data. As a quick glance in any recent issue of a journal such as *Bioinformatics* or *BMC Bioinformatics* will reveal, the empirical correlation and covariance estimators are often rather blindly applied by bioinformaticians to large-scale problems with many variables and few sample points although it is well known that in this setting the standard estimators are not appropriate and may perform extremely poorly. Here, we strongly advise to refrain from using the empirical covariance in the analysis of high-dimensional data such as from microarray or proteomics experiments.

We emphasize that alternatives are readily available in the form of shrinkage

estimators (e.g. Greenland, 2000). Shrinkage formalizes the idea of "borrowing strength across variables" and has proved beneficial in the problem of differential expression (e.g., Smyth, 2004; Cui et al., 2005) and classification of transcriptome data (e.g., Tibshirani et al., 2002; Zhu and Hastie, 2004). In this paper we particularly highlight the shrinkage approach of Ledoit and Wolf (2003) that allows fitting of all necessary tuning parameters in a simple *analytical* fashion. While this method appears to be little known we anticipate that it will be helpful in many "small n, large p" inference problems.

In Section 2 of this paper we present a novel shrinkage estimator for the covariance and correlation matrix (Tab. 1) with guaranteed minimum MSE and positive definiteness that is not only perfectly applicable to "small n, large p" data but can also be computed in time comparable to that of the conventional estimator. By use of the theorem of Ledoit and Wolf (2003) to estimate the optimal shrinkage intensity there is no need to specify any further parameters. Consequently, computationally expensive procedures such as cross-validation are completely avoided. As an added bonus, the proposed estimator is also distribution-free and demands only modest assumptions with regard to the existence of higher moments.

As a specific bioinformatical application we employ this covariance shrinkage estimator in the search for net-like genetic interactions. In Section 3 we show that this leads to large overall gains in the accuracy and in the power to recover the true network structure compared with a precursor approach described in Schäfer and Strimmer (2005a). In addition, our algorithm also outperforms the lasso approach to regularized GGM inference in terms of positive predictive accuracy. Furthermore, network inference by using the shrinkage covariance estimator (Section 2) combined with the heuristic model selection of Section 3 takes only a few minutes even on a slow computer – thus we offer it as a fast alternative to exhaustive GGM search procedures, such as the MCMC method of Dobra et al. (2004).

Further possible uses of the proposed shrinkage covariance estimator in bioinformatics include classification of gene expression profiles. For instance, the SCRDA ("shrunken centroids regularized discriminant analysis") approach (Guo et al., 2004) employs regularized covariance and correlation matrices similar to the one described in Section 2. Hence, it should be straightforward to apply SCRDA also in conjunction with our proposed shrinkage covariance estimator.

We end with a note that "small n, large p" covariance estimation problems have recently arisen also in computational econometrics. Specifically, the inference and modeling of large financial networks (Mantegna and Stanley, 2000; Boginski et al., 2005) requires methods akin to those for gene relevance and association networks.

A. Estimation of the variance and covariance of the components of the S and R matrix

In order to compute the optimal estimated shrinkage intensity $\hat{\lambda}^\star$ (Eq. 8) for the various structured covariance targets listed in Tab. 2, it is necessary to obtain unbiased estimates of the variance and the covariance of individual entries in the matrix $\boldsymbol{S} = (s_{ij})$.

Let x_{ki} be the k-th observation of the variable X_i and $\bar{x_i} = \frac{1}{n}\sum_{k=1}^{n} x_{ki}$ its empirical mean. Now set $w_{kij} = (x_{ki} - \bar{x_i})(x_{kj} - \bar{x_j})$ and $\bar{w}_{ij} = \frac{1}{n}\sum_{k=1}^{n} w_{kij}$. Then the unbiased empirical covariance equals

$$\widehat{\mathrm{Cov}}(x_i, x_j) = s_{ij} = \frac{n}{n-1}\bar{w}_{ij}$$

and, correspondingly, the variance is

$$\widehat{\mathrm{Var}}(x_i) = s_{ii} = \frac{n}{n-1}\bar{w}_{ii}.$$

The empirical unbiased variances and covariances of the *individual entries* of $\boldsymbol{S}$ are computed in a similar fashion.

$$\widehat{\mathrm{Var}}(s_{ij}) = \frac{n^2}{(n-1)^2}\widehat{\mathrm{Var}}(\bar{w}_{ij}) = \frac{n}{(n-1)^2}\widehat{\mathrm{Var}}(w_{ij}) = \frac{n}{(n-1)^3}\sum_{k=1}^{n}(w_{kij} - \bar{w}_{ij})^2.$$

Similarly,

$$\widehat{\mathrm{Cov}}(s_{ij}, s_{lm}) = \frac{n}{(n-1)^3}\sum_{k=1}^{n}(w_{kij} - \bar{w}_{ij})(w_{klm} - \bar{w}_{lm}).$$

Moments of higher order than $\widehat{\mathrm{Var}}(s_{ij})$, in particular variances and covariances of *averages* of s_{ij}, are neglected in estimating the optimal $\hat{\lambda}^\star$ in Tab. 2.

The variance $\mathrm{Var}(r_{ij})$ of the empirical correlation coefficients can be estimated in a similar fashion: simply apply the above formulae to the *standardized* data matrix. We note that this procedure treats the estimated variances as constants and hence introduces a slight but generally negligible error. The same assumption also justifies to ignore the bias of the empirical correlation coefficients in Eq. 10.

B. Available computer software

The shrinkage estimator of the covariance matrix described in this paper is implemented in the R package "corpcor". This package also contains functions for computing (partial) correlations. The analysis and visualisation of the gene expression data was performed using the "GeneTS" R package. Both packages are distributed under the GNU General Public License and are available for download from the CRAN archive at `http://cran.r-project.org`. "GeneTS" is also available from Bioconductor (`http://www.bioconductor.org`) and from `http://www.statistik.lmu.de/~strimmer/software/genets/`.

References

Barabási, A.-L. and Z. N. Oltvai (2004). Network biology: understanding the cell's functional organization. *Nature Rev. Genetics 5*, 101–113.

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc. B 57*, 289–300.

Boginski, V., S. Butenko, and P. M. Pardalos (2005). Statistical analysis of financial networks. *Comp. Stat. Data Anal. 48*, 431–443.

Butte, A. J., P. Tamayo, D. Slonim, T. R. Golub, and I. S. Kohane (2000). Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks. *Proc. Natl. Acad. Sci. USA 97*, 12182–12186.

Cox, D. R. and N. Reid (2004). A note on pseudolikelihood from marginal densities. *Biometrika 91*, 729–737.

Cui, X., J. T. G. Hwang, J. Qiu, N. J. Blades, and G. A. Churchill (2005). Improved statistical tests for differential gene expression by shrinking variance components estimates. *Biostatistics 6*, 59–75.

Daniels, M. J. and R. E. Kass (2001). Shrinkage estimators for covariance matrices. *Biometrics 57*, 1173–1184.

de la Fuente, A., N. Bing, I. Hoeschele, and P. Mendes (2004). Discovery of meaningful associations in genomic data using partial correlation coefficients. *Bioinformatics 20*, 3565–3574.

Dobra, A., C. Hans, B. Jones, J. R. Nevins, G. Yao, and M. West (2004). Sparse graphical models for exploring gene expression data. *J. Multiv. Anal. 90*, 196–212.

Efron, B. (1975). Biased versus unbiased estimation. *Adv. Math. 16*, 259–277.

Efron, B. (1982). Maximum likelihood and decision theory. *Ann. Statist. 10*, 340–356.

Efron, B. (2004). Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *J. Amer. Statist. Assoc. 99*, 96–104.

Efron, B. (2005a). Correlation and large-scale simultaneous significance testing. Preprint, Dept. of Statistics, Stanford University.

Efron, B. (2005b). Local false discovery rates. Preprint, Dept. of Statistics, Stanford University.

Efron, B. and C. N. Morris (1975). Data analysis using Stein's estimator and its generalizations. *J. Amer. Statist. Assoc. 70*, 311–319.

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

Eisen, M. B., P. T. Spellman, P. O. Brown, and D. Botstein (1998). Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA 95*, 14863–14868.

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

Greenland, S. (2000). Principles of multilevel modelling. *Intl. J. Epidemiol. 29*, 158–167.

Guo, Y., T. Hastie, and T. Tibshirani (2004). Regularized discriminant analysis and its application in microarray. Preprint, Dept. of Statistics, Stanford University.

Hastie, T., R. Tibshirani, and J. Friedman (2001). *The Elements of Statistical Learning*. New York: Springer.

Higham, N. J. (1988). Computing a nearest symmetric positive semidefinite matrix. *Linear Algebra Appl. 103*, 103–118.

Hirschberger, M., Y. Qi, and R. E. Steuer (2004). Randomly generating portfolio-selection covariance matrices with specified distributional assumption. Preprint, Terry College of Business, University of Georgia.

Hoerl, A. E. and R. W. Kennard (1970a). Ridge regression: applications to nonorthogonal problems. *Technometrics 12*, 69–82.

Hoerl, A. E. and R. W. Kennard (1970b). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics 12*, 55–67.

Hotelling, H. (1953). New light on the correlation coefficient and its transforms. *J. R. Statist. Soc. B 15*, 193–232.

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

Ledoit, O. and M. Wolf (2004a). Honey, I shrunk the sample covariance matrix. *J. Portfolio Management 30*, 110–119.

Ledoit, O. and M. Wolf (2004b). A well conditioned estimator for large-dimensional covariance matrices. *J. Multiv. Anal. 88*, 365–411.

Leung, P. L. and W. Y. Chan (1998). Estimation of the scale matrix and its eigenvalues in the Wishart and the multivariate F distributions. *Ann. Inst. Statist. Math. 50*, 523–530.

Magwene, P. M. and J. Kim (2004). Estimating genomic coexpression networks using first-order conditional independence. *Genome Biology 5*, R100.

Mantegna, R. N. and H. E. Stanley (2000). *An Introduction to Econophysics: Correlations and Complexity in Finance*. Cambridge, UK: Cambridge University Press.

Meinshausen, N. and P. Bühlmann (2005). High dimensional graphs and variable selection with the lasso. *Ann. Statist.* in press.

Morris, C. N. (1983). Parametric empirical Bayes inference: theory and applications. *J. Amer. Statist. Assoc. 78*, 47–55.

Schäfer, J. and K. Strimmer (2005a). An empirical Bayes approach to inferring large-scale gene association networks. *Bioinformatics 21*, 754–764.

Schäfer, J. and K. Strimmer (2005b). Learning large-scale graphical Gaussian models from genomic data. In J. F. F. Mendes, S. N. Dorogovtsev, A. Povolotsky, F. V. Abreu, and J. G. Oliveira (Eds.), *Science of Complex Networks: From Biology to the Internet and WWW*, Volume 776, AIP Conference Proceedings, Aveiro, PT, August 2004, pp. 263–276. American Institute of Physics.

Schmidt-Heck, W., R. Guthke, S. Toepfer, H. Reischer, K. Duerrschmid, and K. Bayer (2004). Reverse engineering of the stress response during expression of a recombinant protein. In *Proceedings of the EUNITE symposium, 10-12 June 2004, Aachen, Germany*, pp. 407–412. Verlag Mainz.

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

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate distribution. In J. Neyman (Ed.), *Proc. Third. Berkeley Symp. Math. Statist. Probab.*, Volume 1, Berkeley, pp. 197–206. Univ. California Press.

Storey, J. D. (2002). A direct approach to false discovery rates. *J. R. Statist. Soc. B 64*, 479–498.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Statist. Soc. B 58*, 267–288.

Tibshirani, R., T. Hastie, B. Narasimhan, and G. Chi (2002). Diagnosis of multiple cancer type by shrunken centroids of gene expression. *Proc. Natl. Acad. Sci. USA 99*, 6567–6572.

Toh, H. and K. Horimoto (2002). Inference of a genetic network by a combined approach of cluster analysis and graphical Gaussian modeling. *Bioinformatics 18*, 287–297.

Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*. New York: Wiley.

Wille, A., P. Zimmermann, E. Vranová, A. Fürholz, O. Laule, S. Bleuler, L. Hennig, A. Prelić, P. von Rohr, L. Thiele, E. Zitzler, W. Gruissem, and P. Bühlmann (2004). Sparse graphical Gaussian modeling of the isoprenoid gene network in Arabidopsis thaliana. *Genome Biology 5*, R92.

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