

Ridge Estimation of Inverse Covariance Matrices from High-Dimensional Data

Wessel N. van Wieringen^{a,b,*}, Carel F.W. Peeters^a

^a*Department of Epidemiology & Biostatistics, VU University medical center Amsterdam, Postbus 7057, 1007 MB Amsterdam, The Netherlands*

^b*Department of Mathematics, VU University Amsterdam, 1081 HV Amsterdam, The Netherlands*

Abstract

We study ridge estimation of the precision matrix in the high-dimensional setting where the number of variables is large relative to the sample size. We first review two archetypal ridge estimators and note that their penalties do not coincide with common quadratic ridge penalties. Subsequently, starting from a proper ℓ_2 -penalty, analytic expressions are derived for two alternative ridge estimators of the precision matrix. The alternative estimators are compared to the archetypes with regard to eigenvalue shrinkage and risk. The alternatives are also compared to the graphical lasso within the context of graphical modeling. The comparisons may give reason to prefer the proposed alternative estimators.

Keywords: graphical modeling, high-dimensional precision matrix estimation, multivariate normal, ℓ_2 -penalization, precision matrix

1. Introduction

Let \mathbf{Y}_i , $i = 1, \dots, n$, be a p -dimensional random variate drawn from $\mathcal{N}_p(\mathbf{0}, \mathbf{\Sigma})$. The maximum likelihood (ML) estimator of the precision matrix $\mathbf{\Omega} = \mathbf{\Sigma}^{-1}$ maximizes:

$$\mathcal{L}(\mathbf{\Omega}; \mathbf{S}) \propto \ln |\mathbf{\Omega}| - \text{tr}(\mathbf{S}\mathbf{\Omega}), \quad (1)$$

where \mathbf{S} is the sample covariance estimate. If $n > p$, the log-likelihood achieves its maximum for $\hat{\mathbf{\Omega}}^{\text{ML}} = \mathbf{S}^{-1}$.

In the high-dimensional setting where $p > n$, the sample covariance matrix is singular and its inverse is undefined. Consequently, so is $\hat{\mathbf{\Omega}}^{\text{ML}}$. A common workaround is the addition of a penalty to the log-likelihood (1). The ℓ_1 -penalized estimation of the precision matrix was considered almost simultaneously by [1], [2], [3], and [4]. This (graphical) lasso estimate of $\mathbf{\Omega}$ has attracted much attention due to the resulting sparse solution. Juxtaposed to situations in which sparsity is an asset are situations in which one is intrinsically interested in more accurate representations of the high-dimensional precision

*Principal corresponding author

Email addresses: w.vanwieringen@vumc.nl (Wessel N. van Wieringen), cf.peeters@vumc.nl (Carel F.W. Peeters)

Preprint submitted to Elsevier

September 25, 2015

matrix. In addition, the true (graphical) model need not be (extremely) sparse in terms of containing many zero elements. In these cases we may prefer usage of a regularization method that shrinks the estimated elements of the precision matrix proportionally [5] in possible conjunction with some form of post-hoc element selection. It is such estimators we consider.

We thus study ridge estimation of the precision matrix. We first review two archetypal ridge estimators and note that their penalties do not coincide with what is perceived to be the common ridge penalty (Section 2). Subsequently, starting from a common ridge penalty, analytic expressions are derived for alternative ridge estimators of the precision matrix in Section 3. This section, in addition, studies properties of the alternative estimators and proposes a method for choosing the penalty parameter. In Section 4 the alternative estimators are compared to their corresponding archetypes w.r.t. eigenvalue shrinkage. In addition, the risks of the various estimators are assessed under multiple loss functions, revealing the superiority of the proposed alternatives. Section 5 compares the alternative estimators to the graphical lasso in a graphical modeling setting using oncogenomics data. This comparison points to certain favorable behaviors of the proposed alternatives with respect to loss, sensitivity, and specificity. In addition, Section 5 demonstrates that the alternative ridge estimators yield more stable networks vis-à-vis the graphical lasso, in particular for more extreme p/n ratios. This section thus provides empirical evidence in the graphical modeling setting of what is tacitly known from regression (subset selection) problems: ridge penalties coupled with post-hoc selection may outperform the lasso. We conclude with a discussion (Section 6).

2. Archetypal Ridge Estimators

Ridge estimators of the precision matrix currently in use can be roughly divided into two archetypes [cf. 6, 7]. The first archetypal form of ridge estimator commonly is a convex combination of \mathbf{S} and a positive definite (p.d.) target matrix $\mathbf{\Gamma}$: $\hat{\mathbf{\Omega}}^I(\lambda_I) = [(1 - \lambda_I)\mathbf{S} + \lambda_I\mathbf{\Gamma}]^{-1}$, with $\lambda_I \in (0, 1]$. A common (low-dimensional) target choice is $\mathbf{\Gamma}$ diagonal with $(\mathbf{\Gamma})_{jj} = (\mathbf{S})_{jj}$ for $j = 1, \dots, p$. This estimator has the desirable property of shrinking to $\mathbf{\Gamma}^{-1}$ when $\lambda_I = 1$ (maximum penalization). The estimator can be motivated from the bias-variance tradeoff as it seeks to balance the high-variance, low-bias matrix \mathbf{S} with the lower-variance, higher-bias matrix $\mathbf{\Gamma}$. It can also be viewed as resulting from the maximization of the following penalized log-likelihood:

$$\ln |\mathbf{\Omega}| - (1 - \lambda_I)\text{tr}(\mathbf{S}\mathbf{\Omega}) - \lambda_I\text{tr}(\mathbf{\Omega}\mathbf{\Gamma}). \quad (2)$$

The penalized log-likelihood (2) is obtained from the original log-likelihood (1) by the replacement of \mathbf{S} by $(1 - \lambda_I)\mathbf{S}$ and the addition of a penalty. The estimate $\hat{\mathbf{\Omega}}^I(\lambda_I)$ can thus be viewed as a penalized ML estimate.

The second archetype finds its historical base in ridge regression, a technique that started as an ad-hoc modification for dealing with singularity in the least squares normal equations. The archetypal second form of the ridge precision matrix estimate would be $\hat{\mathbf{\Omega}}^{II}(\lambda_{II}) = (\mathbf{S} + \lambda_{II}\mathbf{I}_p)^{-1}$ with $\lambda_{II} \in (0, \infty)$. It can be motivated as an ad-hoc fix of the singularity of \mathbf{S} in the high-dimensional setting, much like how ridge regression was originally introduced by [8]. Alternatively, this archetype too can be viewed as a

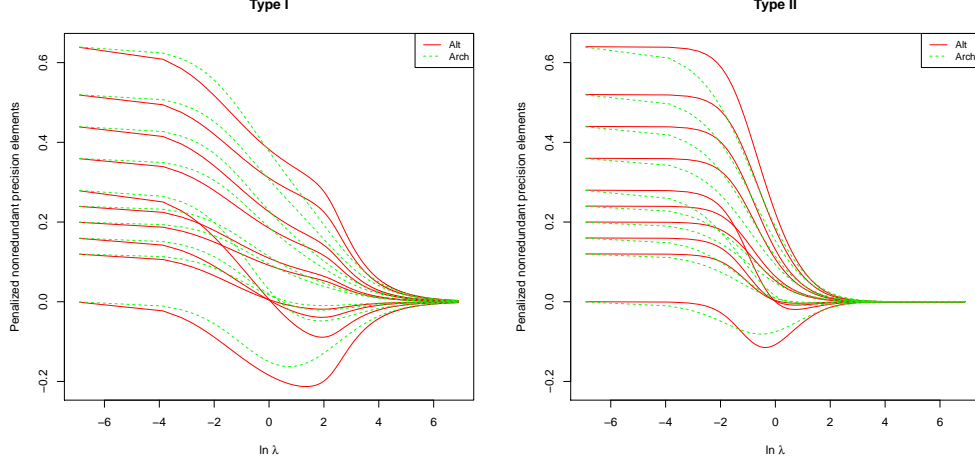


Figure 1: Ridge coefficient paths of nonredundant off-diagonal elements for the archetypal (dashed green) and alternative (solid red) Type I (left panel) and Type II (right panel) ridge estimators. The 5×5 matrix \mathbf{S} was generated as $(\mathbf{S}^{-1})_{j_1, j_2} = [(j_1 \times j_2 + 1) \bmod 21]/25$ if $j_1 \neq j_2$ and $(\mathbf{S}^{-1})_{j_1, j_2} = 1$ if $j_1 = j_2$. The target matrix in the Type I case was taken to be the identity matrix \mathbf{I}_5 . The penalty parameter is generically indicated by λ . For archetypal-to-alternative scaling of the penalty parameters under Type I and Type II estimation see Section 4.1.

penalized estimate, as it maximizes [see also 9]:

$$\ln |\mathbf{\Omega}| - \text{tr}(\mathbf{S}\mathbf{\Omega}) - \lambda_{\text{II}} \text{tr}(\mathbf{\Omega}\mathbf{I}_p). \quad (3)$$

The penalties in (2) and (3) are non-concave (their second order derivatives equal the null-matrix $\mathbf{0}$). This, however, poses no problem under the restriction of a p.d. solution $\mathbf{\Omega}$ as the Hessian of both (2) and (3) equals $-\mathbf{\Omega}^{-2}$. More surprising is that neither penalty of the two current archetypes resembles the precision-analogy of what is commonly perceived as the ridge ℓ_2 -penalty: $\frac{1}{2}\lambda\|\mathbf{\Omega}\|_2^2 = \frac{1}{2}\lambda\sum_{j_1=1}^p\sum_{j_2=1}^p[(\mathbf{\Omega})_{j_1, j_2}]^2$.

The graphical lasso uses a penalty that is in line with the ℓ_1 -penalty of lasso regression. It is a similar objective we have in the remainder. We embark on the derivation of alternative Type I and Type II (graphical) ridge estimators using a proper ℓ_2 -penalty. Consider Figure 1 to get a flavor of the behavior of both the archetypal ridge precision matrix estimators and our alternatives (receiving analytic justification in Section 3). It is seen that ridge estimation based on a proper ridge penalty induces (slight) differences in behavior. Differences that will be shown to point to the preferability of the alternative estimators in Section 4.

3. Alternative Ridge Estimators of the Precision Matrix

In this section we derive analytic expressions for alternative Type I and Type II ridge precision estimators. In addition, we explore their moments (Section 3.3) and consistency (Section 3.4) as well as methods for choosing the penalty parameter (Section 3.5). Proofs (as indeed all proofs in the remainder) are deferred to Appendix A.

3.1. Type I

In this section an analytic expression for an alternative Type I ridge precision estimator is given. Before arriving at a proposition containing some properties of this estimator, we employ the following lemma:

Lemma 1 (Alternative Type I ridge precision estimator). *Amend the log-likelihood (1) with the ℓ_2 -penalty*

$$\frac{\lambda_a}{2} \text{tr} [(\mathbf{\Omega} - \mathbf{T})^T (\mathbf{\Omega} - \mathbf{T})], \quad (4)$$

with \mathbf{T} denoting a symmetric p.d. target matrix, and where $\lambda_a \in (0, \infty)$ denotes a penalty parameter. Under given penalty, an alternative (penalized ML) Type I ridge estimator is obtained as:

$$\hat{\mathbf{\Omega}}^{\text{Ia}}(\lambda_a) = \left\{ \left[\lambda_a \mathbf{I}_p + \frac{1}{4} (\mathbf{S} - \lambda_a \mathbf{T})^2 \right]^{1/2} + \frac{1}{2} (\mathbf{S} - \lambda_a \mathbf{T}) \right\}^{-1}. \quad (5)$$

Proposition 1. *Consider the alternative Type I ridge estimator (5) from Lemma 1. For this estimator, the following properties hold:*

- i. $\hat{\mathbf{\Omega}}^{\text{Ia}}(\lambda_a) \succ 0$, for all $\lambda_a \in (0, \infty)$;
- ii. $\lim_{\lambda_a \rightarrow 0^+} \hat{\mathbf{\Omega}}^{\text{Ia}}(\lambda_a) = \mathbf{S}^{-1}$;
- iii. $\lim_{\lambda_a \rightarrow \infty} \hat{\mathbf{\Omega}}^{\text{Ia}}(\lambda_a) = \mathbf{T}$.

limits of the proposed estimator are the (possibly nonexistent) inverse of the ML estimator \mathbf{S} and a target matrix, respectively. For a fuller understanding of the estimator (5), consider the following remarks.

Remark 1. The target matrix \mathbf{T} from Lemma 1 may in principle be nonnegative definite (n.d.) for the statement to hold. As should be clear from Proposition 1, however, choosing an n.d. target may lead to ill-conditioned estimates in the limit. Moreover, from a shrinkage perspective, the interpretability of a p.d. target may be deemed superior. Hence, Lemma 1 assumes the target matrix to be p.d. (as does the archetypal Type I estimator). Section 3.2 considers as a special case the n.d. choice $\mathbf{T} = \mathbf{0}$, in order to arrive at an alternative for the archetypal Type II estimator.

Remark 2. It may be noticed that the penalty term (4) amounts to a proper ridge penalty as $\frac{\lambda_a}{2} \text{tr} [(\mathbf{\Omega} - \mathbf{T})^T (\mathbf{\Omega} - \mathbf{T})] = \frac{\lambda_a}{2} \|\mathbf{\Omega} - \mathbf{T}\|_2^2$. When $\mathbf{T} = \mathbf{0}$, we obtain $\frac{\lambda_a}{2} \|\mathbf{\Omega}\|_2^2$; a special case that will be considered in Section 3.2.

Remark 3. From Proposition 1 it is clear that (5) is always p.d. when $\lambda_a \in (0, \infty)$. However, as with any regularized covariance or precision estimator, the estimate is not necessarily *well-conditioned* (in terms of, say, the spectral condition number) for any $\lambda_a \in (0, \infty)$ when \mathbf{S} is ill-behaved. To obtain a well-conditioned estimate in such situations, one should choose λ_a not too close to zero. In order to choose an optimal value of λ_a for a problem at hand, one can employ (approximate) cross-validation or information criteria (see Section 3.5).

Remark 4. Lemma 1 considers regularized estimation of the precision matrix. It may also provide an alternative Type I regularized estimator for the covariance matrix, by entertaining

$$[\hat{\Omega}^{\text{Ia}}(\lambda_a)]^{-1} \equiv \hat{\Sigma}^{\text{Ia}}(\lambda_a) = \left[\lambda_a \mathbf{I}_p + \frac{1}{4}(\mathbf{S} - \lambda_a \mathbf{T})^2 \right]^{1/2} + \frac{1}{2}(\mathbf{S} - \lambda_a \mathbf{T}).$$

Then: (i) $\hat{\Sigma}^{\text{Ia}}(\lambda_a) \succ 0$, for all $\lambda_a > 0$; (ii) $\lim_{\lambda_a \rightarrow 0^+} \hat{\Sigma}^{\text{Ia}}(\lambda_a) = \mathbf{S}$; (iii) $\lim_{\lambda_a \rightarrow \infty^-} \hat{\Sigma}^{\text{Ia}}(\lambda_a) = \mathbf{T}^{-1}$. Say one wishes to shrink to a p.d. covariance target \mathbf{C} , one only has to specify $\mathbf{T} = \mathbf{C}^{-1}$ in this case.

Remark 5. We note that (5) can also be obtained without inversion, by noticing

$$\hat{\Omega}^{\text{Ia}}(\lambda_a) = \frac{1}{\lambda_a} \left[\hat{\Sigma}^{\text{Ia}}(\lambda_a) - (\mathbf{S} - \lambda_a \mathbf{T}) \right].$$

The basis for this claim is expression (8) from Section 3.3.

3.2. Type II

An alternative Type II ridge estimator for the precision matrix can be found as a special case of Lemma 1:

Corollary 1 (Alternative Type II ridge precision estimator). *Consider the alternative Type I ridge estimator (5) from Lemma 1. An alternative ridge proper Type II estimator is obtained by choosing $\mathbf{T} = \mathbf{0}$, such that*

$$\hat{\Omega}^{\text{IIa}}(\lambda_a) = \left\{ \left[\lambda_a \mathbf{I}_p + \frac{1}{4} \mathbf{S}^2 \right]^{1/2} + \frac{1}{2} \mathbf{S} \right\}^{-1}. \quad (6)$$

For this estimator, the following properties hold:

- i. $\hat{\Omega}^{\text{IIa}}(\lambda_a) \succ 0$, for all $\lambda_a \in (0, \infty)$;
- ii. $\lim_{\lambda_a \rightarrow 0^+} \hat{\Omega}^{\text{IIa}}(\lambda_a) = \mathbf{S}^{-1}$;
- iii. $\lim_{\lambda_a \rightarrow \infty^-} \hat{\Omega}^{\text{IIa}}(\lambda_a) = \mathbf{0}$.

Similar to the archetypal II estimator, the right and left-hand limits are the (possibly nonexistent) inverse of the ML estimator \mathbf{S} and the null-matrix, respectively. The alternative Type II analogies of Remarks 3–5 hold for (6). Note that the estimator (6) was also considered by [10] in a different setting.

3.3. Moments

The explicit expressions for the alternative (Type I and II) ridge estimators facilitate the study of their properties. For instance, the moments of the ridge covariance and precision estimators can – in principle – be evaluated numerically to any desired degree of accuracy. Consider the following exemplification. With respect to the alternative Type I estimator we write:

$$\hat{\Sigma}^{\text{Ia}}(\lambda_a) = \sqrt{\lambda_a} \left[(\mathbf{I}_p + \mathbf{U}^2)^{1/2} + \mathbf{U} \right],$$

where $\mathbf{U} = (\mathbf{S} - \lambda_a \mathbf{T}) / (2\sqrt{\lambda_a})$. Express the $(1 + x^2)^{1/2}$ term as a binomial series to obtain the series representation of the ridge covariance estimator:

$$\hat{\Sigma}^{Ia}(\lambda_a) = \sqrt{\lambda_a} \mathbf{U} + \sqrt{\lambda_a} \sum_{q=0}^{\infty} \binom{1/2}{q} \mathbf{U}^{2q}.$$

Now, taking the expectation of the right-hand side yields the first moment of the alternative Type I ridge covariance estimator. To evaluate this expectation note that (under normality) \mathbf{S} follows a (singular) Wishart distribution, assume \mathbf{T} to be non-random, and restrict the binomial series to the degree that produces the desired accuracy. It then suffices to plug in the required moments of the Wishart distribution.

From the moments of the ridge covariance estimator one can directly obtain the moments of the ridge precision estimator. Hereto we need the identity:

$$2\sqrt{\lambda_a} \mathbf{U} = \sqrt{\lambda_a} \left[(\mathbf{I}_p + \mathbf{U}^2)^{1/2} + \mathbf{U} \right] - \sqrt{\lambda_a} \left[(\mathbf{I}_p + \mathbf{U}^2)^{1/2} + \mathbf{U} \right]^{-1}, \quad (7)$$

with \mathbf{U} as above. This equality is immediate after noting that all terms have the same eigenvectors and using ready algebra to prove the identity $2x = x + (1 + x^2)^{1/2} - [x + (1 + x^2)^{1/2}]^{-1}$, which applies to each eigenvalue in the eigen-decomposition (see also Section 4.1) of (7) separately. Reformulated we then have:

$$\mathbf{S} - \lambda_a \mathbf{T} = \hat{\Sigma}^{Ia}(\lambda_a) - \lambda_a \hat{\Omega}^{Ia}(\lambda_a). \quad (8)$$

This identity thus yields, via the moments of the alternative Type I ridge covariance matrix, the moments of the alternative Type I ridge precision matrix. The moments of the alternative Type II estimator can be obtained when considering \mathbf{T} to be the null-matrix.

Being able to evaluate the moments facilitates, e.g., the approximation of the bias of the proposed ridge estimators. Hereto assume $\mathbf{Y}_i \sim \mathcal{N}_p(\mathbf{0}, \Sigma)$ for $i = 1, \dots, n$. Define the sample covariance matrix $\mathbf{S} = \frac{1}{n} \sum_{i=1}^n \mathbf{Y}_i \mathbf{Y}_i^T$. Then, it is well-known that $n\mathbf{S}$ follows the Wishart distribution $\mathcal{W}_p(\Sigma, n)$. Recently, [11] have shown how $\mathbb{E}(n^b \mathbf{S}^b)$ may be derived analytically when $b \in \mathbb{Z}$. Their results are exploited here to approximate the bias of the proposed ridge estimators. When we ignore terms of order three and higher and limit ourselves to the type II estimator with $\mathbf{T} = \mathbf{0}$, the expectation may be approximated by (see also Section 1 of the Supplementary Material):

$$\begin{aligned} \mathbb{E} \left[\hat{\Sigma}^{IIa}(\lambda_a) \right] &\approx \frac{1}{2} \mathbb{E}(\mathbf{S}) + \sqrt{\lambda_a} \mathbf{I}_p + \frac{1}{8\sqrt{\lambda_a}} \mathbb{E}(\mathbf{S}^2) \\ &= \frac{1}{2} \Sigma + \sqrt{\lambda_a} \mathbf{I}_p + \frac{1}{8\sqrt{\lambda_a}} \left[\frac{n+1}{n} \Sigma^2 + \frac{1}{n} \text{tr}(\Sigma) \Sigma \right], \end{aligned}$$

in which the expectations of \mathbf{S} and \mathbf{S}^2 are obtained from [11]. Section 1 of the Supplementary Material contains a higher-order approximation and a simulation illustrating the accuracy of the approximation.

3.4. Consistency

We will show that the alternative Type I ridge estimator (5) is consistent under fixed-dimension asymptotics. To make this explicit, we (temporarily) modify the notation. Let

\mathbf{S}_n be the sample covariance matrix with index n indicating the sample size. Furthermore, the penalty parameter is now denoted $\lambda_{a,n}$. This explicates the fact that the penalty parameter is chosen in a data-driven fashion (cf. Section 3.5) and thus depends on the sample size. In particular, it will be assumed that $\lambda_{a,n}$ converges (in some sense) to zero as $n \rightarrow \infty^-$. This reflects the decreasing necessity to regularize the (inverse) covariance estimator as the sample size increases. Finally, let $\hat{\Sigma}_n^{\text{Ia}}(\lambda_{a,n})$ be the alternative ridge covariance estimator (see Remark 4) with \mathbf{S} and λ_a replaced by \mathbf{S}_n and $\lambda_{a,n}$.

In showing consistency, we need the asymptotic unbiasedness of our estimator. This property is warranted by the following lemma:

Lemma 2 (Asymptotic unbiasedness). *Let \mathbf{S}_n be the sample covariance matrix from a sample $\mathbf{Y}_1, \dots, \mathbf{Y}_n$ drawn from $\mathcal{N}_p(\mathbf{0}, \Sigma)$. Denote by $\lambda_{a,n}$ a nonnegative random variable that converges almost surely to zero and by \mathbf{T} a nonrandom p.d. symmetric matrix. Then:*

$$\lim_{n \rightarrow \infty^-} \mathbb{E} \left[\hat{\Sigma}_n^{\text{Ia}}(\lambda_{a,n}) \right] \longrightarrow \lim_{n \rightarrow \infty^-} \mathbb{E}(\mathbf{S}_n) = \Sigma.$$

Simultaneously, the expectation of its inverse $\hat{\Omega}_n^{\text{Ia}}(\lambda_{a,n})$ tends to $\Omega = \Sigma^{-1}$ as $n \rightarrow \infty^-$.

Lemma 2 follows directly from application of the continuous mapping theorem, the Portmanteau lemma and Slutsky's lemma [see, e.g., Theorem 2.3, Lemma 2.2, and Lemma 2.8 in 12]. By virtue of the same asymptotic results, the lemma may be generalized to allow \mathbf{T} to depend on data, as long as the data-dependent target \mathbf{T}_n converges (almost surely) to some \mathbf{T} .

Lemma 2 is conducive in proving the consistency result (note that asymptotic unbiasedness and consistency of the alternative Type II estimator (6) follow as special cases of Lemma 2 and Proposition 2):

Proposition 2 (Consistency). *Let \mathbf{S}_n be the sample covariance matrix from a sample $\mathbf{Y}_1, \dots, \mathbf{Y}_n$ drawn from $\mathcal{N}_p(\mathbf{0}, \Sigma)$. Denote by $\lambda_{a,n}$ a nonnegative random variable that converges almost surely to zero and by \mathbf{T} a nonrandom p.d. symmetric matrix. Then:*

$$\lim_{n \rightarrow \infty^-} \mathbb{E} \left(\left\| \hat{\Sigma}_n^{\text{Ia}}(\lambda_{a,n}) - \Sigma \right\|_F^2 \right) = 0,$$

where $\|\cdot\|_F$ denotes the Frobenius norm. Simultaneously, $\hat{\Omega}_n^{\text{Ia}}(\lambda_{a,n})$ consistently estimates $\Omega = \Sigma^{-1}$.

The consistency result in Proposition 2 takes p to be fixed, and thus does not concern increasing-dimension asymptotics (in which p also tends to infinity). This is motivated by practice. We have an applicatory focus on the reconstruction of (molecular) interaction networks (see also Section 5). The (maximum) number of variates of such systems is fixed. As such, the consistency result above is deemed appropriate.

3.5. Choosing λ_a

A well-informed choice of the penalty parameter λ_a is crucial in applications. The literature contains many proposals for selecting an (in some sense) optimal value for the penalty parameter in (precision) regularization problems. These can be classified [see

13] in methods aiming at model selection consistency (e.g., BIC, EBIC), and methods that aim to maximize predictive power (e.g., cross-validation, AIC). As the ℓ_2 -penalty does not automatically induce sparsity in the estimate, we are not after model selection consistency. Rather, in our case it is natural to seek loss efficiency.

While both cross-validation (CV) and AIC [14] have similar asymptotic properties in terms of minimizing Kullback-Leibler divergence, the data-driven nature of the former makes it prone to have superior behavior in terms of accuracy. The K -fold CV score for a generic regularized estimate $\hat{\Omega}(\lambda)$ based on the generic fixed penalty λ can be given as:

$$\varphi^K(\lambda) = \sum_{k=1}^K n_k \left\{ -\ln |\hat{\Omega}(\lambda)_{-k}| + \text{tr}[\hat{\Omega}(\lambda)_{-k} \mathbf{S}_k] \right\},$$

where n_k is the size of subset k , for $k = 1, \dots, K$ disjoint subsets. Further, \mathbf{S}_k denotes the sample covariance matrix based on subset k , while $\hat{\Omega}(\lambda)_{-k}$ denotes the estimated regularized precision matrix on all samples not in k . Highest predictive accuracy can be obtained by choosing $n_k = 1$, such that $K = n$. This is known as leave-one-out CV (LOOCV). Unfortunately, LOOCV (as K -fold CV in general) is computationally demanding for large p and/or large n .

Recently, [15] and [16] derived, based on the log-likelihood of the precision, an approximate solution to the LOOCV score. Based on their work, the approximate LOOCV score for fixed λ_a , $\tilde{\varphi}^n(\lambda_a)$, is given for the alternative Type I ridge estimator as:

$$\tilde{\varphi}^n(\lambda_a) = -\frac{1}{n} \mathcal{L}[\hat{\Omega}^{\text{Ia}}(\lambda_a); \mathbf{S}] + \frac{1}{2n(n-1)} \sum_{i=1}^n \gamma_i, \quad (9)$$

with

$$\gamma_i = \sum_{j_1=1}^p \sum_{j_2=1}^p \left\{ \left[[\hat{\Omega}^{\text{Ia}}(\lambda_a)]^{-1} - \mathbf{Y}_i \mathbf{Y}_i^T \right] \circ \left[\hat{\Omega}^{\text{Ia}}(\lambda_a) (\mathbf{S} - \mathbf{Y}_i \mathbf{Y}_i^T) \hat{\Omega}^{\text{Ia}}(\lambda_a) \right] \right\}_{j_1, j_2},$$

and where \circ denotes the Hadamard product. Naturally, the approximate LOOCV score for the alternative Type II ridge estimator can be obtained by replacing $\hat{\Omega}^{\text{Ia}}(\lambda_a)$ in (9) by $\hat{\Omega}^{\text{IIa}}(\lambda_a)$. We propose to choose λ_a^* such that $\lambda_a^* = \arg \min_{\lambda_a \in \mathbb{R}^+} \tilde{\varphi}^n(\lambda_a)$, which relates to the minimization of Kullback-Leibler divergence and the maximization of predictive accuracy. The expression $\tilde{\varphi}^n(\lambda)$ is computationally efficient, requiring only a single matrix inversion (as opposed to n inversions for $\varphi^n(\lambda)$). In addition, the Hadamard product has an efficient computational implementation [see 16].

Remark 6. We note that only a single spectral decomposition and a single matrix inversion are required in order to obtain the complete solution path (over any λ_a in the feasible domain) for the alternative Type II estimator and the alternative Type I estimator under a scalar matrix target choice (cf. Section 4.1). This implies that, in these cases, the computation of $\tilde{\varphi}^n(\lambda_a)$ over the (complete) solution path is particularly efficient. This efficiency, coupled with the benefits of knowing the full solution path, may be deemed to rival the benefits of a solution under an analytic choice of λ_a (see also next remark).

Remark 7. There exist analytic solutions to determining an optimal value for the penalty parameter. [17], e.g., determine analytically, under a modified Frobenius loss, the optimal value for the penalty parameter in an archetypal Type I setting under certain choices of \mathbf{T} . For practical applications, however, one still needs to approximate this optimal value, requiring variances and covariances of the individual entries of \mathbf{S} [7]. When the variable to observation ratio grows more extreme, the approximation may propose (overly) conservative or even negative penalty values as optimal [cf. 18, 7], giving us reason to prefer the computationally friendly, data-driven approach from above. In addition, (9) is generic, meaning it can be used under any p.d. choice of \mathbf{T} .

The estimates $\hat{\mathbf{\Omega}}^{\text{Ia}}(\lambda_a^*)$ and $\hat{\mathbf{\Omega}}^{\text{IIa}}(\lambda_a^*)$ may facilitate methods of (high-dimensional) data analysis in need of a precision (or covariance) matrix that is not (necessarily) sparse (cf. Sections 4 and 6). They may also be of interest in situations in which sparsity is required, such as graphical modeling. One may pair, in such situations, the proposed estimates with *a posteriori* methods of support determination (cf. Section 5).

4. Comparing Alternative and Archetypal Ridge Estimation

In this section the proposed alternative Type I and Type II ridge estimators are compared to their corresponding archetypes w.r.t. eigenvalue shrinkage (Section 4.1). Moreover, the alternative and archetypal estimators are subjected to a risk comparison (Section 4.2).

4.1. Eigenvalue Shrinkage

The alternative Type I estimator (5) is, as its archetypal counterpart, rotation equivariant when choosing the target to be a scalar matrix $\mathbf{T} = \psi \mathbf{I}_p$, with (for the alternative Type I estimator) $\psi \in [0, \infty)$. That is, the effect of the ridge penalty on the precision estimate is then equivalent to shrinkage of the eigenvalues of the unpenalized estimate \mathbf{S}^{-1} . To see this, let the eigen-decomposition of \mathbf{S} be \mathbf{VDV}^T where \mathbf{D} is a diagonal matrix with the eigenvalues of \mathbf{S} on the diagonal and \mathbf{V} denotes the matrix that contains the corresponding eigenvectors as columns. The orthogonality of \mathbf{V} implies $\mathbf{VV}^T = \mathbf{V}^T\mathbf{V} = \mathbf{I}_p$. We then rewrite, using $\mathbf{T} = \mathbf{I}_p$ for notational convenience, the inverse of (5) as follows:

$$\begin{aligned} [\hat{\mathbf{\Omega}}^{\text{Ia}}(\lambda_a)]^{-1} &= \left[\lambda_a \mathbf{VV}^T + \frac{1}{4}(\mathbf{VDV}^T - \lambda_a \mathbf{VV}^T)^2 \right]^{1/2} + \frac{1}{2}(\mathbf{VDV}^T - \lambda_a \mathbf{VV}^T) \\ &= \mathbf{V} \left\{ \left[\lambda_a \mathbf{I}_p + \frac{1}{4}(\mathbf{D} - \lambda_a \mathbf{I}_p)^2 \right]^{1/2} + \frac{1}{2}(\mathbf{D} - \lambda_a \mathbf{I}_p) \right\} \mathbf{V}^T, \end{aligned} \quad (10)$$

making clear that the ridge penalty deals with singularity and ill-conditioning through shrinkage of the eigenvalues of \mathbf{S}^{-1} . The alternative Type II estimator (6) also has the property of being rotation equivariant. This can be seen by:

$$[\hat{\mathbf{\Omega}}^{\text{IIa}}(\lambda_a)]^{-1} = \mathbf{V} \left[\left(\lambda_a \mathbf{I}_p + \frac{1}{4} \mathbf{D}^2 \right)^{1/2} + \frac{1}{2} \mathbf{D} \right] \mathbf{V}^T. \quad (11)$$

The equivariance property can be used in the comparison of eigenvalue shrinkage between the archetypes and alternatives. The following claims summarize:

Proposition 3. *Let the regularization parameters of the archetypal and alternative Type I ridge estimators – λ_I and λ_a respectively – map to the same scale. That is, choose $\lambda_I = 1 - 1/(\lambda_a + 1)$. In addition, consider a p.d. scalar matrix as the low-dimensional target matrix \mathbf{T} and let the archetypal Type I estimator have the same target in the precision sense, i.e., $\mathbf{\Gamma}^{-1} = \mathbf{T}$. Then the alternative estimator $\hat{\mathbf{\Omega}}^{Ia}(\lambda_a)$ displays shrinkage of the eigenvalues of \mathbf{S}^{-1} that is at least as heavy as the shrinkage propagated by the archetypal estimator $\hat{\mathbf{\Omega}}^I(\lambda_I)$.*

Proposition 4. *Let the regularization parameters of the archetypal and alternative Type II ridge estimators – λ_{II} and λ_a respectively – map to the same scale. That is, choose $\lambda_a = \lambda_{II}^2$. Then the archetypal estimator $\hat{\mathbf{\Omega}}^{II}(\lambda_{II})$ displays shrinkage of the eigenvalues of \mathbf{S}^{-1} that is at least as heavy as the shrinkage propagated by the alternative estimator $\hat{\mathbf{\Omega}}^{IIa}(\lambda_a)$.*

Corollary 2. *The eigenvalue inequality of Proposition 4 implies:*

$$\mathcal{L}[\hat{\mathbf{\Omega}}^{II}(\lambda_{II}); \mathbf{S}] \leq \mathcal{L}[\hat{\mathbf{\Omega}}^{IIa}(\lambda_a); \mathbf{S}].$$

The alternative Type I estimator displays *faster* shrinkage to the target \mathbf{T} than the archetypal Type I estimator. The alternative estimator then can be expected to have lower risk (in terms of, say, quadratic loss) than its archetypal counterpart when the (low-dimensional) target is an adequate representation of the true precision matrix. In such cases it can be shown under mild assumptions that, analogous to Corollary 2, $\mathcal{L}[\hat{\mathbf{\Omega}}^I(\lambda_I); \mathbf{S}] \leq \mathcal{L}[\hat{\mathbf{\Omega}}^{Ia}(\lambda_a); \mathbf{S}]$. In absence of a natural target \mathbf{T} , Type II estimators are an option. It is seen from proposition 4 that, as opposed to the Type I situation, the alternative Type II estimator displays *slower* shrinkage to the null-matrix than the archetypal Type II estimator. As the limiting null-matrix can indeed never be a good representation of the true precision matrix, the alternative Type II estimator can also be expected to have lower risk than its archetypal counterpart. The behavior of the alternative Type I and Type II estimators with regard to shrinkage rate may initially seem contradictory when evaluating Propositions 3 and 4. It is not if we notice that the penalty parameter λ_a is more influential in the Type I alternative as its effect is not diluted by a null \mathbf{T} . The topics of Loss and Risk are explored in the next subsection.

4.2. Risk

The risks of the alternative Type I and Type II estimators for the precision matrix are compared to that of Type I and II archetypes. Let $\mathbf{\Omega}$ denote a generic $(p \times p)$ population precision matrix and let $\hat{\mathbf{\Omega}}(\lambda)$ denote a generic ridge estimator of the precision matrix under generic regularization parameter λ . The following loss functions are then considered in risk evaluation:

a. Squared Frobenius loss, given by:

$$L_F[\hat{\mathbf{\Omega}}(\lambda), \mathbf{\Omega}] = \|\hat{\mathbf{\Omega}}(\lambda) - \mathbf{\Omega}\|_F^2;$$

b. Quadratic loss, given by:

$$L_Q[\hat{\mathbf{\Omega}}(\lambda), \mathbf{\Omega}] = \|\hat{\mathbf{\Omega}}(\lambda)\mathbf{\Omega}^{-1} - \mathbf{I}_p\|_F^2.$$

The risk \mathcal{R}_f of the estimator $\hat{\Omega}(\lambda)$ given a loss function L_f , $f \in \{F, Q\}$, is then defined as the expected loss:

$$\mathcal{R}_f[\hat{\Omega}(\lambda)] = \mathbb{E}\{L_f[\hat{\Omega}(\lambda), \Omega]\},$$

which is approximated by the median of losses over repeated simulation runs.

The risk is evaluated on data sets drawn from a multivariate normal distribution with four different (population) precision matrices:

1. Ω^{random} with no conditional dependencies, generated as $\Omega^{\text{random}} = \frac{1}{n} \mathbf{Y}^T \mathbf{Y}$ from the $(n \times p)$ -dimensional matrix \mathbf{Y} with $n = 10,000$ and each Y_{ij} drawn from $\mathcal{N}(0, 1)$;
2. Ω^{chain} representing a conditional independence graph with a chain topology. Its element are $(\Omega^{\text{chain}})_{j,j} = 1$, $(\Omega^{\text{chain}})_{j,j+1} = 0.25 = (\Omega^{\text{chain}})_{j+1,j}$ for $j = 1, \dots, p-1$, and zero otherwise;
3. Ω^{star} representing a conditional independence graph with a star topology. Its element are $(\Omega^{\text{star}})_{j,j} = 1$, $(\Omega^{\text{star}})_{1,j+1} = 1/(j+1) = (\Omega^{\text{star}})_{j+1,1}$ for $j = 1, \dots, p-1$, and zero otherwise;
4. Ω^{clique} representing a conditional independence graph with a clique structure. The structure consists of five equally sized blocks along the diagonal, each with unit diagonal elements and off-diagonal elements equal to 0.25.

Throughout the simulation the dimension of p is fixed at $p = 100$ while the sample size varies: $n = 5, 10$ and 25 . This represents varying degrees of high-dimensionality. For each combination of precision matrix and sample size one hundred data sets are drawn. For each draw the sample covariance matrix is calculated. The penalized estimates of the precision matrix are obtained for a large grid of the penalty parameter using the Type II null-matrix target ($\mathbf{T} = \mathbf{0}$), a diagonal target ($\text{diag}[\mathbf{T}] = 1/\text{diag}[\mathbf{S}]$), and a target equal to the true precision matrix ($\mathbf{T} = \Omega$). Note that in the comparison for the latter two Type I situations the archetypal target $\mathbf{\Gamma}$ is taken to be \mathbf{T}^{-1} , so that the archetypal and alternative estimators have the same target in the precision sense. For each penalized precision estimate the quadratic and Frobenius loss are evaluated and subsequently the risk (under given loss function) is approximated by the median loss over the hundred draws. Figure 2 shows, for the star topology, the estimated risks under quadratic loss for Type I ridge estimators ($\text{diag}[\mathbf{T}] = 1/\text{diag}[\mathbf{S}]$ and $\mathbf{T} = \Omega$) plotted against the penalty parameter (see Section 2 of the Supplementary Material for visualizations of all risk comparisons).

The simulation results (as summarized in Figure 2 and Section 2 of the Supplement) show that the alternative Type I ridge estimator outperforms its archetypal counterpart with respect to both loss types (when shrinking towards either of the non-zero targets). This behavior holds irrespective of the generated population precision matrix, the p/n ratio, and the choice of target. The superior performance of the alternative Type I estimator is strongest for small to medium-sized values of the penalty parameter (this will correspond, in practice, to the most relevant part of the domain). For large values of the penalty parameter the loss difference vanishes. This due to the fact that both alternative and archetype shrink to the same target. For both estimators the spot-on target ($\mathbf{T} = \Omega$) yields a lower loss for large values of λ than the diagonal target. The gain of employing a spot-on target increases, as can be expected, with the p/n ratio. With regard to Type II estimation the estimated risks of the alternative and archetypal

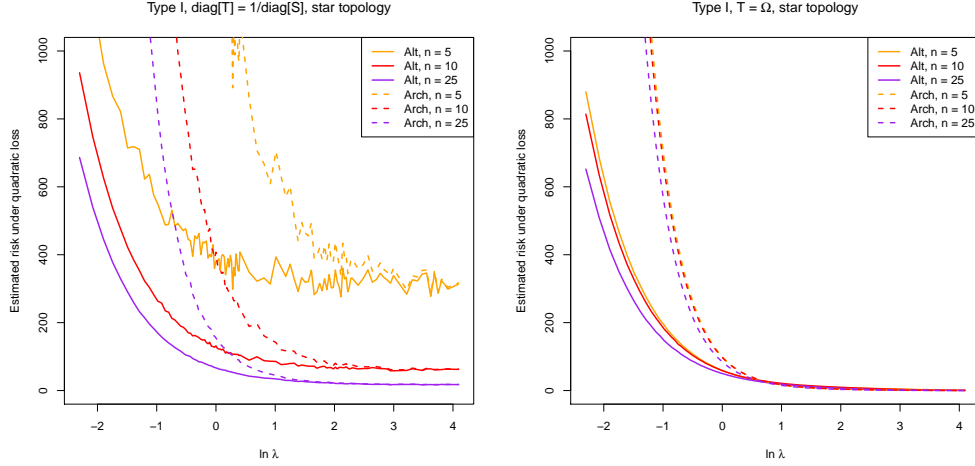


Figure 2: Estimated risk vs. penalty parameter. All panels display, for the star topology, the estimated risks under quadratic loss for Type I ridge estimators. The left panel compares the alternative and archetypal Type I ridge estimators when the target is taken to be $\text{diag}[\mathbf{T}] = 1/\text{diag}[\mathbf{S}]$. The right hand panel compares the alternative and archetypal Type I ridge estimators when $\mathbf{T} = \mathbf{\Omega}$. The dashed lines represent the archetypal estimator while the solid lines represent the alternative estimator. The orange, red and purple line colorings represent the various sample sizes ($n = 5, 10, 25$, respectively). Note that the fluctuations in the estimated risks in the left-hand panel are due to the data dependency of the target. Also note that, for purposes of comparability, the scales of the λ parameter under the various estimators were chosen in accordance with the eigenvalue comparison in Section 4.1.

estimators are similar, although the alternative estimator performs marginally better. In all, the alternative ridge precision estimators outperform their archetypal counterparts in this simulation study.

5. Comparing Alternative Ridge and Graphical Lasso Estimation

A contemporary use for precision matrices is found in network reconstruction through graphical modeling. Graphical modeling refers to a class of probabilistic models that uses graphs to express conditional (in)dependence relations between random variables. In this section we investigate how well the proposed ridge estimators of the precision matrix uncover conditional (in)dependencies from high-dimensional data. The performance of the alternative ridge estimators is contrasted with the graphical lasso [3]; the lasso estimator of the precision matrix. Two versions of each estimator are considered. On the ridge side the Type II alternative ridge precision estimator with $\mathbf{T} = \mathbf{0}$ and the Type I alternative ridge estimator with $\text{diag}[\mathbf{T}] = 1/\text{diag}[\mathbf{S}]$ are considered. The concordant graphical lasso precision estimators employ penalization and no penalization of the diagonal elements, respectively [see the `glasso` package: 19]. In order to avoid any bias towards either method of estimation, the comparison makes use of real data while adhering to the *ceteris paribus* principle with regard to penalty parameter selection (see also below). In the remainder of this section we will first review graphical modeling (Section 5.1) and the data (Section 5.2), before focusing the comparison on loss (Section 5.3), sensitivity

and specificity (Section 5.4), and network stability (Section 5.5), respectively.

5.1. Graphical Modeling

We consider graphs $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ consisting of a finite set \mathcal{V} of vertices and set of edges \mathcal{E} . The vertices of the graph correspond to a collection of random variables with probability distribution \mathcal{P} , i.e., $\{Y_1, \dots, Y_p\} \sim \mathcal{P}$. Edges in \mathcal{E} consist of pairs of distinct vertices such that $Y_j - Y_{j'} \in \mathcal{E}$. The basic assumption is: $\{Y_1, \dots, Y_p\} \sim \mathcal{N}_p(\mathbf{0}, \mathbf{\Sigma})$, with $\mathbf{\Sigma} \succ 0$. We thus focus on Gaussian graphical modeling by considering pairs $(\mathcal{G}, \mathcal{P} \sim \mathcal{N})$. (See Figure 6 for a visual example of a graphical model).

In this Gaussian case, conditional independence between a pair of variables corresponds to zero entries in the precision matrix. Indeed, let $\hat{\mathbf{\Omega}}$ denote a generic estimate of the precision matrix and consider its transformation to a partial correlation matrix $\hat{\mathbf{P}}$. Then the following relations can be shown to hold for all pairs $\{Y_j, Y_{j'}\} \in \mathcal{V}$ with $j \neq j'$ [see, e.g., 20]:

$$(\hat{\mathbf{P}})_{jj'} = 0 \iff (\hat{\mathbf{\Omega}})_{jj'} = 0 \iff Y_j \perp\!\!\!\perp Y_{j'} | \mathcal{V} \setminus \{Y_j, Y_{j'}\} \iff Y_j \neq Y_{j'},$$

where $\mathcal{V} \setminus \{\cdot\}$ denotes set-minus notation and where \neq indicates the absence of an edge. Hence, model selection efforts in Gaussian graphical models focus on determining the support of the precision matrix.

The graphical lasso [3] performs, next to shrinkage, automatic selection of conditional dependencies. As the alternative ridge estimators will not generally produce sparse estimates, they will need to rely on an additional procedure for support determination. Here, we resort to a multiple testing procedure. Specifically, we use the local false discovery rate (lfDR) procedure [21] proposed by [7]. Let $\hat{\mathbf{P}}^{Ia}(\lambda_a)$ denote the regularized precision estimate $\hat{\mathbf{\Omega}}^{Ia}(\lambda_a)$ scaled to partial correlation form. For support determination, we assume that the nonredundant off-diagonal partial correlation coefficients (indexed by, say, $j < j'$) follow a mixture distribution:

$$f \left\{ [\hat{\mathbf{P}}^{Ia}(\lambda_a)]_{jj'} \right\} = \eta_0 f_0 \left\{ [\hat{\mathbf{P}}^{Ia}(\lambda_a)]_{jj'}; \kappa \right\} + (1 - \eta_0) f_{\mathcal{E}} \left\{ [\hat{\mathbf{P}}^{Ia}(\lambda_a)]_{jj'} \right\},$$

with mixture weight $\eta_0 \in [0, 1]$, and where $f_0\{\cdot\}$ denotes the distribution of a null-edge while $f_{\mathcal{E}}\{\cdot\}$ denotes the distribution of a present edge. The former density can be found to be a scaled beta-density [22, 23, 7]:

$$f_0 \left\{ [\hat{\mathbf{P}}^{Ia}(\lambda_a)]_{jj'}; \kappa \right\} = \left| [\hat{\mathbf{P}}^{Ia}(\lambda_a)]_{jj'} \right| \mathcal{B} \left\{ [\hat{\mathbf{P}}^{Ia}(\lambda_a)]_{jj'}^2; \frac{1}{2}, \frac{\kappa - 1}{2} \right\},$$

with κ degrees of freedom (note that in the last expression $|\cdot|$ is used to denote the absolute value). In the $p > n$ situation κ has to be estimated, next to η_0 and $f_{\mathcal{E}}\{\cdot\}$. See [24] and [7] for details on obtaining estimates of these unknowns. Having these estimates at hand, the lfDR is given as [7]:

$$P \left(Y_j \neq Y_{j'} | [\hat{\mathbf{P}}^{Ia}(\lambda_a)]_{jj'} \right) = \frac{\hat{\eta}_0 f_0 \left\{ [\hat{\mathbf{P}}^{Ia}(\lambda_a)]_{jj'}; \hat{\kappa} \right\}}{\hat{\eta}_0 f_0 \left\{ [\hat{\mathbf{P}}^{Ia}(\lambda_a)]_{jj'}; \hat{\kappa} \right\} + (1 - \hat{\eta}_0) \hat{f}_{\mathcal{E}} \left\{ [\hat{\mathbf{P}}^{Ia}(\lambda_a)]_{jj'} \right\}},$$

conveying the empirical posterior probability that the edge between Y_j and $Y_{j'}$ is null given $[\hat{\mathbf{P}}^{Ia}(\lambda_a)]_{jj'}$. Another useful quantity is $1 - \text{IFDR}$, indicating the analogous probability that an edge is present. Again, similar probabilistic statements can be made with the alternative Type II estimator when replacing in the above $\hat{\mathbf{P}}^{Ia}(\lambda_a)$ by $\hat{\mathbf{P}}^{IIa}(\lambda_a)$. In Sections 5.4 and 5.5 an edge will be selected when $1 - \text{IFDR} \geq .99$.

While the two-step procedure of regularization followed by subsequent support determination does not have the appeal of simultaneous estimation and model selection, it does have the advantage that it enables probabilistic statements about the inclusion (or exclusion) of edges. An additional advantage is that the procedure may lead to a better representation of individual partial correlation or precision elements after sparsification: The closest, in a least-squares sense, p.d. sparsified representation of $\hat{\mathbf{\Omega}}^{Ia}(\lambda_a)$ (or $\hat{\mathbf{\Omega}}^{IIa}(\lambda_a)$), is indeed $\hat{\mathbf{\Omega}}^{Ia}(\lambda_a)$ (or $\hat{\mathbf{\Omega}}^{IIa}(\lambda_a)$) with the zero-structure imposed as follows from the lFDR test [cf. 25].

5.2. Data

The performance of the ridge and lasso precision estimators is evaluated on gene expression data of three pathways from five oncogenomics studies. The Bioconductor repository [26] offers five curated breast cancer data sets [27] generated on the same microarray platform (Affymetrix hgu 133 platform). These datasets will be indicated as follows: Mainz, Transbig, UNT, UPP, VDX. The data of these studies have been preprocessed in a uniform manner [see 28]. Cancer of the breast is a hormone-related cancer, with a central role for estrogen. Breast cancerous tissue may have many estrogen receptors (ER+ breast cancer) or few estrogen receptors (ER- breast cancer). The genomic pathways of ER+ and ER- breast cancers differ. Thus, to remove further heterogeneity among the data sets, they are limited to ER+ samples. The chosen pathways, p53, apoptosis, and mTOR, are defined by KEGG [29]. The p53 gene is a tumor suppressor gene. Cellular stress signals such as DNA damage can activate the p53-pathway, resulting in a multilayered tumor suppressive mechanism [30]. The genetic p53-pathway is defined to consist of those genes mediating the path from cellular stress signal to p53-induced tumor suppressive response. Alterations of the p53 pathway are found in most human cancers [31]. Apoptosis refers to the process of regulated cell death. The ability of cancerous cells to resist apoptosis is considered to be one of the hallmarks of human cancer [32]. The mTOR protein is a kinase (a phosphate transferring enzyme) that is frequently overexpressed in human cancers. This may lead to oncogenic signaling, making the cancerous cell self-sufficient in survival and multiplication [30], another hallmark of human cancer [32]. The underlying conditional dependency structure of the respective pathways is not fully known but is (generally) believed to be (relatively) sparse.

For each data set the probe sets that interrogate genes mapping to the p53, apoptosis, and mTOR pathways are selected. Whenever multiple probe sets map to the same gene, their expression levels have been averaged sample-wise over the instances. The resulting dimensions of the $n \times p$ pathway data sets are: $n = 162$ (Mainz), $n = 134$ (Transbig), $n = 86$ (UNT), $n = 213$ (UPP), $n = 209$ (VDX), and $p = 67$ (p53), $p = 83$ (apoptosis), $p = 47$ (mTOR). See Section 3 of the Supplement for R code on extracting the mentioned data.

The pathway data are not high-dimensional in the sense $p > n$. High-dimensionality is achieved by subsampling with sample sizes $n = 5, 10$ and 25 . One hundred subsamples are

drawn of each mentioned sample size for each combination of pathway and breast cancer data set. Optimal values of the penalty parameter for both versions of the alternative ridge and lasso estimators are obtained for each subsample by way of LOOCV. The ridge and lasso precision estimates for a subsample then correspond to these optimal penalty parameter values. Finally, the estimates are standardized to have unit diagonal (the standardized precision matrix is equal to the partial correlation matrix up to the sign of off-diagonal entries).

5.3. Loss Comparison

The standardized precision estimates are evaluated in terms of quadratic and Frobenius loss (as defined in Section 4.2). This requires the standardized population precision matrix, which is unknown. As a proxy we take the sample version obtained from the data with all samples, e.g., the standardized population precision matrix for the p53-pathway in the UPP data is defined as the (67×67) -dimensional standardized sample precision matrix over all $n = 213$ samples. The results of the loss evaluation are displayed in Figure 3 and Section 4 of the Supplementary Material.

Figure 3 and Section 4.1 of the Supplement show that the quadratic loss of the lasso estimate of the standardized precision matrix exceeds that of its ridge counterpart. In general, this is a consistent observation over the sample sizes ($n = 5, 10$ and 25), the pathways, and the data sets. This behavior also holds for the Frobenius loss and holds irrespective of the choice of target. In several cases the loss difference between the estimators decreases as n increases. This should not surprise, as the loss difference is expected to vanish for large n under fixed p (also note that, naturally, loss decreases with increasing n). Thus, the alternative ridge estimators of the standardized precision matrix yield a lower loss than the corresponding lasso estimators, in particular for the larger p/n ratios.

5.4. Sensitivity and Specificity

The evaluation of sensitivity and specificity of edge selection requires knowledge of the true conditional dependencies. Such knowledge is absent as the (causal) biological mechanisms underlying the pathway are mostly unknown (or at least uncertain). Hence, we resort to defining a ‘consensus truth’, comprised of those conditional dependencies that appear in the top $100\alpha\%$ of at least 4 out of the 5 breast cancer data sets by both methods (graphical lasso and alternative ridge paired with lFDR edge selection). The top $100\alpha\%$ constitutes of the $\lceil \frac{1}{2}p(p-1)\alpha \rceil$ edges with the largest selection frequency over the hundred respective subsamples (see Section 5.2), with $\alpha = \{0.005, 0.01, 0.015, \dots, 0.20\}$. This yields a nested sequence of ‘consensus truths’. The range of α corresponds to what is believed to be biologically plausible. Thus, with observed selected edges and the ‘consensus truths’ at hand, sensitivity and specificity are estimated per subsample over the range of α . The median sensitivity (specificity) over the hundred subsamples over all data sets is taken as the estimate of the sensitivity (specificity) for a particular combination of \mathbf{T} , n , α , and pathway. Figure 4 and Section 4.2 of the Supplementary Material visualize estimated sensitivity and specificity against α .

First note that the sensitivity is high for both the ridge estimator and the graphical lasso (in both Type I and Type II situations), due to the stringent definition of ‘consensus truth’. Furthermore, the Type I ridge estimator outperforms the corresponding lasso in

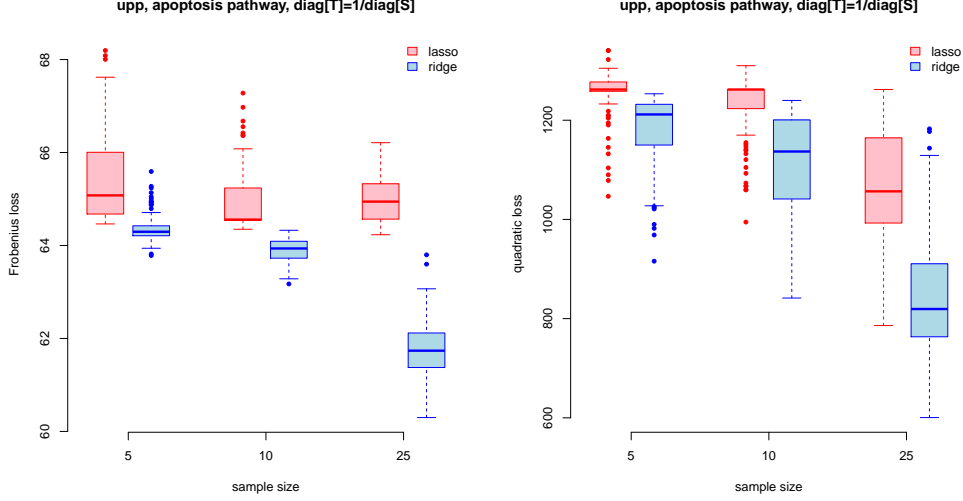


Figure 3: Loss comparison between the Type I alternative ridge estimator with $\text{diag}[\mathbf{T}] = 1/\text{diag}[\mathbf{S}]$ and the corresponding graphical lasso estimator on the UPP apoptosis-pathway data. The left-hand panel depicts Frobenius loss while the right-hand panel depicts quadratic loss.

terms of sensitivity. For the Type II setting it is seen that the graphical lasso fares slightly better with regard to sensitivity. These behaviors are reversed when evaluating specificity: The lasso fares better in the Type I setting, while the ridge outperforms the lasso in the Type II situation. These observations hold for all pathways. These findings can (at least in part) be traced to the utilization of lFDR edge selection on the ridge estimators (cf. Section 5.1). Using $\mathbf{T} = \mathbf{0}$ will (tend to), by enforcing more uniformity among the partial correlation values, emphasize the null-edge distribution, leading to improved specificity and (somewhat) diminished sensitivity. A p.d. target \mathbf{T} , on the other hand, will tend to preserve data signal, and will subsequently lead to improved sensitivity and (somewhat) diminished specificity. These behaviors might suggest the following (also taking into account the loss behavior and the stability of performance over respective sample sizes): Give preference to the Type I alternative ridge estimator when emphasizing the true positive rate, and give preference to the Type II alternative ridge estimator when emphasizing the true negative rate.

5.5. Stability

The performance of the (Type I and II) ridge and lasso precision estimators can also be evaluated in terms of network stability. Define an edge stable when it is selected in the union of the top $100\alpha\%$ over the respective subsample sizes $n = 5, 10$, and 25 . When plotting the number of stable edges against α (see Figure 5 and Section 4.3 of the Supplement), it is clear that the number of stable edges shows a faster increase, with increasing α , for the ridge estimators than for the graphical lasso. This effect is especially pronounced for the Type I ridge setting. The ridge estimators also sort more stable behavior over the respective data sets.

Analogous behavior can be shown with regard to the effect of sample size. Figure 6 contains conditional independence graphs for the Type I alternative ridge estimator with $\text{diag}[\mathbf{T}] = 1/\text{diag}[\mathbf{S}]$ and the corresponding graphical lasso on the UPP apoptosis-pathway data. A represented edge means that it was selected at least 50 times over the 100 subsamples. It may be observed that the pairing of the alternative ridge estimator with IFDR support determination selects more stable (in terms of network-structure change) networks over the respective sample sizes. While usage of IFDR edge selection on the ridge regularized precision matrix tends to gain in conservativeness with growing

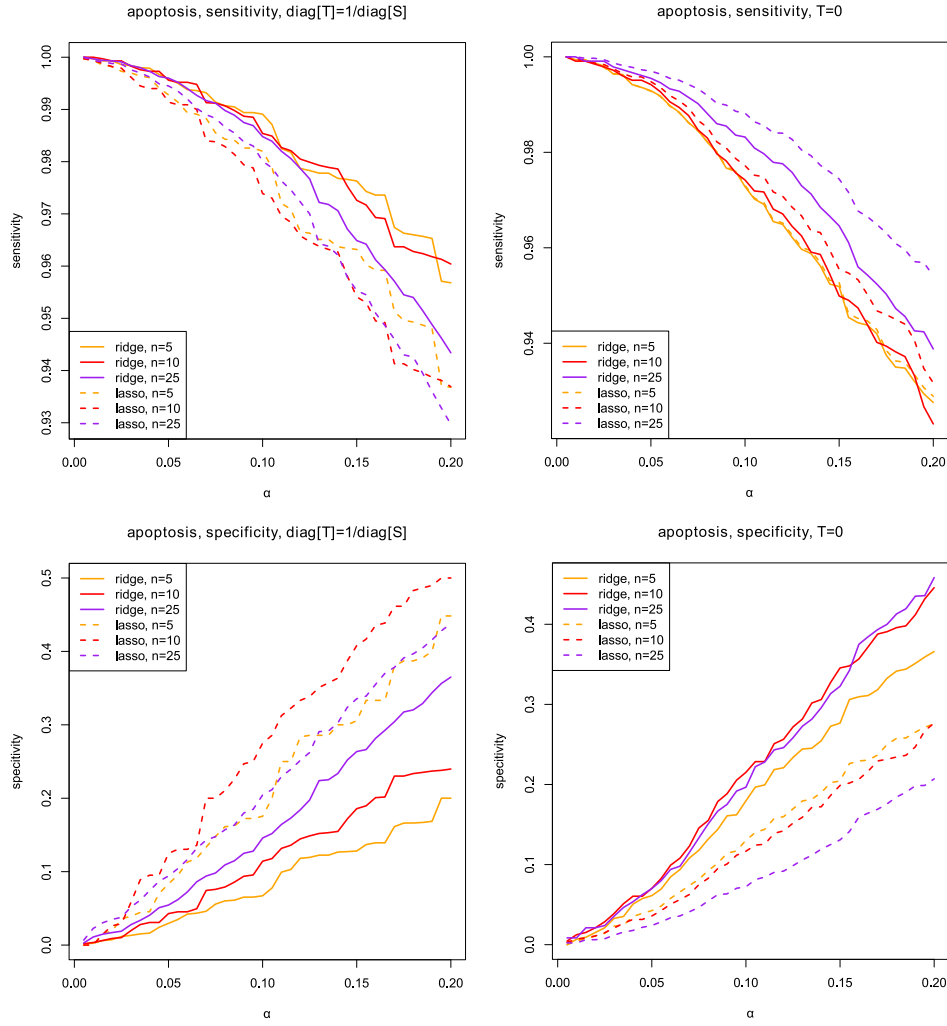


Figure 4: Sensitivity and specificity comparison between the alternative ridge and graphical lasso estimators on the apoptosis-pathway data. The upper panels depict sensitivity results while the lower panels depict specificity results. The left-hand panels depict results for $\text{diag}[\mathbf{T}] = 1/\text{diag}[\mathbf{S}]$ while the right-hand panels depict results for $\mathbf{T} = \mathbf{0}$.

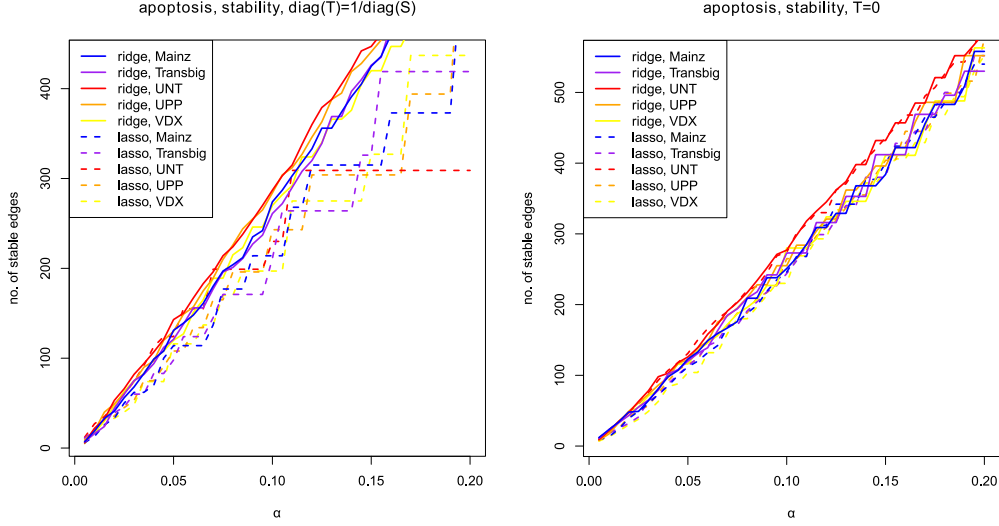


Figure 5: The number of stable edges plotted against α for the apoptosis-pathway data. The left-hand panel depicts results for $\text{diag}[\mathbf{T}] = 1/\text{diag}[\mathbf{S}]$ while the right-hand panel depicts results for $\mathbf{T} = \mathbf{0}$.

n , the network-structure changes over the respective sample sizes are much less dramatic vis-à-vis the graphical lasso. This picture of stability holds for the remaining pathways under $\text{diag}[\mathbf{T}] = 1/\text{diag}[\mathbf{S}]$. For the Type II comparison the alternative ridge estimator tends to be more conservative than the graphical lasso for the higher sample sizes (cf. explanation Section 5.4). This can again be taken as an indication that when the network data at disposal do contain a sizeable signal, it is preferable to choose a non-null target \mathbf{T} for better signal preservation.

5.6. The Graphical Lasso as Reference

One may argue that the comparability may be obscured when the number of selected edges differs considerably between methods. We thus, in addition to the exercises above, take interest in comparing the alternative ridge estimators with the graphical lasso when the latter dictates the number of edges the former may select. Say the graphical lasso selects, within a certain subsample, τ edges; then for the corresponding ridge precision estimator the τ edges are selected with the largest absolute partial correlation. It is obvious that thresholding the ridge precision estimator on the basis of the graphical lasso will favor the latter. In this setup the ridge estimators thus prove their strength through non-inferiority. The results in Section 5 of the Supplementary Material show that, indeed, the alternative ridge estimators display non-inferiority with respect to sensitivity, specificity, and stability in this situation.

Summarizing on the basis of the results in Sections 5.3–5.6: The alternative Type I (Type II) ridge estimator paired with post-hoc edge selection is a contender in a graphical modeling setting, especially when the p/n ratio tends to get more extreme and/or when emphasis is placed on the true positive (negative) rate.

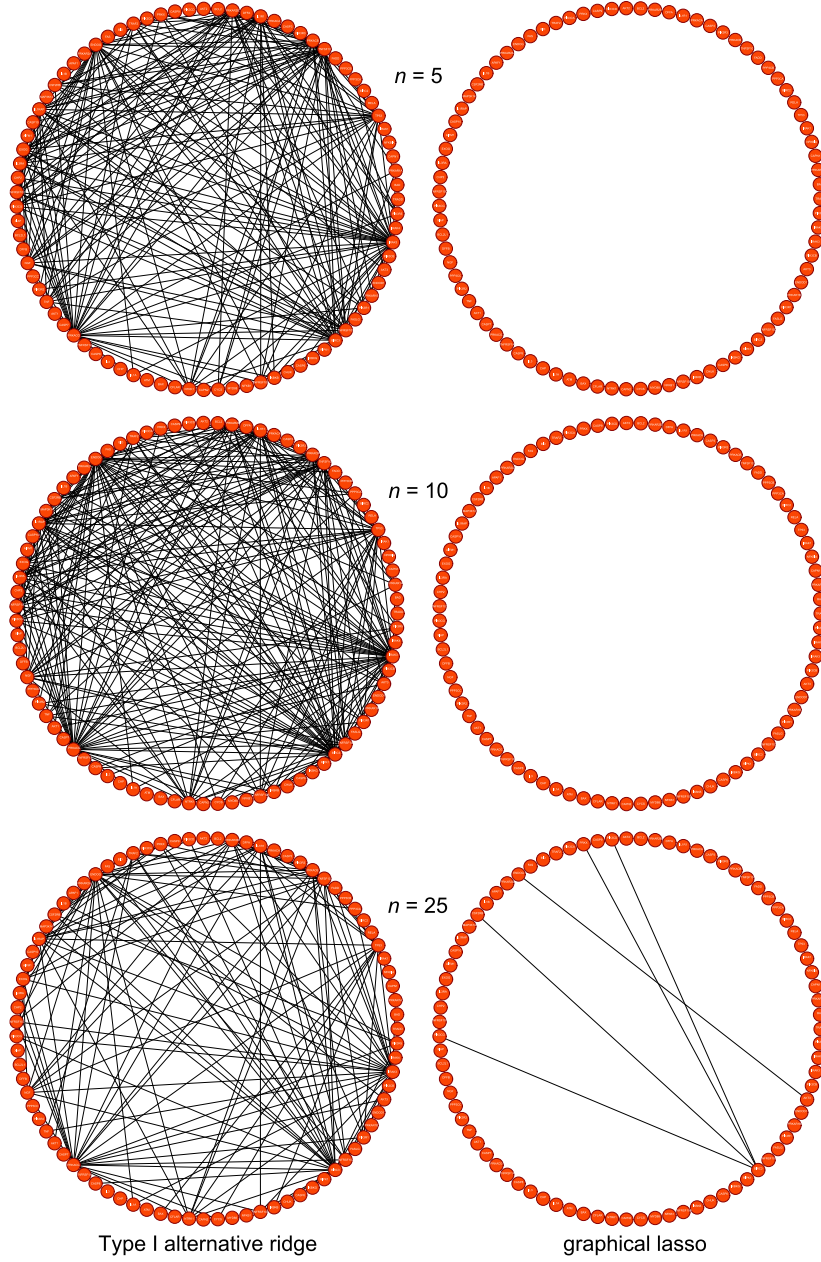


Figure 6: Conditional independence graphs for the Type I alternative ridge estimator using IFDR edge selection (left-hand figures) and the corresponding graphical lasso (right-hand figures) on the UPP apoptosis-pathway data. For an edge to be represented in the conditional independence graphs above, it must have been selected at least 50 times over the 100 replications (given sample size $n = 5, 10$ and 25 , respectively).

6. Discussion

We studied ridge estimation of the precision matrix. Estimators currently in use can be roughly divided into two archetypes whose penalties do not coincide with the common ridge penalty. Starting from the common ridge penalty we derived an analytic expression of the ridge estimator of the inverse covariance matrix, on the basis of which alternatives were formulated for the two archetypes. The alternative estimators were shown to outperform the archetypes in terms of risk. An illustration using pathway data also showed that the alternative ridge estimators perform better than the corresponding graphical lasso estimators in terms of loss. They also tend to select more stable networks, especially in situations where the variable to sample ratio is more extreme. The provided expressions can also be of use in the study of theoretical properties of penalized inverse covariance estimators.

The proposed estimators can facilitate methods and approaches of data analysis leaning on the estimation of precision (or covariance) matrices in high-dimensional situations. For example, the estimators may be used in supporting covariance regularized regression [10], discriminant analysis, or canonical correlation analysis. In addition, in the context of graphical modeling, the proposed estimators can be paired with post-hoc methods for determining the support of the precision matrix, such as IFDR multiple testing [7]. Furthermore, regularized (inverse) covariance matrices stemming from the proposed estimators can be used as input in covariance structure modeling efforts [33] (including factor analysis and structural equation modeling as special cases), when p is large relative to n .

We see various inroads for further research. One would be to study the proposed estimators from a Bayesian perspective. In addition, the Type I estimator may lend itself for a natural framework of Bayesian updating regarding graphical modeling, where the target is determined by previous rounds of fitting the estimator followed by subsequent support determination. Another option would be to extend the proposed estimators with a condition number constraint [34], so that it can be formalized which values for the penalty parameter can be considered ‘too small’. Also, the results from the numerical studies may be further supported with results on increasing-dimension asymptotics of the proposed estimators. From a more applied perspective it may be deemed interesting to compare multiple post-hoc methods for determining the support of the precision matrix. These issues are the focal points of current research.

The ridge estimators employed in this paper are implemented in the R-package `rags2ridges` [35] along with supporting functions to employ these estimators in a graphical modeling setting. The package is freely available from the Comprehensive R Archive Network (<http://cran.r-project.org/>) [36].

Acknowledgements

The research leading to these results has received funding from the European Community’s Seventh Framework Programme (FP7, 2007-2013), Research Infrastructures action, under the grant agreement No. FP7-269553 (EpiRadBio project). The authors would also like to thank Mark van de Wiel, Poul Svante Eriksen, and Grégory Nuel, whose constructive comments have led to an improvement in presentation.

Appendix A. Proofs

This appendix contains proofs for Lemma 1, Propositions 1, 2, 3 and 4, as well as Corollaries 1 and 2. Consider first the following theorem:

Theorem 1. [37, p. 115] *Let \mathbf{H} be a p.d. Hermitian matrix. There exists a unique p.d. Hermitian matrix h such that $h^2 = \mathbf{H}$. If \mathbf{H} is real-valued, then so is h . The matrix h is called the square root of \mathbf{H} , and is denoted by $h = \mathbf{H}^{1/2}$.*

This theorem is of use in the proof of Lemma 1:

PROOF OF LEMMA 1. Define the penalized log-likelihood:

$$\mathcal{L}^p(\mathbf{\Omega}; \mathbf{S}, \mathbf{T}, \lambda_a) \propto \ln |\mathbf{\Omega}| - \text{tr}(\mathbf{S}\mathbf{\Omega}) - \frac{\lambda_a}{2} \text{tr}[(\mathbf{\Omega} - \mathbf{T})^T(\mathbf{\Omega} - \mathbf{T})].$$

Now, take the derivative of $\mathcal{L}^p(\mathbf{\Omega}; \mathbf{S}, \mathbf{T}, \lambda_a)$ w.r.t. $\mathbf{\Omega}$:

$$\begin{aligned} \frac{\partial \mathcal{L}^p(\mathbf{\Omega}; \mathbf{S}, \mathbf{T}, \lambda_a)}{\partial \mathbf{\Omega}} &= 2 [\mathbf{\Omega}^{-1} - (\mathbf{S} - \lambda_a \mathbf{T}) - \lambda_a \mathbf{\Omega}] - [\mathbf{\Omega}^{-1} - (\mathbf{S} - \lambda_a \mathbf{T}) - \lambda_a \mathbf{\Omega}] \circ \mathbf{I}_p \\ &= [\mathbf{\Omega}^{-1} - (\mathbf{S} - \lambda_a \mathbf{T}) - \lambda_a \mathbf{\Omega}] \circ (2\mathbf{J}_p - \mathbf{I}_p), \end{aligned} \quad (\text{A.1})$$

where \mathbf{J}_p denotes the all-ones matrix. It is immediate that (A.1) is $\mathbf{0}$ only when

$$\mathbf{\Omega}^{-1} - (\mathbf{S} - \lambda_a \mathbf{T}) - \lambda_a \mathbf{\Omega} = \mathbf{0}. \quad (\text{A.2})$$

We will approach the problem in (A.2) from a square-completion angle.

Post-multiply (A.2) by $\mathbf{\Omega}^{-1}$. Subsequently adding $\frac{1}{4}(\mathbf{S} - \lambda_a \mathbf{T})^2$ to both sides of the equality sign gives that $\mathbf{\Omega}$ must satisfy:

$$\lambda_a \mathbf{I}_p + \frac{1}{4}(\mathbf{S} - \lambda_a \mathbf{T})^2 := \mathbf{\Omega}^{-2} - (\mathbf{S} - \lambda_a \mathbf{T})\mathbf{\Omega}^{-1} + \frac{1}{4}(\mathbf{S} - \lambda_a \mathbf{T})^2. \quad (\text{A.3})$$

Notice that under pre-multiplication by $\mathbf{\Omega}^{-1}$ the matrix $\mathbf{\Omega}$ is also implied to satisfy:

$$\lambda_a \mathbf{I}_p + \frac{1}{4}(\mathbf{S} - \lambda_a \mathbf{T})^2 := \mathbf{\Omega}^{-2} - \mathbf{\Omega}^{-1}(\mathbf{S} - \lambda_a \mathbf{T}) + \frac{1}{4}(\mathbf{S} - \lambda_a \mathbf{T})^2. \quad (\text{A.4})$$

Adding (A.3) and (A.4) and subsequently dividing by 2 thus yields:

$$\lambda_a \mathbf{I}_p + \frac{1}{4}(\mathbf{S} - \lambda_a \mathbf{T})^2 = \mathbf{\Omega}^{-2} - \frac{1}{2}\mathbf{\Omega}^{-1}(\mathbf{S} - \lambda_a \mathbf{T}) - \frac{1}{2}(\mathbf{S} - \lambda_a \mathbf{T})\mathbf{\Omega}^{-1} + \frac{1}{4}(\mathbf{S} - \lambda_a \mathbf{T})^2.$$

Now, complete the square to obtain:

$$\lambda_a \mathbf{I}_p + \frac{1}{4}(\mathbf{S} - \lambda_a \mathbf{T})^2 = \left[\mathbf{\Omega}^{-1} - \frac{1}{2}(\mathbf{S} - \lambda_a \mathbf{T}) \right]^2. \quad (\text{A.5})$$

The left-hand side of (A.5) is p.d., which implies that the right-hand side is p.d. By Theorem 1, both sides then have a unique square root that is p.d. and symmetric. Taking this square root on both sides results in:

$$\left[\lambda_a \mathbf{I}_p + \frac{1}{4}(\mathbf{S} - \lambda_a \mathbf{T})^2 \right]^{1/2} = \mathbf{\Omega}^{-1} - \frac{1}{2}(\mathbf{S} - \lambda_a \mathbf{T}).$$

Finally, solving for $\mathbf{\Omega}$ gives the desired expression (5). □

PROOF OF PROPOSITION 1.

(i) Let $d(\cdot)_{jj}$ denote the j 'th eigenvalue of the matrix term in brackets (\cdot) . Then

$$d\left\{[\hat{\mathbf{\Omega}}^{Ia}(\lambda_a)]^{-1}\right\}_{jj} = d\left[\frac{1}{2}(\mathbf{S} - \lambda_a \mathbf{T})\right]_{jj} + \sqrt{\left\{d\left[\frac{1}{2}(\mathbf{S} - \lambda_a \mathbf{T})\right]_{jj}\right\}^2 + \lambda_a} > 0,$$

when $\lambda_a > 0$. Hence, $\hat{\mathbf{\Omega}}^{Ia}(\lambda_a)$ is p.d. for any $\lambda_a \in (0, \infty)$.

(ii) The right-hand limit is immediate as:

$$\hat{\mathbf{\Omega}}^{Ia}(0) = \left\{\left[0\mathbf{I}_p + \frac{1}{4}(\mathbf{S} - 0\mathbf{T})^2\right]^{1/2} + \frac{1}{2}(\mathbf{S} - 0\mathbf{T})\right\}^{-1} = \mathbf{S}^{-1}.$$

(iii) For the left-hand limit we note that, when λ_a approaches ∞ ,

$$[\hat{\mathbf{\Omega}}^{Ia}(\lambda_a)]^{-1} = \left[\lambda_a \mathbf{I}_p + \frac{1}{4}(\mathbf{S} - \lambda_a \mathbf{T})^2\right]^{1/2} + \frac{1}{2}(\mathbf{S} - \lambda_a \mathbf{T}) \longrightarrow \mathbf{T}^{-1},$$

must hold for the property to hold. We will first embark on rewriting this implied convergence behavior to a standard form. Note that we can rewrite such that, equivalently,

$$\begin{aligned} & \mathbf{T}^{1/2} \left\{ \left[\lambda_a \mathbf{I}_p + \frac{1}{4}(\mathbf{S} - \lambda_a \mathbf{T})^2 \right]^{1/2} + \frac{1}{2}(\mathbf{S} - \lambda_a \mathbf{T}) \right\} \mathbf{T}^{1/2} \\ &= \left[\lambda_a \mathbf{T}^2 + \frac{1}{4}(\tilde{\mathbf{S}} - \lambda_a \mathbf{T}^2)^2 \right]^{1/2} + \frac{1}{2}(\tilde{\mathbf{S}} - \lambda_a \mathbf{T}^2) \longrightarrow \mathbf{I}_p, \end{aligned}$$

where $\tilde{\mathbf{S}} = \mathbf{T}^{1/2} \mathbf{S} \mathbf{T}^{1/2}$, must hold for the property to hold. Note that the term $[\lambda_a \mathbf{T}^2 + \frac{1}{4}(\tilde{\mathbf{S}} - \lambda_a \mathbf{T}^2)^2]^{1/2} + \frac{1}{2}(\tilde{\mathbf{S}} - \lambda_a \mathbf{T}^2)$ can be rewritten as:

$$\left[\frac{1}{4} \left(\lambda_a \mathbf{T}^2 + 2\mathbf{I}_p - \tilde{\mathbf{S}} \right)^2 + \left(\tilde{\mathbf{S}} - \mathbf{I}_p \right)^2 \right]^{1/2} - \frac{1}{2} \left(\lambda_a \mathbf{T}^2 + 2\mathbf{I}_p - \tilde{\mathbf{S}} \right) + \mathbf{I}_p,$$

implying that the problem can be reduced to proving

$$\lim_{\lambda_a \rightarrow \infty^-} \left[\mathbf{B}^2(\lambda_a) + \left(\tilde{\mathbf{S}} - \mathbf{I}_p \right)^2 \right]^{1/2} - \mathbf{B}(\lambda_a) = \mathbf{0}, \quad (\text{A.6})$$

where $\mathbf{B}(\lambda_a) = \frac{1}{2} \left(\lambda_a \mathbf{T}^2 + 2\mathbf{I}_p - \tilde{\mathbf{S}} \right)$.

To prove this invoke Weyl's eigenvalue inequality [38]. Let \mathbf{A} , \mathbf{B} , $\mathbf{C} = \mathbf{A} + \mathbf{B}$ be real, symmetric $p \times p$ matrices with eigenvalues $\alpha_1 \geq \alpha_2 \geq \dots \geq \alpha_p$, $\beta_1 \geq \beta_2 \geq \dots \geq \beta_p$, and $\gamma_1 \geq \gamma_2 \geq \dots \geq \gamma_p$, respectively. Weyl's result then states:

$$\alpha_j + \beta_p \leq \gamma_j \leq \alpha_j + \beta_1 \quad \text{for all } j.$$

Applying this inequality to $\mathbf{C}(\lambda) = \lambda \mathbf{A} + \mathbf{B}$ with $\lambda > 0$ (where λ is used generically) and \mathbf{A} and \mathbf{B} as before, we obtain:

$$\lambda \alpha_j + \beta_p \leq \gamma_j(\lambda) \leq \lambda \alpha_j + \beta_1 \quad \text{for all } j.$$

Divide by λ and let λ tend to infinity (from the left), which is immediately seen to imply:

$$\lim_{\lambda \rightarrow \infty^-} \frac{1}{\lambda} \gamma_j(\lambda) = \alpha_j \quad \text{for all } j.$$

Put differently, the eigenvalues of $\mathbf{C}(\lambda)$ tend to those of $\lambda \mathbf{A}$. Application of Weyl's eigenvalue inequality and the consequence derived above thus warrant that

$$\lim_{\lambda_a \rightarrow \infty^-} \left[\mathbf{B}^2(\lambda_a) + \left(\tilde{\mathbf{S}} - \mathbf{I}_p \right) \right]^{1/2} - \mathbf{B}(\lambda_a) = \mathbf{0},$$

as indeed needed to be proven. \square

PROOF OF COROLLARY 1. We first need to show that (6) stems properly as the unique maximizer of the log-likelihood (1) amended with (4) under $\mathbf{T} = \mathbf{0}$. We thus define the penalized log-likelihood:

$$\mathcal{L}^p(\mathbf{\Omega}; \mathbf{S}, \lambda_a) \propto \ln |\mathbf{\Omega}| - \text{tr}(\mathbf{S}\mathbf{\Omega}) - \frac{\lambda_a}{2} \text{tr}(\mathbf{\Omega}^T \mathbf{\Omega}).$$

Taking the derivative of $\mathcal{L}^p(\mathbf{\Omega}; \mathbf{S}, \lambda_a)$ w.r.t. $\mathbf{\Omega}$ gives:

$$\frac{\partial \mathcal{L}^p(\mathbf{\Omega}; \mathbf{S}, \lambda_a)}{\partial \mathbf{\Omega}} = (\mathbf{\Omega}^{-1} - \mathbf{S} - \lambda_a \mathbf{\Omega}) \circ (2\mathbf{J}_p - \mathbf{I}_p),$$

which is $\mathbf{0}$ only when

$$\mathbf{\Omega}^{-1} - \mathbf{S} - \lambda_a \mathbf{\Omega} = \mathbf{0}.$$

A strategy analogous to the one used in the proof of Lemma 1 will give the desired expression (6). With regard to the properties of this estimator:

(i) Let $d(\cdot)_{jj}$ denote the j 'th eigenvalue of the matrix term in brackets (\cdot) . Notice

$$d \left\{ [\hat{\mathbf{\Omega}}^{\text{II}a}(\lambda_a)]^{-1} \right\}_{jj} = d \left(\frac{1}{2} \mathbf{S} \right)_{jj} + \sqrt{\left[d \left(\frac{1}{2} \mathbf{S} \right)_{jj} \right]^2 + \lambda_a} > 0,$$

when $\lambda_a > 0$, implying $\hat{\mathbf{\Omega}}^{\text{II}a}(\lambda_a)$ is p.d. for any $\lambda_a \in (0, \infty)$.

(ii) The right-hand limit is immediate as:

$$\hat{\mathbf{\Omega}}^{\text{II}a}(0) = \left\{ \left[0\mathbf{I}_p + \frac{1}{4}\mathbf{S}^2 \right]^{1/2} + \frac{1}{2}\mathbf{S} \right\}^{-1} = \mathbf{S}^{-1}.$$

(iii) For the left-hand limit we note that as λ_a approaches ∞ ,

$$[\hat{\mathbf{\Omega}}^{\text{II}a}(\lambda_a)]^{-1} = \left[\lambda_a \mathbf{I}_p + \frac{1}{4}\mathbf{S}^2 \right]^{1/2} + \frac{1}{2}\mathbf{S}$$

becomes a diagonally dominant matrix with near infinite diagonal values. The inverse of which must necessarily approach the null-matrix. \square

PROOF OF PROPOSITION 2. First, note:

$$\begin{aligned}\mathbb{E} \left(\|\hat{\Sigma}_n^{Ia}(\lambda_{a,n}) - \Sigma\|_F^2 \right) &= \mathbb{E} \left\{ \text{tr} \left[\left(\hat{\Sigma}_n^{Ia}(\lambda_{a,n}) - \Sigma \right)^T \left(\hat{\Sigma}_n^{Ia}(\lambda_{a,n}) - \Sigma \right) \right] \right\} \\ &= \text{tr} \left\{ \text{Var} \left[\hat{\Sigma}_n^{Ia}(\lambda_{a,n}) \right] \right\} \\ &\quad - \text{tr} \left\{ \Sigma \mathbb{E} \left[\hat{\Sigma}_n^{Ia}(\lambda_{a,n}) - \Sigma \right] \right\} \text{tr} \left\{ \mathbb{E} \left[\hat{\Sigma}_n^{Ia}(\lambda_{a,n}) - \Sigma \right] \Sigma \right\}.\end{aligned}$$

By virtue of Lemma 2 and the continuity of the trace the latter two terms vanish as $n \rightarrow \infty^-$. It remains to be shown that the first term converges to zero. To this end note that the almost sure convergence of $\lambda_{a,n}$ to zero implies $\lim_{n \rightarrow \infty^-} P[\lambda_{a,n} d(\mathbf{T})_{11} < d(\mathbf{S}_n)_{pp}] = 1$ and $\mathbf{S}_n - \lambda_{a,n} \mathbf{T} \succcurlyeq \mathbf{0}$ with probability 1 as n tends to infinity. Thus, in the limit:

$$\begin{aligned}\hat{\Sigma}_n^{Ia}(\lambda_{a,n}) &= \left[\lambda_{a,n} \mathbf{I}_p + \frac{1}{4} (\mathbf{S}_n - \lambda_{a,n} \mathbf{T})^2 \right]^{1/2} + \frac{1}{2} (\mathbf{S}_n - \lambda_{a,n} \mathbf{T}) \\ &\preccurlyeq \left\{ \left[\sqrt{\lambda_{a,n}} \mathbf{I}_p + \frac{1}{2} (\mathbf{S}_n - \lambda_{a,n} \mathbf{T}) \right]^2 \right\}^{1/2} + \frac{1}{2} (\mathbf{S}_n - \lambda_{a,n} \mathbf{T}) \\ &= \mathbf{S}_n + \sqrt{\lambda_{a,n}} \mathbf{I}_p - \lambda_{a,n} \mathbf{T}.\end{aligned}$$

From this it follows that

$$\text{tr} \left[\left(\mathbf{S}_n + \sqrt{\lambda_{a,n}} \mathbf{I}_p - \lambda_{a,n} \mathbf{T} \right)^2 \right] \succcurlyeq \text{tr} \left\{ \left[\hat{\Sigma}_n^{Ia}(\lambda_{a,n}) \right]^2 \right\}$$

as $n \rightarrow \infty^-$, which in turn gives:

$$\text{tr} \left[\text{Var} \left(\mathbf{S}_n + \sqrt{\lambda_{a,n}} \mathbf{I}_p - \lambda_{a,n} \mathbf{T} \right) \right] \succcurlyeq \text{tr} \left\{ \text{Var} \left[\hat{\Sigma}_n^{Ia}(\lambda_{a,n}) \right] \right\} \quad \text{as } n \rightarrow \infty^-.$$

The assumptions on $\lambda_{a,n}$, \mathbf{S}_n and their covariance imply that the left-hand side tends to zero. Finally, the dominated convergence theorem warrants that the right-hand side too converges to zero as $n \rightarrow \infty^-$. \square

The proof of Proposition 3 will be based on the target $\mathbf{T} = \mathbf{I}_p$. The extension to a general p.d. target scalar matrix is straightforward as it is a direct consequence, but notationally slightly more cumbersome.

PROOF OF PROPOSITION 3. Note that $\hat{\Omega}^I(\lambda_I)$ can be decomposed as:

$$[\hat{\Omega}^I(\lambda_I)]^{-1} = \mathbf{V}[(1 - \lambda_I)\mathbf{D} + \lambda_I \mathbf{I}_p] \mathbf{V}^T.$$

Juxtaposing this expression with (10) while writing $d_{jj} = (\mathbf{D})_{jj}$, we are after establishing

$$\sqrt{\lambda_a + \frac{1}{4} (d_{jj} - \lambda_a)^2} + \frac{1}{2} (d_{jj} - \lambda_a) \stackrel{?}{>=<} \frac{1}{1 + \lambda_a} d_{jj} + \frac{\lambda_a}{1 + \lambda_a},$$

which after some ready algebra can be rewritten as:

$$\sqrt{\varphi_{jj}(\lambda_a)^2 + d_{jj} - 1} \stackrel{?}{>=<} \frac{1}{1 + \lambda_a} (d_{jj} - 1) + \varphi_{jj}(\lambda_a),$$

with $\varphi_{jj}(\lambda_a) = \frac{1}{2}\lambda_a - \frac{1}{2}d_{jj} + 1$. Squaring both sides and simplifying the problem becomes:

$$d_{jj} - 1 \stackrel{?}{=} \frac{1}{(1 + \lambda_a)^2} (d_{jj} - 1)^2 + d_{jj} - 1 - \frac{1}{1 + \lambda_a} (d_{jj} - 1)^2,$$

which reduces to establishing the sign of:

$$\frac{(d_{jj} - 1)^2}{(1 + \lambda_a)^2} - \frac{(d_{jj} - 1)^2}{1 + \lambda_a}.$$

The solution to which is readily found to be:

$$0 \geq \frac{(d_{jj} - 1)^2}{(1 + \lambda_a)^2} - \frac{(d_{jj} - 1)^2}{1 + \lambda_a} = -\frac{\lambda_a (d_{jj} - 1)^2}{(1 + \lambda_a)}.$$

Consequently, the alternative estimator $\hat{\boldsymbol{\Omega}}^{\text{Ia}}(\lambda_a)$ displays shrinkage of the eigenvalues of \mathbf{S}^{-1} that is at least as heavy as the shrinkage propagated by the archetypal estimator $\hat{\boldsymbol{\Omega}}^{\text{I}}(\lambda_{\text{I}})$. \square

PROOF OF PROPOSITION 4. Note that the decomposition of the original ridge estimator of the second type is

$$[\hat{\boldsymbol{\Omega}}^{\text{II}}(\lambda_{\text{II}})]^{-1} = \mathbf{V}(\lambda_{\text{II}}\mathbf{I}_p + \mathbf{D})\mathbf{V}^{\text{T}}.$$

Then, when writing $d_{jj} = (\mathbf{D})_{jj}$ while juxtaposing the above expression with (11), we have:

$$\lambda_{\text{II}} + d_{jj} \geq \sqrt{\lambda_{\text{II}}^2 + \frac{1}{4}d_{jj}^2} + \frac{1}{2}d_{jj},$$

as follows directly from $(\lambda_{\text{II}} + \frac{1}{2}d_{jj})^2 \geq \lambda_{\text{II}}^2 + \frac{1}{4}d_{jj}^2$. This indicates the archetypal estimator $\hat{\boldsymbol{\Omega}}^{\text{II}}(\lambda_{\text{II}})$ displaying shrinkage of the eigenvalues of \mathbf{S}^{-1} that is at least as heavy as the shrinkage propagated by the alternative estimator $\hat{\boldsymbol{\Omega}}^{\text{IIa}}(\lambda_a)$. \square

PROOF OF COROLLARY 2. Note:

$$\mathcal{L}[\hat{\boldsymbol{\Omega}}^{\text{II}}(\lambda_{\text{II}}); \mathbf{S}] \propto \ln |\hat{\boldsymbol{\Omega}}^{\text{II}}(\lambda_{\text{II}})| - \text{tr}[\mathbf{S}\hat{\boldsymbol{\Omega}}^{\text{II}}(\lambda_{\text{II}})] \propto -\sum_{j=1}^p \ln(\lambda_{\text{II}} + d_{jj}) - \sum_{j=1}^p \frac{d_{jj}}{\lambda_{\text{II}} + d_{jj}}.$$

Similarly:

$$\mathcal{L}[\hat{\boldsymbol{\Omega}}^{\text{IIa}}(\lambda_a); \mathbf{S}] \propto -\sum_{j=1}^p \ln[\gamma_{jj}(\lambda_{\text{II}})] - \sum_{j=1}^p \frac{d_{jj}}{\gamma_{jj}(\lambda_{\text{II}})},$$

where $\gamma_{jj}(\lambda_{\text{II}}) = \sqrt{\lambda_{\text{II}}^2 + \frac{1}{4}d_{jj}^2} + \frac{1}{2}d_{jj}$. It then suffices to show that

$$\ln(\lambda_{\text{II}} + d_{jj}) - \ln[\gamma_{jj}(\lambda_{\text{II}})] + \frac{d_{jj}}{\lambda_{\text{II}} + d_{jj}} - \frac{d_{jj}}{\gamma_{jj}(\lambda_{\text{II}})} \geq 0.$$

Using $\ln(1+x) \geq x/(1+x)$ and $d_{jj} + \lambda_{\Pi} \geq \gamma_{jj}(\lambda_{\Pi}) \geq d_{jj}$ (Proposition 4), the manipulations below prove this:

$$\begin{aligned} \ln\left(\frac{\lambda_{\Pi} + d_{jj}}{\gamma_{jj}(\lambda_{\Pi})}\right) + \frac{d_{jj}}{\lambda_{\Pi} + d_{jj}} - \frac{d_{jj}}{\gamma_{jj}(\lambda_{\Pi})} &\geq \\ \frac{\lambda_{\Pi} + d_{jj} - \gamma_{jj}(\lambda_{\Pi})}{\lambda_{\Pi} + d_{jj}} + \frac{d_{jj}}{\lambda_{\Pi} + d_{jj}} - \frac{d_{jj}}{\gamma_{jj}(\lambda_{\Pi})} &= \\ \frac{[\gamma_{jj}(\lambda_{\Pi}) - d_{jj}][d_{jj} + \lambda_{\Pi} - \gamma_{jj}(\lambda_{\Pi})]}{\gamma_{jj}(\lambda_{\Pi})(\lambda_{\Pi} + d_{jj})} &\geq 0. \end{aligned}$$

□

References

References

- [1] M. Yuan, Y. Lin, Model selection and estimation in the Gaussian graphical model, *Biometrika* 94 (2007) 19–35.
- [2] O. Banerjee, L. El Ghaoui, A. d’Aspremont, Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data, *Journal of Machine Learning Research* 9 (2008) 485–516.
- [3] J. Friedman, T. Hastie, R. Tibshirani, Sparse inverse covariance estimation with the graphical lasso, *Biostatistics* 9 (2008) 432–441.
- [4] M. Yuan, Efficient computation of ℓ_1 regularized estimates in Gaussian graphical models, *Journal of Computational and Graphical Statistics* 17 (2008) 809–826.
- [5] W. J. Fu, Penalized regressions: The bridge versus the lasso, *Journal of Computational and Graphical Statistics* 7 (1998) 397–416.
- [6] O. Ledoit, M. Wolf, A well-conditioned estimator for large-dimensional covariance matrices, *Journal of Multivariate Analysis* 88 (2004) 365–411.
- [7] J. Schäfer, K. Strimmer, A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics, *Statistical Applications in Genetics and Molecular Biology* 4 (2005) art. 32.
- [8] A. E. Hoerl, R. Kennard, Ridge regression: Biased estimation for nonorthogonal problems, *Technometrics* 12 (1970) 55–67.
- [9] D. Warton, Penalized normal likelihood and ridge regularization of correlation and covariance matrices, *Journal of the American Statistical Association* 103 (2008) 340–349.
- [10] D. M. Witten, R. Tibshirani, Covariance-regularized regression and classification for high-dimensional problems, *Journal of the Royal Statistical Society, Series B* 71 (2009) 615–636.
- [11] G. Letac, H. Massam, All invariant moments of the Wishart distribution, *Scandinavian Journal of Statistics* 31 (2004) 295–318.
- [12] A. W. van der Vaart, *Asymptotic Statistics*, Cambridge University Press, Cambridge, 1998.
- [13] A. Abbruzzo, I. Vujačić, E. Wit, A. M. Mineo, Generalized information criterion for model selection in penalized graphical models, [arXiv:1403.1249v1 \[stat.ME\]](#) (2014).
- [14] H. Akaike, Information theory and an extension of the maximum likelihood principle, in: B. N. Petrov, F. Csaki (Eds.), *Second International Symposium on Information Theory*, Akademiai Kiado, Budapest, 1973, pp. 267–281.
- [15] H. Lian, Shrinkage tuning parameter selection in precision matrices estimation, *Journal of Statistical Planning and Inference* 141 (2011) 2839–2848.
- [16] I. Vujačić, A. Abbruzzo, E. C. Wit, A computationally fast alternative to cross-validation in penalized Gaussian graphical models, [arXiv:1309.621v2 \[stat.ME\]](#) (2014).
- [17] O. Ledoit, M. Wolf, Improved estimation of the covariance matrix of stock returns with an application to portfolio selection, *Journal of Empirical Finance* 10 (2003) 603–621.
- [18] M. J. Daniels, R. E. Kass, Shrinkage estimators for covariance matrices, *Biometrics* 57 (2001) 1173–1184.

- [19] J. Friedman, T. Hastie, R. Tibshirani, **glasso**: Graphical lasso-estimation of Gaussian graphical models, R package, version 1.7 (2011).
URL <http://CRAN.R-project.org/package=glasso>
- [20] J. Whittaker, Graphical Models in Applied Multivariate Statistics, John Wiley & Sons Ltd., Chichester, 1990.
- [21] B. Efron, R. Tibshirani, J. D. Storey, V. Tusher, Empirical Bayes analysis of a microarray experiment, *Journal of the American Statistical Association* 96 (2001) 1151–1160.
- [22] H. Hotelling, New light on the correlation coefficient and its transforms, *Journal of the Royal Statistical Society, Series B* 15 (1953) 193–232.
- [23] J. Schäfer, K. Strimmer, An empirical Bayes approach to inferring large-scale gene association networks, *Bioinformatics* 21 (2005) 754–764.
- [24] B. Efron, Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction, Cambridge University Press, Cambridge, 2010.
- [25] S. Boyd, L. Xiao, Least-squares covariance matrix adjustment, *SIAM Journal on Matrix Analysis and Applications* 27 (2005) 532–546.
- [26] R. C. Gentleman, V. J. Carey, D. M. Bates, B. M. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, L. Cheng, M. Maechler, A. J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J. Y. H. Yang, J. Zhang, Bioconductor: Open software development for computational biology and bioinformatics, *Genome Biology* 5 (2004) R80.
- [27] M. Schröder, B. Haibe-Kains, A. Culhane, C. Sotiriou, G. Bontempi, J. Quackenbush, breastCancerMAINZ; breastCancerTRANSBIG; breastCancerUNT; breastCancerUPP; breastCancerVDX, R packages, versions 1.0.6 (2011).
URL <http://compbio.dfci.harvard.edu/>
- [28] B. Haibe-Kains, C. Desmedt, S. Loi, A. C. Culhane, G. Bontempi, J. Quackenbush, C. Sotiriou, A three-gene model to robustly identify breast cancer molecular subtypes, *Journal of the National Cancer Institute* 104 (2012) 311–325.
- [29] M. Kanehisa, S. Goto, KEGG: Kyoto Encyclopedia of Genes and Genomes, *Nucleic Acids Research* 28 (2000) 27–30.
- [30] L. Pecorino, Molecular Biology of Cancer: Mechanisms, Targets and Therapeutics, 3rd Edition, Oxford University Press, Oxford, 2012.
- [31] B. Vogelstein, S. Sur, C. Prives, p53: The most frequently altered gene in human cancers, *Nature Education* 3 (2010) 6.
- [32] D. Hanahan, R. A. Weinberg, The hallmarks of cancer, *Cell* 100 (2000) 57–70.
- [33] K. G. Jöreskog, Analysis of covariance structures, *Scandinavian Journal of Statistics* 8 (1981) 65–92.
- [34] J. H. Won, J. Lim, S. J. Kim, B. Rajaratnam, Condition-number-regularized covariance estimation, *Journal of the Royal Statistical Society, Series B* 75 (2013) 427–450.
- [35] C. F. W. Peeters, W. N. van Wieringen, **rags2ridges**: Ridge estimation of precision matrices from high-dimensional data, R package, Version 1.3 (2014).
URL <http://cran.r-project.org/web/packages/rags2ridges/index.html>
- [36] R Development Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria (2011).
URL <http://www.R-project.org/>
- [37] D. Serre, Matrices: Theory and Applications, Springer, New York, 2002.
- [38] H. Weyl, Das asymptotische Verteilungsgesetz der Eigenwerte linearer partieller Differentialgleichungen (mit einer Anwendung auf die Theorie der Hohlraumstrahlung), *Mathematische Annalen* 71 (1912) 441–479.