RIDGE ESTIMATION OF INVERSE COVARIANCE MATRICES
FROM HIGH-DIMENSIONAL DATA

BY WESSEL N. VAN WIERINGEN†‡ AND CAREL F.W. PEETERS†

VU University medical center† VU University Amsterdam‡

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 utilized penalties do not coincide with common ridge penalties. Subsequently, starting from a common ridge 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.

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

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

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

In the high-dimensional setting where $p > n$, the sample covariance matrix is singular and its inverse is undefined. Consequently, so is $\hat{\Omega}_{\text{ML}}$. A common workaround is the addition of a penalty to the log-likelihood (1). The $\ell_1$-penalized estimation of the precision matrix was first considered by [1] and [4]. This (graphical) lasso estimate of $\Omega$ 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 matrix. In addition, the true (graphical) model need not be 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 [6] 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 utilized 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 (Section 3). In Section 4 the alternative

---

*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).

Keywords and phrases: Graphical modeling; Multivariate normal; Penalized estimation; Precision matrix.

1
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. A graphical modeling application on oncogenomics data (Section 5) subsequently demonstrates that the alternative ridge estimators yield a lower loss (and select more stable networks) vis-à-vis the graphical lasso, in particular for more extreme \( p/n \) ratios. 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. [11, 14]). The first archetypal form of ridge estimator commonly is a convex combination of \( S \) and a positive definite (p.d.) target matrix \( T \): 

\[
\hat{\Omega} I(\lambda I) = [(1 - \lambda I)S + \lambda I T]^{-1}, \text{ with } \lambda I \in (0, 1].
\]

A common (low-dimensional) target choice is \( T \) diagonal with \( (T)_{jj} = (S)_{jj} \) for \( j = 1, \ldots, p \). This estimator has the desirable property of shrinking to \( T^{-1} \) 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 \( S \) with the lower-variance, higher-bias matrix \( T \). It can also be viewed as resulting from the maximization of the following penalized log-likelihood:

\[
\ln |\Omega| - (1 - \lambda I) \text{tr}(S\Omega) - \lambda I \text{tr}(\Omega T).
\]

The penalized log-likelihood (2) is obtained from the original log-likelihood (1) by the replacement of \( S \) by \( (1 - \lambda I)S \) and the addition of a penalty. The estimate \( \hat{\Omega} I(\lambda I) \) 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 

\[
\hat{\Omega} II(\lambda II) = (S + \lambda II I_p)^{-1} \text{ with } \lambda II \in (0, \infty). \]

It can be motivated as an ad-hoc fix of the singularity of \( S \) in the high-dimensional setting, much like how ridge regression was originally introduced by [8]. Alternatively, this archetype too can be viewed as a penalized estimate, as it maximizes (see also [19]):

\[
\ln |\Omega| - \text{tr}(S\Omega) - \lambda II \text{tr}(\Omega I_p).
\]

The penalties in (2) and (3) are non-concave (their second order derivatives equal the null-matrix \( 0 \)). This, however, poses no problem under the restriction of a p.d. solution \( \Omega \) as the Hessian of both (2) and (3) equals \(-\Omega^{-2}\). Of more concern is that neither penalty of the two current archetypes resembles the precision-analogy of what is commonly perceived as the ridge \( \ell_2 \)-penalty:

\[
\frac{1}{2} \lambda \|\Omega\|_2^2 = \frac{1}{2} \lambda \sum_{j_1, j_2=1}^{p} |(\Omega)_{j_1, j_2}|^2.
\]

The graphical lasso utilizes a penalty that is in line with the \( \ell_1 \)-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 common ridge \( \ell_2 \)-penalty. Consider Figure 1 to get a flavor of the behavior of both the archetypal graphical ridge 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.
Figure 1. Ridge coefficient paths of nonredundant off-diagonal elements for the archetypal and alternative (see Section 3) Type I (left panel) and Type II (right panel) ridge estimators. The 5×5 matrix S was generated as \((S^{-1})_{j_1,j_2} = [(j_1 \times j_2 + 1) \mod 21]/25\) if \(j_1 \neq j_2\) and \((S^{-1})_{j_1,j_2} = 1\) if \(j_1 = j_2\). The target matrix in the Type I case was taken to be the identity matrix \(I_5\). The penalty parameter is generically indicated by λ. For archetype-to-alternative scaling of the penalty parameters under Type I and Type II estimation see Section 4.1.

3. Alternative ridge estimators of the precision matrix.

3.1. Type I. In this section an analytic expression for an alternative Type I ridge precision estimator is given. Before arriving at the proposition containing the result, we employ the following lemma, whose proof (as indeed all proofs) is deferred to Appendix A:

**Lemma 1.** Amend the log-likelihood (1) with the \(\ell_2\)-penalty

\[
\frac{1}{2} tr \left[ (\Omega - T)^T \Lambda (\Omega - T) \right],
\]

with \(T\) denoting a p.d. symmetric target matrix, and \(\Lambda\) denoting a p.d. symmetric matrix bearing penalty parameters. Under given penalty, the penalized ML ridge proper estimator amounts to

\[
\hat{\Omega}^a(\Lambda) = \left\{ \Lambda + \frac{1}{4} (S - \Lambda T)^2 \right\}^{1/2} + \frac{1}{2} (S - \Lambda T)^{-1}.
\]  

**Proposition 1 (Alternative Type I ridge precision estimator).** Consider the generic alternative ridge estimator (5) from Lemma 1. Choose \(\Lambda \equiv \lambda_a I_p\), \(\lambda_a \in (0, \infty)\). Under this choice an alternative Type I ridge estimator is obtained as

\[
\hat{\Omega}^{Ia}(\lambda_a) = \left\{ \lambda_a I_p + \frac{1}{4} (S - \lambda_a T)^2 \right\}^{1/2} + \frac{1}{2} (S - \lambda_a T)^{-1},
\]
for which the following properties hold:

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

Analogous to the archetypal I estimator, the right and left-hand limits of the proposed estimator are the (imaginary) inverse of the ML estimator \( S \) and a target matrix, respectively. For a fuller understanding of the estimator (6), consider the following remarks.

Remark 1. It may be noticed that the penalty term (4) under the choice \( \Lambda \equiv \lambda_a I_p \) amounts to a proper ridge penalty as
\[
\frac{1}{2} \text{tr} \left[ (\Omega - T)^T (\Omega - T) \right] = \frac{1}{2} \|\Omega - T\|_F^2.
\]
When \( T = 0 \), we obtain \( \frac{1}{2} \|\Omega\|_F^2 \); a special case that will be considered in Section 3.2.

Remark 2. From Proposition 1 it is clear that (6) 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 \( S \) is ill-behaved. To obtain a well-conditioned estimate in such situations, one should choose \( \lambda_a \) not too close to zero. In order to choose an optimal value of \( \lambda_a \) for a problem at hand, one can employ (approximate) cross-validation or utilize information criteria.

Remark 3. Proposition 1 considers regularized estimation of the precision matrix. It may also provide an alternative Type I regularized estimator for the covariance matrix, by entertaining

\[
[\hat{\Sigma}^{Ia}(\lambda_a)]^{-1} = \left[ \lambda_a I_p + \frac{1}{4} (S - \lambda_a T)^2 \right]^{1/2} + \frac{1}{2} (S - \lambda_a T).
\]

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

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

Corollary 1 (Alternative Type II ridge precision estimator). Consider the alternative Type I ridge estimator (6) from Proposition 1. An alternative ridge proper Type II estimator is obtained by choosing \( T = 0 \), such that

\[
\hat{\Omega}^{IIa}(\lambda_a) = \left\{ \left[ \lambda_a I_p + \frac{1}{4} S^2 \right]^{1/2} + \frac{1}{2} S \right\}^{-1}.
\]

For this estimator, the following properties hold:
(i) \( \hat{\Omega}^{IIa}(\lambda_a) > 0 \), for all \( \lambda_a \in (0, \infty) \);
(ii) \( \lim_{\lambda_a \to 0^+} \hat{\Omega}^{IIa}(\lambda_a) = S^{-1} \);
(iii) \( \lim_{\lambda_a \to \infty^-} \hat{\Omega}^{IIa}(\lambda_a) = 0 \).

Analogous to the archetypal II estimator, the right and left-hand limits are the (imaginary) inverse of the ML estimator \( S \) and the null-matrix, respectively. The alternative Type II analogies of Remarks 2 and 3 hold for (7).

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:

\[
\hat{\Sigma}^{Ia}(\lambda_a) = \sqrt{\lambda_a} \left[(I_p + U^2)^{1/2} + U\right],
\]

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

\[
\hat{\Sigma}^{Ia}(\lambda_a) = \sqrt{\lambda_a} U + \sqrt{\lambda_a} \sum_{k=0}^{\infty} \left(\frac{1}{2}\right)_k U^{2k}.
\]

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) \( S \) follows a (singular) Wishart distribution, assume \( T \) 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:

\[
2\sqrt{\lambda_a} U = \sqrt{\lambda_a} \left[(I_p + U^2)^{1/2} + U\right] - \sqrt{\lambda_a} \left[(I_p + U^2)^{1/2} + U\right]^{-1},
\]

with \( U \) 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 (8) separately. Reformulated we then have:

\[
S - \lambda_a T = \hat{\Sigma}^{Ia}(\lambda_a) - \lambda_a \hat{\Omega}^{Ia}(\lambda_a).
\]

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 \( T \) to be the null-matrix.
4. Comparisons.

4.1. Eigenvalue shrinkage. The alternative Type I estimator \((\hat{\Omega}_{Ia}(\lambda_a))^{-1}\) is, as its archetypal counterpart, rotation equivariant. That is, the effect of the ridge penalty on the precision estimate is equivalent to shrinkage of the eigenvalues of the unpenalized estimate \(S^{-1}\). To see this, let the eigen-decomposition of \(S\) be \(VDV^T\) where \(D\) is a diagonal matrix with the eigenvalues of \(S\) on the diagonal and \(V\) denotes the matrix that contains the corresponding eigenvectors as columns. The orthogonality of \(V\) implies \(VV^T = V^TV = I_p\). We then rewrite, using \(T = I_p\) for notational convenience, the inverse of (6) as follows:

\[
[\hat{\Omega}_{Ia}(\lambda_a)]^{-1} = \left[\lambda_aVV^T + \frac{1}{4}(VDV^T - \lambda_aVV^T)^2\right]^{1/2} + \frac{1}{2}(VDV^T - \lambda_aVV^T)
\]

(9)

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

\[
[\hat{\Omega}_{II}(\lambda)]^{-1} = V \left[\left(\lambda_a I_p + \frac{1}{4}D - \lambda_a I_p\right)^2\right]^{1/2} + \frac{1}{2}(D - \lambda_a I_p) V^T.
\]

(10)

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

**Proposition 2.** Let the regularization parameters of the archetypal and alternative Type I ridge estimators – \(\lambda_I\) and \(\lambda_a\) respectively – map to the same scale. That is, choose \(\lambda_I = 1 - 1/(\lambda_a + 1)\). In addition, consider a diagonal matrix as the low-dimensional target matrix \(T\). Then the alternative estimator \(\hat{\Omega}_{Ia}(\lambda_a)\) displays shrinkage of the eigenvalues of \(S^{-1}\) that is at least as heavy as the shrinkage propagated by the archetypal estimator \(\hat{\Omega}_{II}(\lambda_I)\).

**Proposition 3.** Let the regularization parameters of the archetypal and alternative Type II ridge estimators – \(\lambda_{II}\) and \(\lambda_a\) respectively – map to the same scale. That is, choose \(\lambda_a = \lambda_{II}^2\). Then the archetypal estimator \(\hat{\Omega}_{II}(\lambda_{II})\) displays shrinkage of the eigenvalues of \(S^{-1}\) that is at least as heavy as the shrinkage propagated by the alternative estimator \(\hat{\Omega}_{Ia}(\lambda_a)\).

**Corollary 2.** The eigenvalue inequality of Proposition 3 implies:

\[
\mathcal{L}[\hat{\Omega}_{II}(\lambda_{II}); S] \leq \mathcal{L}[\hat{\Omega}_{Ia}(\lambda_a); S].
\]

The alternative Type I estimator displays faster shrinkage to the target \(T\) 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, $L[\hat{\Omega}_I(\lambda_I); S] \leq L[\hat{\Omega}_{1a}(\lambda_a); S]$. In absence of a natural target $T$, Type II estimators are an option. It is seen from proposition 3 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 2 and 3. It is not if we notice that the penalty parameter $\lambda_a$ is more influential in the Type I alternative as its effect is not diluted by a null $T$. 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 $\Omega$ denote a generic ($p \times p$) population precision matrix and let $\hat{\Omega}(\lambda)$ denote a generic ridge estimator of the precision matrix under generic regularization parameter $\lambda$. The following loss functions are then considered in risk evaluation:

a. Squared Frobenius loss, given by:

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

b. Quadratic loss, given by:

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

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

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

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. $\Omega^{random}$ with no conditional dependencies, generated as $\Omega^{random} = \frac{1}{n}Y^TY$ from the ($n \times p$)-dimensional matrix $Y$ with $n = 10,000$ and each $Y_{ij}$ drawn from $\mathcal{N}(0, 1)$;
2. $\Omega^{chain}$ representing a conditional independence graph with a chain topology. Its element are $\Omega^{chain}_{j,j} = 1$, $\Omega^{chain}_{j,j+1} = 0.25 = \Omega^{chain}_{j+1,j}$ for $j = 1, \ldots, p - 1$, and zero otherwise;
3. $\Omega^{star}$ representing a conditional independence graph with a star topology. Its element are $\Omega^{star}_{j,j} = 1$, $\Omega^{star}_{1,j+1} = 1/(j + 1) = \Omega^{star}_{j+1,1}$ for $j = 1, \ldots, p - 1$, and zero otherwise;
4. $\Omega^\text{clique}$ 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 ($T = 0$), a diagonal target ($\text{diag}[T] = 1/\text{diag}[S]$), and a target equal to the true precision matrix ($T = \Omega$). 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 ($\text{diag}[T] = 1/\text{diag}[S]$ and $T = \Omega$) plotted against the penalty parameter (see Supplement A for visualizations of all risk comparisons).

![Figure 2](image.png)

**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 $\text{diag}[T] = 1/\text{diag}[S]$. The right hand panel compares the alternative and archetypal Type I ridge estimators when $T = \Omega$. 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 $\lambda$ parameter under the various estimators were chosen in accordance with the eigenvalue comparison in Section 4.1.

The simulation results (as summarized in Figure 2 and Supplement A) 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 ($T = \Omega$) yields a lower loss for large values of $\lambda$ 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 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. Illustration.

5.1. Data and goal. Within the context of Gaussian graphical modeling the support of the precision matrix can be used to read off the conditional (in)dependencies among the variates. In this section we investigate how well the proposed ridge estimators of the precision matrix uncover the conditional (in)dependencies from high-dimensional data. The performance of the alternative ridge estimators are also contrasted with the graphical lasso [4] in terms of loss. The graphical lasso is the lasso estimator of the precision matrix, which, next to shrinkage, performs automatic selection of conditional dependencies. Two versions of each estimator are considered. On the ridge side the Type II alternative ridge precision estimator with $T = 0$ and the Type I alternative ridge estimator with $\text{diag}(T) = 1/\text{diag}(S)$ are considered. The concordant graphical lasso precision estimators employ penalization and no penalization of the diagonal elements, respectively (see the glasso package [5]).

The performance of the ridge and lasso precision estimators is evaluated on gene expression data of two pathways from two oncogenomics studies. The data originate from The Cancer Genome Atlas (TCGA), in particular from the breast and colorectal cancer studies [15, 16]. The chosen pathways, p53 and ErbB signaling, are defined by KEGG [10]. 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 [12]. 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 [18]. The ErbB family of tyrosine kinases (phosphate transferring enzymes) is frequently overexpressed in human cancers. This may lead to oncogenic signaling, making the cancerous cell self-sufficient in survival and multiplication [12], one of the hallmarks of human cancer [7]. The underlying conditional dependency structure of both pathways is not fully known but is (generally) believed to be (relatively) sparse. Normalized data for the genes of both pathways are retrieved from the cBioPortal for Cancer Genomics [2]. Samples and genes with more than 10% missings have been removed and remaining missing data are imputed by means of
the \( k \)-nearest neighbor method of [17]. The resulting dimensions of the pathway data are \( n = 526 \) (breast) or \( n = 195 \) (colorectal), and \( p = 87 \) (ErbB) or \( p = 67 \) (p53).

5.2. Loss comparison. 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. For each sample size one hundred subsamples of the pathway data are drawn. Optimal values of the penalty parameter for both versions of the alternative ridge and lasso estimators are obtained for each subsample by way of leave-one-out cross-validation. The ridge and lasso precision estimates for a subsample then

![Box plots for loss comparison]

**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 breast cancer data. The upper panels depict loss for the p53-pathway while the lower panels depict loss for the ErbB-pathway. The left-hand panels depict quadratic loss while the right-hand panels depict Frobenius loss.
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).

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 breast cancer data is defined as the $(67 \times 67)$-dimensional standardized sample precision matrix over all $n = 526$ samples. The results of the loss evaluation are displayed in Figure 3 and Supplement B.

Figure 3 and Section B.1 of Supplement B 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.

5.3. Conditional independence graphs. To construct a conditional independence graph one needs to determine the support of the (standardized) precision matrix. While the graphical lasso performs automated model selection, the alternative ridge estimators 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 (FDR) procedure [3] proposed by [14] that fits a mixture distribution to the regularized partial correlations. This allows one to determine the posterior probability that a partial correlation is non-null (and thus that the corresponding edge is present) given the observed regularized estimate of that partial correlation. While such a two-step procedure does not have the appeal of simultaneous estimation and model selection, it does have the advantage that it enables probabilistic statements about the (number of) selected edges. In the remainder we will select edges that have a posterior probability of being non-null of at least .8.

Figure 4 contains conditional independence graphs for the Type I alternative ridge estimator with $\text{diag}[T] = 1/\text{diag}[S]$ and the corresponding graphical lasso on the p53-pathway breast cancer 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 local FDR support determination selects more stable (in terms of network-structure change) networks over the respective sample sizes. While usage of local FDR edge selection on the ridge regularized precision matrix tends to gain in conservativeness with growing $n$ (see Section B.2 of Supplement B), 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 ErbB-pathway and the colorectal data under $\text{diag}[T] = 1/\text{diag}[S]$ (see Section B.3 of Supplement B). For the Type II comparison the alternative ridge estimator
Figure 4. Conditional independence graphs for the Type I alternative ridge estimator using local FDR edge selection (left-hand figures) and the corresponding graphical lasso (right-hand figures) on the p53-pathway breast cancer 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).
tends to be more conservative than the graphical lasso for the higher sample sizes (Section B.3 of Supplement B). This should not surprise as the Type II alternative estimator proportionally shrinks the full estimate, in the limit, to the stringent null matrix. This can 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 $T$ for better signal preservation. In all, the alternative (Type I) 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.

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 utilized in supporting covariance regularized regression [21], 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 local FDR multiple testing [14]. Furthermore, regularized (inverse) covariance matrices stemming from the proposed estimators can be used as input in covariance structure modeling efforts [9] (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 [22], so that it can be formalized which values for the penalty parameter can be considered ‘too small’. From a more applied perspective it may be deemed interesting to compare multiple methods for determining the optimal value of the penalty parameter. These issues are the focal points of current research.

The ridge estimators employed in this paper are implemented in the R-package rags2ridges 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/) [13].

References.
[1] O. Banerjee, L. El Ghaoui, and A. d’Aspremont. Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *Journal of Machine Learning Research*, 9:485–516, 2008.

[2] E. Cerami, J. Gao, U. Dogrusoz, B. E. Gross, S. O. Sumer, B. A. Aksoy, A. Jacobsen, C. J. Byrne, M. L. Heuer, E. Larsson, Y. Antipin, B. Reva, A. P. Goldberg, C. Sander, and N. Schultz. The cbio Cancer Genomics Portal: An open platform for exploring multidimensional cancer genomics data. *Cancer Discovery*, 2:401–404, 2012.

[3] B. Efron, R. Tibshirani, J. D. Storey, and V. Tusher. Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association*, 96:1151–1160, 2001.

[4] J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9:432–441, 2008.

[5] J. Friedman, T. Hastie, and R. Tibshirani. *glasso: Graphical lasso-estimation of Gaussian graphical models*, 2011. R package version 1.7.

[6] W. J. Fu. Penalized regressions: The Bridge versus the Lasso. *Journal of Computational and Graphical Statistics*, 7:397–416, 1998.

[7] D. Hanahan and R. A. Weinberg. The hallmarks of cancer. *Cell*, 100:57–70, 2000.

[8] A. E. Hoerl and R. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12:55–67, 1970.

[9] K. G. Jöreskog. Analysis of covariance structures. *Scandinavian Journal of Statistics*, 8:65–92, 1981.

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

[11] O. Ledoit and M. Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88:365–411, 2004.

[12] L. Pecorino. *Molecular Biology of Cancer: Mechanisms, Targets and Therapeutics*. Oxford: Oxford University Press, 3rd edition, 2012.

[13] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2011. ISBN 3-900051-07-0.

[14] J. Schäfer and K. Strimmer. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology*, 4:art. 32, 2005.

[15] The Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature*, 490:61–70, 2012.

[16] The Cancer Genome Atlas Research Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, 487:330–337, 2012.

[17] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman. Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17:520–525, 2001.

[18] B. Vogelstein, S. Sur, and C. Prives. p53: The most frequently altered gene in human cancers. *Nature Education*, 3:6, 2010.

[19] D.I. Warton. Penalized normal likelihood and ridge regularization of correlation and covariance matrices. *Journal of the American Statistical Association*, 103:340–349, 2008.

[20] H. Weyl. Das asymptotische Verteilungsgesetz der Eigenwerte linearer partieller Differentialgleichungen (mit einer Anwendung auf die Theorie der Hohlraumstrahlung). *Mathematische Annalen*, 71:441–479, 1912.

[21] D. M. Witten and R. Tibshirani. Covariance-regularized regression and classification for high-dimensional problems. *Journal of the Royal Statistical Society, Series B*, 71:615–636, 2009.

[22] J. H. Won, J. Lim, S. J. Kim, and B. Rajaratnam. Condition-number-regularized covariance estimation. *Journal of the Royal Statistical Society, Series B*, 75:427–450, 2013.
APPENDIX A: PROOFS

A.1. Proof Lemma 1.

**Proof.** Define the penalized log-likelihood:

\[ L^p(\Omega; S) \propto \ln |\Omega| - \text{tr}(S\Omega) - \frac{1}{2}\text{tr}[(\Omega - T)^T\Lambda(\Omega - T)]. \]

Now, take the derivative of \( L^p(\Omega; S) \) w.r.t. \( \Omega \):

\[ \frac{\partial L^p(\Omega; S)}{\partial \Omega} = \Omega^{-1} - (S - \Lambda T) - \Lambda \Omega. \]  

Equate (11) to zero and post-multiply by \( \Omega^{-1} \). Subsequently adding \( \frac{1}{4}(S - \Lambda T)^2 \) to both sides of the equality sign gives:

\[ \Lambda + \frac{1}{4}(S - \Lambda T)^2 = \Omega^{-2} - (S - \Lambda T)\Omega^{-1} + \frac{1}{4}(S - \Lambda T)^2. \]

Notice that \( \frac{1}{2}\text{tr}[(\Omega - T)^T\Lambda(\Omega - T)] = \frac{1}{2}\text{tr}[(\Omega - T)\Lambda(\Omega - T)^T] \), so that \( \Omega \) also satisfies (under pre-multiplication by \( \Omega^{-1} \)):

\[ \Lambda + \frac{1}{4}(S - \Lambda T)^2 = \Omega^{-2} - \Omega^{-1}(S - \Lambda T) + \frac{1}{4}(S - \Lambda T)^2. \]

Adding (12) and (13) and subsequently dividing by 2 yields:

\[ \Lambda + \frac{1}{4}(S - \Lambda T)^2 = \Omega^{-2} - \frac{1}{2}\Omega^{-1}(S - \Lambda T) - \frac{1}{2}(S - \Lambda T)\Omega^{-1} + \frac{1}{4}(S - \Lambda T)^2. \]

Complete the square to obtain:

\[ \Lambda + \frac{1}{4}(S - \Lambda T)^2 = \left[ \Omega^{-1} - \frac{1}{2}(S - \Lambda T) \right]^2. \]

Taking the square root on both sides and solving for \( \Omega \) results in the desired expression (5).

A.2. Proof Proposition 1.

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

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

when \( \lambda_a > 0 \). Hence, \( \hat{\Omega}^{1a}(\lambda_a) \) is p.d. for any \( \lambda_a \in (0, \infty) \).
(ii) The right-hand limit is immediate as:

\[ \hat{\Omega}^a(0) = \left\{ \begin{array}{c} 0 + \frac{1}{4} (S - 0T)^2 \end{array} \right\}^{1/2} + \frac{1}{2} (S - 0T)^{-1} = S^{-1}. \]

(iii) For the left-hand limit we note that, when \( \lambda_a \) approaches \( \infty \),

\[ [\hat{\Omega}^a(\lambda_a)]^{-1} = \left( \lambda_a I_p + \frac{1}{4} (S - \lambda_a T)^2 \right)^{1/2} + \frac{1}{2} (S - \lambda_a T) \to T^{-1}, \]

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,

\[ T^{1/2} \left\{ \lambda_a I_p + \frac{1}{4} (S - \lambda_a T)^2 \right\}^{1/2} + \frac{1}{2} (S - \lambda_a T) \to I_p, \]

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

\[ \left[ \frac{1}{4} (\lambda_a T^2 + 2I_p - \hat{S})^2 + (\hat{S} - I_p) \right]^{1/2} - \frac{1}{2} \left( \lambda_a T^2 + 2I_p - \hat{S} \right) + I_p, \]

implying that the problem can be reduced to proving

\[ \lim_{\lambda_a \to \infty} \left[ B^2(\lambda_a) + (\hat{S} - I_p) \right]^{1/2} - B(\lambda_a) = 0, \]

where \( B(\lambda_a) = \frac{1}{2} \left( \lambda_a T^2 + 2I_p - \hat{S} \right) \).

To prove this invoke Weyl’s eigenvalue inequality [20]. Let \( A, B, C = A + B \) be real, symmetric \( p \times p \) matrices with eigenvalues \( \alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_p, \beta_1 \geq \beta_2 \geq \cdots \geq \beta_p \), and \( \gamma_1 \geq \gamma_2 \geq \cdots \geq \gamma_p \), 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 \( C(\lambda) = \lambda A + B \) with \( \lambda > 0 \) (where \( \lambda \) is used generically) and \( A \) and \( B \) 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 \( \lambda \) and let \( \lambda \) tend to infinity (from the left), which is immediately seen to imply:

\[ \lim_{\lambda \to \infty} \frac{1}{\lambda} \gamma_j(\lambda) = \alpha_j \quad \text{for all } j. \]

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

\[ \lim_{\lambda_a \to \infty} \left[ B^2(\lambda_a) + (\hat{S} - I_p) \right]^{1/2} - B(\lambda_a) = 0, \]

as indeed needed to be proven. \( \square \)
A.3. Proof Corollary 1.

Proof.  
(i) Let $d(\cdot)_{jj}$ denote the $j$’th eigenvalue of the matrix term in brackets $(\cdot)$. Notice
\[
d\{ [\hat{\Omega}^{\text{IIa}}(\lambda_a)]^{-1} \}_{jj} = d\left( \frac{1}{2} S \right)_{jj} + \sqrt{\left[ d\left( \frac{1}{2} S \right)_{jj} \right]^2 + \lambda_a} > 0,
\]
when $\lambda_a > 0$, implying $\hat{\Omega}^{\text{IIa}}(\lambda_a)$ is p.d. for any $\lambda_a \in (0, \infty)$.
(ii) The right-hand limit is immediate as:
\[
\hat{\Omega}^{\text{IIa}}(0) = \left\{ \left[ 0I_p + \frac{1}{4} S^2 \right]^{1/2} + \frac{1}{2} S \right\}^{-1} = S^{-1}.
\]
(iii) For the left-hand limit we note that as $\lambda_a$ approaches $\infty$,
\[
[\hat{\Omega}^{\text{IIa}}(\lambda_a)]^{-1} = \left[ \lambda_a I_p + \frac{1}{4} S^2 \right]^{1/2} + \frac{1}{2} S
\]
becomes a diagonally dominant matrix with near infinite diagonal values. The inverse of which must necessarily approach the null-matrix.

A.4. Proof Proposition 2. The proof will be based on the target $T = I_p$. The extension to a general p.d. diagonal target matrix is straightforward as it is a direct consequence, but notationally slightly more cumbersome.

Proof. Note that $\hat{\Omega}^{\text{I}}(\lambda_I)$ can be decomposed as:
\[
[\hat{\Omega}^{\text{I}}(\lambda_I)]^{-1} = V[ (1 - \lambda_I)D + \lambda_I I_p ] V^T.
\]
Juxtaposing this expression with (9) while writing $d_{jj} = (D)_{jj}$, we are after establishing
\[
\sqrt{\lambda_a + \frac{1}{4} (d_{jj} - \lambda_a)^2 + \frac{1}{2} (d_{jj} - \lambda_a) > \sqrt{ \lambda_a } d_{jj} + \frac{\lambda_a}{1 + \lambda_a} ,
\]
which after some ready algebra can be rewritten as:
\[
\sqrt{\varphi_{jj}(\lambda_a)^2 + d_{jj} - 1 > \sqrt{ \lambda_a } (d_{jj} - 1) + \varphi_{jj}(\lambda_a) ,
\]
with $\varphi_{jj}(\lambda_a) = \frac{1}{2} \lambda_a - \frac{1}{2} d_{jj} + 1$. Squaring both sides and simplifying the problem becomes:
\[
d_{jj} - 1 > \sqrt{ \frac{1}{(1 + \lambda_a)^2} (d_{jj} - 1)^2 + d_{jj} - 1 - \frac{1}{1 + \lambda_a} (d_{jj} - 1)^2} ,
\]
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 \( \hat{\Omega}_{Ia}^H(\lambda_a) \) displays shrinkage of the eigenvalues of \( S^{-1} \) that is at least as heavy as the shrinkage propagated by the archetypal estimator \( \hat{\Omega}_I^H(\lambda_I) \).

A.5. Proof Proposition 3.

Proof. Note that the decomposition of the original ridge estimator of the second type is
\[
[\hat{\Omega}^H(\lambda_{II})]^{-1} = V(\lambda_{II} I_p + D)V^T.
\]
Then, when writing \( d_{jj} = (D)_{jj} \) while juxtaposing the above expression with (10), we have:
\[
\lambda_{II} + d_{jj} \geq \sqrt{\lambda_{II}^2 + \frac{1}{4}d_{jj}^2 + \frac{1}{2}d_{jj}},
\]
as follows directly from \( (\lambda_{II} + \frac{1}{2}d_{jj})^2 \geq \lambda_{II}^2 + \frac{1}{4}d_{jj}^2 \). This indicates the archetypal estimator \( \hat{\Omega}^H(\lambda_{II}) \) displaying shrinkage of the eigenvalues of \( S^{-1} \) that is at least as heavy as the shrinkage propagated by the alternative estimator \( \hat{\Omega}_{IIa}^H(\lambda_a) \).

A.6. Proof Corollary 2.

Proof. Note:
\[
\mathcal{L}[\hat{\Omega}^H(\lambda_{II}); S] \propto \ln |\hat{\Omega}^H(\lambda_{II})| - \text{tr}[S\hat{\Omega}^H(\lambda_{II})] \propto -\sum_{j=1}^p \ln(\lambda_{II} + d_{jj}) - \sum_{j=1}^p \frac{d_{jj}}{\lambda_{II} + d_{jj}}.
\]
Similarly:
\[
\mathcal{L}[\hat{\Omega}_{IIa}^H(\lambda_a); S] \propto -\sum_{j=1}^p \ln[\gamma_{jj}(\lambda_{II})] - \sum_{j=1}^p \frac{d_{jj}}{\gamma_{jj}(\lambda_{II})},
\]
where \( \gamma_{jj}(\lambda_{II}) = \sqrt{\lambda_{II}^2 + \frac{1}{4}d_{jj}^2 + \frac{1}{2}d_{jj}} \). It then suffices to show that
\[
\ln(\lambda_{II} + d_{jj}) - \ln[\gamma_{jj}(\lambda_{II})] + \frac{d_{jj}}{\lambda_{II} + d_{jj}} - \frac{d_{jj}}{\gamma_{jj}(\lambda_{II})} \geq 0.
\]
Using $\ln(1 + x) \geq x/(1 + x)$ and $d_{jj} + \lambda_{II} \geq \gamma_{jj}(\lambda_{II}) \geq d_{jj}$ (Proposition 3), the manipulations below prove this:

$$
\ln\left(\frac{\lambda_{II} + d_{jj}}{\gamma_{jj}(\lambda_{II})}\right) + \frac{d_{jj}}{\lambda_{II} + d_{jj}} - \frac{d_{jj}}{\gamma_{jj}(\lambda_{II})} \geq \lambda_{II} + d_{jj} - \frac{\gamma_{jj}(\lambda_{II})}{\gamma_{jj}(\lambda_{II})(\lambda_{II} + d_{jj})} = 0.
$$

SUPPLEMENTARY MATERIAL

**Supplement A: Figures risk comparison**

(). Supplement containing all figures of the risk comparison of Section 4.2

**Supplement B: Figures alternative ridge and graphical lasso comparison**

(). Supplement containing all figures of the comparison between the alternative ridge estimators and the graphical lasso in Section 5

W.N. van Wieringen  
Dept. of Epidemiology & Biostatistics  
VU University Medical Center  
Amsterdam, The Netherlands  
E-mail: w.vanwieringen@vumc.nl  
(Corresponding author)

C.F.W. Peeters  
Dept. of Epidemiology & Biostatistics  
VU University Medical Center  
Amsterdam, The Netherlands  
E-mail: cf.peeters@vumc.nl