Taylor & Francis
Taylor & Francis Group

Journal of Computational and Graphical Statistics

ISSN: 1061-8600 (Print) 1537-2715 (Online) Journal homepage: https://www.tandfonline.com/loi/ucgs20

The Generalized Ridge Estimator of the Inverse Covariance Matrix

Wessel N. van Wieringen

To cite this article: Wessel N. van Wieringen (2019) The Generalized Ridge Estimator of the Inverse Covariance Matrix, Journal of Computational and Graphical Statistics, 28:4, 932-942, DOI: 10.1080/10618600.2019.1604374

To link to this article: https://doi.org/10.1080/10618600.2019.1604374

JOURNAL OF COMPUTATIONAL AND GRAPHICAL STATISTICS
2019, VOL. 28, NO. 4, 932–942
https://doi.org/10.1080/10618600.2019.1604374

Taylor & Francis
Taylor & Francis Group

The Generalized Ridge Estimator of the Inverse Covariance Matrix

Wessel N. van Wieringen[a,b]

[a]Department of Epidemiology and Biostatistics, Amsterdam School of Public Health, Amsterdam UMC, location VUmc, Amsterdam, The Netherlands; [b]Department of Mathematics, VU University Amsterdam, Amsterdam, The Netherlands

ABSTRACT

The ridge inverse covariance estimator is generalized to allow for entry-wise penalization. An efficient algorithm for its evaluation is proposed. Its computational accuracy is benchmarked against implementations of specific cases the generalized ridge inverse covariance estimator encompasses. The proposed estimator shrinks toward a user-specified, nonrandom target matrix and is shown to be positive definite and consistent. It is pointed out how the generalized ridge inverse covariance estimator can be used to obtain a generalization of the graphical lasso estimator as well as of its elastic net counterpart. The usage of the presented estimator is illustrated in graphical modeling of omics data. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received September 2017
Revised March 2019

KEYWORDS
Graphical lasso; Multivariate normality; Nonzero centered penalty; Penalized estimation; Precision matrix.

1. Introduction

Interest in graphical models that combine a probabilistic description (through a multivariate distribution) of a system with a graph that depicts the system's structure (capturing dependence relationships), has surged in recent years. In its trail this has renewed the attention to the estimation of precision matrices as they harbor the conditional (in)dependencies among jointly distributed variates. In particular, with the advent of high-dimensional data, for which traditional precision estimators are not well-defined, this brought about several novel precision estimators. Generally, these novel estimators overcome the undersampling by maximization of the log-likelihood augmented with a so-called penalty. A penalty discourages large (in some sense) values among the elements of the precision matrix estimate. This reduces the risk of overfitting but also yields a well-defined penalized precision matrix estimator. In this work, penalized precision estimators are generalized to allow for (i) different penalization among the elements of the precision matrix and (ii) the incorporation of prior knowledge on these elements.

A Gaussian graphical model of a collection of random variables $Y_1, \ldots, Y_p$ assumes they are jointly distributed as a multivariate normal law. When these random variables are stacked in a p-dimensional random vector denoted $\mathbf{Y}$, the Gaussian graphical model amounts to $\mathbf{Y} \sim \mathcal{N}(\mathbf{0}_p, \boldsymbol{\Omega}^{-1})$ with $\boldsymbol{\Omega}$ the inverse covariance matrix—also called precision matrix or simply precision. In principle, the mean may be nonzero but that is of no concern here. The covariance matrix is parametrized by its precision matrix as the latter harbors the conditional (in)dependencies among the p random variables. It is well-known that $(\boldsymbol{\Omega})_{j,j'} = 0$ implies that the corresponding random variables Y_j and $Y_{j'}$ are independent conditionally on all other random variables contained in $\mathbf{Y}$. Similarly, a nonzero in $\boldsymbol{\Omega}$ indicates conditional dependence.

The estimation of the Gaussian graphical model boils down to the estimation of the precision matrix. This requires a draw of $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ independent random vectors all following the same $\mathcal{N}(\mathbf{0}_p, \boldsymbol{\Omega}^{-1})$-law. When $n > p$, the precision matrix $\boldsymbol{\Omega}$ may be estimated by the inverse of the sample covariance matrix $\mathbf{S} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{Y}_i\mathbf{Y}_i^\top$. The latter becomes singular when p is close to n or even exceeds it ($p > n$). The precision $\boldsymbol{\Omega}$ then needs to be estimated in penalized fashion. Estimators have been proposed for both dominant penalization flavors, lasso and ridge (Banerjee, Ghaoui, and d'Aspremont 2008; Friedman, Hastie, and Tibshirani 2008; Witten, Friedman, and Simon 2011; van Wieringen and Peeters 2016; Kuismin, Kemppainen, and Sillanpää 2017). The lasso precision estimator, referred as the graphical lasso, maximizes—in its most general form presented by Friedman, Hastie, and Tibshirani (2008)—the log-likelihood augmented with a penalty of the form $\|\boldsymbol{\Lambda} \circ \boldsymbol{\Omega}\|_1$, in which $\boldsymbol{\Lambda}$ is a matrix of penalty parameters and the $\circ$-operator is the Hadamard or element-wise matrix product. The graphical lasso thus allows for an element-wise penalization of the absolute values of the precision matrix. No analytic expression of the graphical lasso estimator is known and Friedman, Hastie, and Tibshirani (2008) present an efficient algorithm for its evaluation. The ridge precision estimator maximizes—in its most general form presented by van Wieringen and Peeters (2016)—the log-likelihood augmented by $\frac{1}{2}\lambda_2\|\boldsymbol{\Omega} - \mathbf{T}\|_F^2$, where λ_2 is the penalty parameter, $\mathbf{T}$ a nonrandom $p \times p$-dimensional, symmetric, (semi)positive definite target matrix, and $\|\cdot\|_F^2$ denotes the (squared) Frobenius norm $\|\mathbf{X}\|_F^2 = \text{tr}(\mathbf{X}^\top\mathbf{X}) = \sum_{j_1,j_2=1}^{p}(\mathbf{X})_{j_1,j_2}^2$. The target matrix is chosen prior to estimation

CONTACT Wessel N. van Wieringen w.vanwieringen@vumc.nl Department of Mathematics, VU University Amsterdam, De Boelelaan 1081a, 1081 HV Amsterdam, The Netherlands.

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/r/JCGS.

Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JCGS.

and serves as an initial guess toward which the precision estimate is shrunken. This ridge precision estimator penalizes the elements of $\boldsymbol{\Omega}$ in equal amount by a common penalty parameter λ_2. Conveniently, an analytic expression of this estimator exists

$$\widehat{\boldsymbol{\Omega}}(\lambda_2) = \{\tfrac{1}{2}(\mathbf{S} - \lambda_2\mathbf{T}) + [\lambda_2\mathbf{I}_{pp} + \tfrac{1}{4}(\mathbf{S} - \lambda_2\mathbf{T})^2]^{1/2}\}^{-1}.$$

When $\mathbf{T} = \mathbf{0}_{pp}$, the ridge penalization of the precision estimator boils down to penalization of the eigenvalues of the sample covariance matrix (van Wieringen and Peeters 2016; Kuismin, Kemppainen, and Sillanpää 2017). The most general forms of the lasso and ridge estimators of the precision matrix thus generalize in two different directions: element-wise penalization and shrinkage toward a nonnull target, respectively. In this work, we are after precision estimators that allow for the generalization into these two directions simultaneously.

The aim for generalization of the existing penalized precision estimators can, apart from theoretical interests, also be motivated practically. Element-wise penalization allows for differentiation in the amount of shrinkage exerted on each element. When the variates can be endowed with an order induced by either time or space (as would be the case in a longitudinal study or in spatial statistics), it seems reasonable to assume that neighboring variates exhibit stronger dependencies then those further apart. Such decay of the dependency with time or distance may be enforced by penalizing the precision matrix' bands close to the diagonal less than those further away. On the other hand, the target may come in use when the data arrive in batches or when multiple studies of the same system are conducted. The first batch or study yields a precision matrix estimate, which may serve as an informed guess for the next batch or study. That is, an existing precision matrix estimate can be used as a target matrix when analyzing novel data. As such, the possibility of including a target in the penalty thus provides the way for updating in a high-dimensional context. In some cases it may even be that part of the precision matrix is considered known. This knowledge may be included in the estimation via the target and element-wise penalization. For the known elements a very large penalty parameter is to be chosen, thus forcing these precision matrix elements to be equal to those of the target matrix. The remaining and unknown precision matrix elements are then endowed with their own penalty that is chosen via (say) cross-validation.

Here we present an algorithm to find the generalized ridge precision estimator, for which several theoretical properties, for example, positive definiteness and consistency, are shown. Next it is pointed out that iterative application of the generalized ridge precision estimator produces the generalized graphical lasso estimator of the inverse covariance matrix. The computation accuracy of the proposed algorithm is benchmarked against specific encompassing cases. Moreover, simulations are presented that identify situations where the generalized ridge precision estimator improves upon its regular counterpart. The paper closes with an illustration of the usage of the presented estimator in graphical modeling of omics data.

2. Estimation

Here generalized ridge estimation of the precision matrix is presented. The common penalty parameter on all elements of $\boldsymbol{\Omega}$ (as employed in van Wieringen and Peeters (2016) or Kuismin, Kemppainen, and Sillanpää (2017)) is replaced by element-specific penalty parameters. These are gathered in a $p \times p$-dimensional, symmetric matrix denoted $\boldsymbol{\Lambda}$ with the (j_1, j_2)th element, denoted $\lambda_{j_1,j_2} \geq 0$, is the penalty parameter to the corresponding element of $\boldsymbol{\Omega}$. Augmenting the log-likelihood of the sample with the generalized ridge penalty gives the loss function

$$\log(|\boldsymbol{\Omega}|) - \text{tr}(\mathbf{S}\boldsymbol{\Omega}) - \tfrac{1}{2}\|\sqrt{\boldsymbol{\Lambda}} \circ (\boldsymbol{\Omega} - \mathbf{T})\|_F^2, \tag{1}$$

where—following Marcus and Sandy (1991)—$\sqrt{\boldsymbol{\Lambda}}$ denotes the Hadamard or entry-wise square root of $\boldsymbol{\Lambda}$ (not to be confused with the matrix square root, denoted $\boldsymbol{\Lambda}^{1/2}$, see Higham 2008). The generalized ridge estimator of the precision matrix $\boldsymbol{\Omega}$ is the maximum of the loss function (1). To find its maximum equate its derivative with respect to $\boldsymbol{\Omega}$ to zero, and solve the following estimating equation for $\boldsymbol{\Omega}$

$$\boldsymbol{\Omega}^{-1} - \mathbf{A} - \boldsymbol{\Lambda} \circ \boldsymbol{\Omega} = \mathbf{0}_{pp}, \tag{2}$$

where the matrix $\mathbf{A} \equiv \mathbf{A}(\mathbf{S}, \boldsymbol{\Lambda}, \mathbf{T}) = \mathbf{S} - \boldsymbol{\Lambda} \circ \mathbf{T}$ comprises known matrices only. When $\boldsymbol{\Lambda} = \lambda\mathbf{1}_{pp}$, where $\mathbf{1}_{pp}$ denotes a $p \times p$ dimensional matrix filled with ones, an explicit expression for the minimizer of the loss function exists (see van Wieringen and Peeters 2016). In general the estimating Equation (2) seems not to have an explicit solution.

The solution of the estimating Equation (2) is found by iteratively running over the columns—and rows due to the symmetry—of the (inverse) covariance matrix, considering all-but-one column/row of $\boldsymbol{\Omega}$ temporarily known. To this end each matrix in the estimating Equation (2) is partitioned as a 2-by-2 block matrix with identical block structure

$$\boldsymbol{\Omega} = \begin{pmatrix} \boldsymbol{\Omega}_{11} & \boldsymbol{\Omega}_{12} \\ \boldsymbol{\Omega}_{12}^\top & \boldsymbol{\Omega}_{22} \end{pmatrix}, \quad \mathbf{A} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{12}^\top & \mathbf{A}_{22} \end{pmatrix}, \quad \text{and} \quad \boldsymbol{\Lambda} = \begin{pmatrix} \boldsymbol{\Lambda}_{11} & \boldsymbol{\Lambda}_{12} \\ \boldsymbol{\Lambda}_{12}^\top & \boldsymbol{\Lambda}_{22} \end{pmatrix}.$$

Furthermore, the inverse of the thus partitioned precision matrix is known explicitly

$$\boldsymbol{\Omega}^{-1} = \begin{pmatrix} \boldsymbol{\Omega}_{11} & \boldsymbol{\Omega}_{12} \\ \boldsymbol{\Omega}_{12}^\top & \boldsymbol{\Omega}_{22} \end{pmatrix}^{-1} = \begin{pmatrix} \boldsymbol{\Omega}_{11}^{-1} + \mathbf{F}\,\mathbf{E}^{-1}\,\mathbf{F}^\top & -\mathbf{F}\mathbf{E}^{-1} \\ -\mathbf{E}^{-1}\,\mathbf{F}^\top & \mathbf{E}^{-1} \end{pmatrix},$$

where $\mathbf{E} = \boldsymbol{\Omega}_{22} - \boldsymbol{\Omega}_{12}^\top\,\boldsymbol{\Omega}_{11}^{-1}\,\boldsymbol{\Omega}_{12}$ and $\mathbf{F} = \boldsymbol{\Omega}_{11}^{-1}\,\boldsymbol{\Omega}_{12}$. Now without loss of generality assume that the $(p-1) \times (p-1)$ matrix $\boldsymbol{\Omega}_{22}$ known, and thus that $\boldsymbol{\Omega}_{11}$ is a scalar while $\boldsymbol{\Omega}_{12}$ is a vector of length $p-1$, both unknown. Then, after substitution of the block partitioned matrices and the above inverse of a 2×2 block matrix in the estimating Equation (2) one obtains directly $\mathbf{E}^{-1} = \mathbf{A}_{22} + \boldsymbol{\Lambda}_{22} \circ \boldsymbol{\Omega}_{22}$. Aforementioned substitutions also yield the estimating equation for $\boldsymbol{\Omega}_{12}$

$$\mathbf{0}_{p-1} = -\boldsymbol{\Omega}_{11}^{-1}\,\mathbf{E}^{-1}\boldsymbol{\Omega}_{12}^\top - \mathbf{A}_{12}^\top - \text{diag}(\boldsymbol{\Lambda}_{12})\boldsymbol{\Omega}_{12}^\top,$$

where $\text{diag}(\boldsymbol{\Lambda}_{12})$ denotes a diagonal matrix with diagonal comprising the vector $\boldsymbol{\Lambda}_{12}$. In the display above one recognizes a "ridge regression"-type estimating equation. Solve for $\boldsymbol{\Omega}_{12}^\top$ and arrive at

$$\boldsymbol{\Omega}_{12} = -\boldsymbol{\Omega}_{11}\mathbf{A}_{12}[\,\mathbf{E}^{-1} + \boldsymbol{\Omega}_{11}\text{diag}(\boldsymbol{\Lambda}_{12})]^{-1}. \tag{3}$$

Similarly, singling out $\boldsymbol{\Omega}_{11}$ from the estimating Equation (2)—in which the explicit expression for partitioned $\boldsymbol{\Omega}^{-1}$ has been substituted—gives

$$0 = \boldsymbol{\Omega}_{11}^{-1} + \boldsymbol{\Omega}_{11}^{-2}\boldsymbol{\Omega}_{12}\,\mathbf{E}^{-1}\,\boldsymbol{\Omega}_{12}^{\top} - \mathbf{A}_{11} - \boldsymbol{\Lambda}_{11}\boldsymbol{\Omega}_{11}.$$

In this substitute expression (3) for $\boldsymbol{\Omega}_{12}$ to arrive at $f(\boldsymbol{\Omega}_{11}) = 0$ with

$$\begin{aligned} f(\boldsymbol{\Omega}_{11}) = \boldsymbol{\Omega}_{11}^{-1} &- \mathbf{A}_{11} - \boldsymbol{\Lambda}_{11}\boldsymbol{\Omega}_{11} \\ &+ \mathbf{A}_{12}[\,\mathbf{E}^{-1} + \boldsymbol{\Omega}_{11}\mathrm{diag}(\boldsymbol{\Lambda}_{12})]^{-1}\,\mathbf{E}^{-1} \\ &\times [\mathbf{E}^{-1} + \boldsymbol{\Omega}_{11}\mathrm{diag}(\boldsymbol{\Lambda}_{12})]^{-1}\mathbf{A}_{12}^{\top}. \end{aligned}$$

As $\boldsymbol{\Omega}_{11}$ is a scalar, this can be solved by standard root finding machinery. Note that $\boldsymbol{\Omega}_{11}$ is a diagonal element of $\boldsymbol{\Omega}$ and thereby positive. Moreover, $\lim_{\boldsymbol{\Omega}_{11}\downarrow 0} f(\boldsymbol{\Omega}_{11}) = \infty$, $\lim_{\boldsymbol{\Omega}_{11}\to\infty} f(\boldsymbol{\Omega}_{11}) = -\infty$, and $\partial f(\boldsymbol{\Omega}_{11})/\partial\boldsymbol{\Omega}_{11} < 0$ for $\boldsymbol{\Omega}_{11} > 0$. Together these boundary conditions and f being strictly monotone decreasing on $\mathbb{R}_{>0}$ imply that there is a unique zero. The root search may be sped up when limiting the domain. Hereto note that the last summand that comprises f assumes values in the interval $[0, \mathbf{A}_{12}\mathbf{E}\mathbf{A}_{12}^{\top}]$. Replacing this last summand by its boundary values and solving for $\boldsymbol{\Omega}_{11}$, reveals that the root of $f(\boldsymbol{\Omega}_{11}) = 0$ is in the interval $[\frac{1}{2}\mathbf{A}_{11} + (\boldsymbol{\Lambda}_{11} + \frac{1}{4}\mathbf{A}_{11}^2)^{1/2}]^{-1} < \boldsymbol{\Omega}_{11} < \{\frac{1}{2}(\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{E}\mathbf{A}_{12}^{\top}) + [\boldsymbol{\Lambda}_{11} + \frac{1}{4}(\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{E}\mathbf{A}_{12}^{\top})^2]^{1/2}\}^{-1}$. With an estimate of $\boldsymbol{\Omega}_{11}$ at hand, Equation (3) allows the updating of $\boldsymbol{\Omega}_{12}$. Put together and running over the columns, while initiated by—but other choices possible—$\boldsymbol{\Omega}^{(0)} = \mathbf{T} + \mathbf{S}_d$ (with $\mathbf{S}_d$ diagonal and $\mathrm{diag}(\mathbf{S}_d) = \mathrm{diag}(\mathbf{S})$), this gives rise to Algorithm 1 for solving the generalized ridge precision estimating Equation (2).

A direct implementation of the above may be rather slow: the evaluation of $f(\cdot)$ for any choice of $\boldsymbol{\Omega}_{11}$ requires—at least—the calculation of a $(p-1)\times(p-1)$ matrix inverse. This can be sped up considerably. Hereto define $\mathbf{B} = \mathbf{A}_{12}\mathrm{diag}(\boldsymbol{\Lambda}_{12})^{-1/2}$ and $\mathbf{C} = \mathrm{diag}(\boldsymbol{\Lambda}_{12})^{-1/2}\mathbf{E}^{-1}\mathrm{diag}(\boldsymbol{\Lambda}_{12})^{-1/2}$ and denote the eigen-decomposition of $\mathbf{C}$ by $\mathbf{V}_c\mathbf{D}_c\mathbf{V}_c^{\top}$ where $\mathbf{V}_c$ and $\mathbf{D}_c$ contain the eigen-vectors as columns and -values on the diagonal, respectively. This enables the rewriting of Equation (3) and the function $f(\cdot)$ to

$$\begin{aligned} \boldsymbol{\Omega}_{12} &= -\boldsymbol{\Omega}_{11}\mathbf{B}\mathbf{V}_c(\mathbf{D}_c + \boldsymbol{\Omega}_{11}\mathbf{I}_{pp})^{-1}\mathbf{V}_c\mathrm{diag}(\boldsymbol{\Lambda}_{12})^{-1/2}, \\ f(\boldsymbol{\Omega}_{11}) &= \boldsymbol{\Omega}_{11}^{-1} - \mathbf{A}_{11} - \boldsymbol{\Lambda}_{11}\boldsymbol{\Omega}_{11} \\ &\quad + \mathbf{B}\mathbf{V}_c(\mathbf{D}_c + \boldsymbol{\Omega}_{11}\mathbf{I}_{pp})^{-2}\mathbf{D}_c(\mathbf{B}\mathbf{V}_c)^{\top}, \end{aligned}$$

respectively. In particular, with the eigen-decomposition of $\mathbf{C}$ available, the root of $f(\cdot) = 0$ now requires only a single evaluation of the vector $\mathbf{B}\mathbf{V}_c$, after which $f(\cdot)$ is evaluated for each $\boldsymbol{\Omega}_{11}$ by simple Hadamard multiplications of same-sized vectors. The evaluation of many inverses in the search for the root of $f(\cdot)$ is thus avoided. In addition, this benefits the computation of the suggested upper bound for the root of $f(\cdot)$ as well as $\boldsymbol{\Omega}_{12}$ when having found the root of $f(\boldsymbol{\Omega}_{11}) = 0$.

The resulting generalized ridge precision estimate of $\boldsymbol{\Omega}$ is positive definite.

Proposition 1. Let $\mathbf{T} \succ 0$, that is, $\mathbf{T}$ is positive definite, and $(\boldsymbol{\Lambda})_{j_1,j_2} > 0$ for $j_1, j_2 = 1,\ldots,p$. Then, the generalized ridge inverse covariance estimator, as produced by Algorithm 1, is positive definite.

Proof. The proof shows (i) that, when $\mathbf{E}^{-1} \succ 0$, the updating of a single row/column results in a positive definite updated precision matrix, denoted $\tilde{\boldsymbol{\Omega}}$, and (ii) that, in the updating of the next row/column the redefined submatrix $\mathbf{E}^{-1}$ derived from the updated $\tilde{\boldsymbol{\Omega}}$ is positive definite. To wrap up the proof invoke (i) and (ii) iteratively.

With respect to (i) first note that the initiation $\boldsymbol{\Omega}_{22}^{(0)} = \mathbf{T}_{22} + (\mathbf{S}_d)_{22}$ implies $\mathbf{E}^{-1} = \mathbf{S}_{22} + (\mathbf{S}_d)_{22}$, which ensures $\mathbf{E}^{-1} \succ 0$ due to the fact that $\mathbf{S}_{22} \succeq 0$, that is, $\mathbf{S}_{22}$ is semipositive definite, and $(\mathbf{S}_d)_{22} \succ 0$ (see Harville 2008, Lemma 14.2.4). Now the positive definiteness of $\tilde{\boldsymbol{\Omega}}^{-1}$ is shown. Hereto express the determinant of $\boldsymbol{\Omega}^{-1}$ as the product of determinant of submatrix $(\tilde{\boldsymbol{\Omega}}^{-1})_{22}$ and that of its Schur complement

$$\begin{aligned} |\tilde{\boldsymbol{\Omega}}^{-1}| &= |(\tilde{\boldsymbol{\Omega}}^{-1})_{22}||(\tilde{\boldsymbol{\Omega}}^{-1})_{11} - (\tilde{\boldsymbol{\Omega}}^{-1})_{12}(\tilde{\boldsymbol{\Omega}}^{-1})_{22}(\tilde{\boldsymbol{\Omega}}^{-1})_{12}^{\top}| \\ &= |\mathbf{E}^{-1}||\tilde{\boldsymbol{\Omega}}_{11}^{-1} + \tilde{\mathbf{F}}\mathbf{E}^{-1}\tilde{\mathbf{F}}^{\top} - \tilde{\mathbf{F}}\mathbf{E}^{-1}\mathbf{E}\mathbf{E}^{-1}\tilde{\mathbf{F}}^{\top}| \\ &= |\mathbf{E}^{-1}||\tilde{\boldsymbol{\Omega}}_{11}^{-1}| > 0, \end{aligned}$$

where $\tilde{\mathbf{F}} = \tilde{\boldsymbol{\Omega}}_{11}^{-1}\tilde{\boldsymbol{\Omega}}_{12}$. Consequently, by Corollary 14.8.6 of Harville (2008) $\tilde{\boldsymbol{\Omega}}^{-1} \succ 0$ and so is $\tilde{\boldsymbol{\Omega}}$ as the eigenvalues of the latter are the reciprocal of those of the former.

For (ii) it is shown that any $(p-1)\times(p-1)$ submatrix $\mathbf{E}^{-1}$ defined by $\mathbf{A}_{22} + \boldsymbol{\Lambda}_{22} \circ \tilde{\boldsymbol{\Omega}}_{22}$ is positive definite. From the estimating Equation (2) note that $\mathbf{A}_{22} + \boldsymbol{\Lambda}_{22} \circ \tilde{\boldsymbol{\Omega}}_{22} = (\tilde{\boldsymbol{\Omega}}^{-1})_{22}$. Furthermore, the positive definiteness of $\tilde{\boldsymbol{\Omega}}^{-1} \succ 0$ carries over to its principal submatrix (Harville 2008, Corollary 14.2.14) and thereby to the new $\mathbf{E}^{-1}$ derived from it. □

The limiting (in $\boldsymbol{\Lambda}$) joint shrinkage behavior of the resulting generalized ridge precision estimate of $\boldsymbol{\Omega}$ is as expected: it converges to $\mathbf{T}$.

Algorithm 1 Generalized ridge precision estimation

Require: $\mathbf{S}$; $\mathbf{T}$; ▷ Data; Target matrix;
$\boldsymbol{\Lambda}$; $\widehat{\boldsymbol{\Omega}}^{(0)}$; ▷ Penalty parameter matrix; Init. estimate;
K; ε. ▷ Max. # iteration; Max. succ. diff.

Set $\boldsymbol{\Omega}^{(0)} = \mathbf{T} + \mathbf{S}_d$. ▷ Initiate
for $k = 1$ to K **do** ▷ Iterate
  for $j = 1$ to p **do** ▷ Run over rows of $\boldsymbol{\Omega}$
    ∘ Find $\widehat{\boldsymbol{\Omega}}_{jj}^{(k)}$ as the positive root of
    $f(\boldsymbol{\Omega}_{jj}^{(k)}) = 0$ using latest estimate of $\boldsymbol{\Omega}_{\backslash j,\backslash j}$.
    ∘ Calculate $\widehat{\boldsymbol{\Omega}}_{j,\backslash j}^{(k)}$ from Equation (3) using
    $\widehat{\boldsymbol{\Omega}}_{jj}^{(k)}$ and latest estimate of $\boldsymbol{\Omega}_{\backslash j,\backslash j}$.
    ∘ Set $\widehat{\boldsymbol{\Omega}}_{j,\backslash j}^{(k)}$ equal to $(\widehat{\boldsymbol{\Omega}}_{j,\backslash j}^{(k)})^{\top}$.
  end for
  if $\|\widehat{\boldsymbol{\Omega}}^{(k)} - \widehat{\boldsymbol{\Omega}}^{(k-1)}\|_F^2 < \varepsilon$ **then** ▷ Convergence Assessment
    `terminate`
  end if
end for
return $\widehat{\boldsymbol{\Omega}}^{(k)}$

Proposition 2. Define $\lambda_{\min} = \min_{j_1,j_2}(\boldsymbol{\Lambda})_{j_1,j_2}$. Then: $\lim_{\lambda_{\min}\to\infty} \widehat{\boldsymbol{\Omega}}(\boldsymbol{\Lambda}) = \mathbf{T}$.

Proof. Define $\lambda_{\max} = \max_{j_1,j_2}(\boldsymbol{\Lambda})_{j_1,j_2}$ and let $\lambda_0 \in \mathbb{R}_{>0}$ be such that$\lambda_{\max} = \lambda_{\min} + \lambda_0$. By construction of the estimators, it holds that

$$\begin{aligned}
&\mathcal{L}[\mathbf{S}, \widehat{\boldsymbol{\Omega}}(\lambda_{\min}\mathbf{1}_{pp})] - \lambda_{\min}\|\widehat{\boldsymbol{\Omega}}(\lambda_{\min}\mathbf{1}_{pp}) - \mathbf{T}\|_F^2 \\
&\quad \geq \mathcal{L}[\mathbf{S}, \widehat{\boldsymbol{\Omega}}(\boldsymbol{\Lambda})] - \|\sqrt{\boldsymbol{\Lambda}} \circ [\widehat{\boldsymbol{\Omega}}(\boldsymbol{\Lambda}) - \mathbf{T}]\|_F^2 \\
&\quad \geq \mathcal{L}[\mathbf{S}, \widehat{\boldsymbol{\Omega}}(\lambda_{\max}\mathbf{1}_{pp})] - \lambda_{\max}\|\widehat{\boldsymbol{\Omega}}(\lambda_{\max}\mathbf{1}_{pp}) - \mathbf{T}\|_F^2 \\
&\quad = \mathcal{L}[\mathbf{S}, \widehat{\boldsymbol{\Omega}}(\lambda_{\max}\mathbf{1}_{pp})] - \lambda_{\min}\|\widehat{\boldsymbol{\Omega}}(\lambda_{\max}\mathbf{1}_{pp}) - \mathbf{T}\|_F^2 \\
&\qquad -\lambda_0\|\widehat{\boldsymbol{\Omega}}(\lambda_{\max}\mathbf{1}_{pp}) - \mathbf{T}\|_F^2.
\end{aligned}$$

Divide the above display by $\lambda_{\min}$, subsequently let $\lambda_{\min}$ tend to infinity, and use Proposition 1 of van Wieringen and Peeters (2016) (which states that $\lim_{\lambda\to\infty} \boldsymbol{\Omega}(\lambda\mathbf{I}_{pp}) = \mathbf{T}$) to conclude that

$$\begin{aligned}
0 &= \lim_{\lambda_{\min}\to\infty} \|\widehat{\boldsymbol{\Omega}}(\lambda_{\min}\mathbf{1}_{pp}) - \mathbf{T}\|_F^2 \\
&\geq \lim_{\lambda_{\min}\to\infty} \|\sqrt{\boldsymbol{\Lambda}} \circ [\widehat{\boldsymbol{\Omega}}(\boldsymbol{\Lambda}) - \mathbf{T}]\|_F^2 \\
&\geq \lim_{\lambda_{\min}\to\infty} \|\widehat{\boldsymbol{\Omega}}[\lambda_{\min} + \lambda_0)\mathbf{1}_{pp}] - \mathbf{T}\|_F^2 \;=\; 0.
\end{aligned}$$

From this the claimed limit follows. □

The element-wise shrinkage limit, $(\boldsymbol{\Lambda})_{j_1,j_2} \to \infty$ while $(\boldsymbol{\Lambda})_{j_3,j_4} < \infty$ for all $(j_3,j_4) \neq (j_1,j_2)$, of an element of $\widehat{\boldsymbol{\Omega}}(\boldsymbol{\Lambda})$ is sketched, based on numerical (not shown) and analytic explorations with convenient parameter choices. Intuitively, it is clear that element-wise shrinkage forces an element of the generalized ridge precision estimator to the corresponding element of the target matrix $\mathbf{T}$. However, element-wise shrinkage affects other elements of the generalized ridge precision estimator, with shrinkage of either diagonal or off-diagonal elements bearing different consequences. First, concentrate on the element-wise shrinkage of a diagonal element. Consider $\mathbf{S} = (1-\rho)\mathbf{I}_{pp} + \rho\mathbf{1}_{pp}$ with $\rho \in (-1/p, 1)$ and $\mathbf{T} = \alpha\mathbf{I}_{pp}$ with $0 < \alpha \ll 1$. Then, $[\boldsymbol{\Omega}(\boldsymbol{\Lambda})]_{jj}$ shrinks to α as $(\boldsymbol{\Lambda})_{jj} \to \infty$ while the other elements of $\boldsymbol{\Lambda}$ are kept fixed. The off-diagonal elements in the corresponding row and column will then shrink to virtually zero, as the resulting estimate would otherwise possibly contradict Proposition 1. This effect can—with the specified parameter choices substituted—also be deduced analytically from the estimating equations of $\boldsymbol{\Omega}_{11}$ and $\boldsymbol{\Omega}_{12}$. Element-wise shrinkage of an off-diagonal element affects the other off-diagonal elements that share the same row or column. With the particular parameter choices above they are shrunken to zero. The diagonal element, however, is affected only to a much lesser degree. For general parameter choices element-wise shrinkage still affects other elements of $\widehat{\boldsymbol{\Omega}}(\boldsymbol{\Lambda})$ but effects are more intricated.

The proposed estimator is consistent, in the traditional sense with fixed dimension p and the sample size n tending to infinity. Temporarily add the subscript n to the penalty parameter and estimator, that is, $\boldsymbol{\Lambda}_n$ and $\boldsymbol{\Omega}_n(\boldsymbol{\Lambda}_n)$, to explicate their sample size dependence.

Proposition 3. Let the entries of $\boldsymbol{\Lambda}_n$ converge (in probability) to zero as n tends to infinity. Then: $\widehat{\boldsymbol{\Omega}}_n(\boldsymbol{\Lambda}_n) \xrightarrow{P} \boldsymbol{\Omega}$ as $n \to \infty$.

The proof of Proposition 3 invokes Theorem 5.7 of van der Vaart (2000), specifying sufficient conditions for consistency, and it is left to verify the conditions of the theorem. The latter is done in Aflakparast, de Gunst, and van Wieringen (2018) for a different but related model, and here the proof is almost analogous and therefore omitted.

Theoretical results on the validity of the assumption of $\boldsymbol{\Lambda}_n$ entry-wise tending to zero (in probability) are nonexistent. However, van Wieringen and Peeters (2016) showed numerically and for $\boldsymbol{\Lambda} = \lambda\mathbf{1}_{pp}$ that—when λ is chosen by leave-one-out cross-validation—the assumption in unproblematic. In the supplementary materials, the tenability of the assumption is evaluated for the generalized ridge precision estimator with a penalty matrix parameterized by either two or three parameters. The results of this additional study corroborate with the assumption.

3. Relation to Other Precision Estimators

The generalized ridge estimator is well-suited to evaluate other precision estimators that maximize the log-likelihood augmented with a concave penalty. This is due to the quadratic form of the generalized ridge penalty. Its two parameters, $\mathbf{T}$ and $\boldsymbol{\Lambda}$, can be chosen such that it provides an approximation to any convex penalty. The resulting generalized estimator is then an approximation to the sought-for one. Should this approximation be unsatisfactory, it may be improved upon. This requires the construction of a sequence of generalized ridge loss functions that converges to the desired concavely penalized log-likelihood. The corresponding sequence of generalized ridge precision estimators then converges to the desired estimator. This is illustrated in the next two subsections for the existing shrinkage estimator of Ledoit and Wolf (2004) and to the generalized version of the graphical lasso and elastic net precision estimators.

3.1. Linear Shrinkage

The generalized ridge precision estimator may be used to obtain the inverse of the linear shrunken covariance estimator of Ledoit and Wolf (2004). This requires to view the latter as a penalized precision estimator. Hereto consider a precision estimator defined as the maximum of a loss function related to (1) but with the generalized ridge penalty replaced by a concave function of the same form as the log-likelihood it penalizes

$$\mathcal{L}(\mathbf{S}; \boldsymbol{\Omega}) - \lambda\mathcal{L}(\mathbf{T}^{-1}; \boldsymbol{\Omega}). \tag{4}$$

This loss function is maximized by $\widehat{\boldsymbol{\Omega}}(\nu) = [(1-\nu)\mathbf{S} + \nu\mathbf{T}^{-1}]^{-1}$ with $\nu = \lambda(1+\lambda)^{-1} \in (0,1)$, which coincides with the inverse of the shrunken covariance estimator of Ledoit and Wolf (2004). The connection between the two precision estimators becomes apparent from the quadratic approximation—around an initial precision estimate $\boldsymbol{\Omega}_0$—of the second term, the penalty, of loss

function (4)

$$\begin{aligned}
&\mathcal{L}(\mathbf{S},\boldsymbol{\Omega}) - \lambda\mathcal{L}(\mathbf{T}^{-1},\boldsymbol{\Omega}) \\
&\quad\approx \mathcal{L}(\mathbf{S},\boldsymbol{\Omega}) - \lambda\mathcal{L}(\mathbf{T}^{-1},\boldsymbol{\Omega}_0) \\
&\qquad -\lambda[\text{vec}(\boldsymbol{\Omega}_0^{-1} - \mathbf{T}^{-1})]^\top \text{vec}(\boldsymbol{\Omega} - \boldsymbol{\Omega}_0) \\
&\qquad -\tfrac{1}{2}\lambda[\text{vec}(\boldsymbol{\Omega} - \boldsymbol{\Omega_0})]^\top(\boldsymbol{\Omega}_0^{-1}\otimes\boldsymbol{\Omega}_0^{-1})\text{vec}(\boldsymbol{\Omega} - \boldsymbol{\Omega}_0) \\
&\quad\propto \mathcal{L}(\mathbf{S},\boldsymbol{\Omega}) - \tfrac{1}{2}\lambda\left\|\sqrt{\boldsymbol{\Omega}_0^{-1}\otimes\boldsymbol{\Omega}_0^{-1}}\circ(\boldsymbol{\Omega} - \check{\mathbf{T}})\right\|_F^2,
\end{aligned}$$

where $\check{\mathbf{T}} = \boldsymbol{\Omega}_0 - [1/(\boldsymbol{\Omega}_0^{-1}\otimes\boldsymbol{\Omega}_0^{-1})]\circ(\boldsymbol{\Omega}_0^{-1} - \mathbf{T}^{-1})\}$ in which $1/\mathbf{A}$ denotes the Hadamard (element-wise) inverse. This is exactly a generalized ridge precision estimator. A similar algorithm to that described in the next section for the generalized lasso precision estimator may now be conceived, but with the target also updated at each iteration. It is omitted here as the connection presented above is only of theoretical interest as the latter comes as an analytic expression (of which a computationally fast implementation is readily obtained), but it shows the wide applicability of the generalized ridge precision estimator.

3.2. Graphical Lasso and Elastic Net

A sequence of generalized ridge estimators can also be constructed for the evaluation of estimators based on concave, but not strict, penalties that are not continuously differentiable, for example, the graphical lasso. This is illustrated for a generalized version of the graphical lasso that maximizes the log-likelihood augmented with $\|\boldsymbol{\Lambda}\circ(\boldsymbol{\Omega} - \mathbf{T})\|_1$. With a nonzero target matrix $\mathbf{T}$ the generalized graphical lasso no longer shrinks the elements of $\boldsymbol{\Omega}$ to zero. This estimator therefore does not perform selection in the classical sense. Instead, it assesses whether the data give rise to adopt a novel (estimated) parameter value for a particular precision element, or to stick to the original one supplied through $\mathbf{T}$. This touches upon another point: the (sequence of) generalized ridge precision estimator, which—like the inverse of the shrinkage covariance estimators of Ledoit and Wolf (2004)—is a shrinkage estimator, is able to approximate one that exhibits different behavior. The former shrinks the eigenvalues, and with a nonzero target also alters the eigenvectors, of the sample covariance matrix. The generalized graphical lasso estimator shrinks too, although elements-wise, and thereby the eigenvalues. Generally, this implies less shrinkage than its ridge counterpart. But the estimators differ in the potential of the generalized graphical lasso estimator to select, which its ridge counterpart does not. As such, this section illustrates the potential of the generalized ridge precision estimator to evaluate (iteratively) other precision estimators that exhibit selection properties. This selection property also delineates the use of the generalized graphical lasso estimator. It, in combination with a nonzero $\mathbf{T}$, seems most relevant when it is reasonable to assume that—up to a few elements—most knowledge of $\mathbf{T}$ is correct. A practical situation for its application may be conceived when experimenting with a well-studied model system for which an accurate quantitative description is available (in the form of $\mathbf{T}$) and a perturbation of the system is expected to affect the system only locally (a few elements of $\mathbf{T}$). Otherwise, when knowledge of $\mathbf{T}$ is considered to be more vague and viewed as a suggestion, in the sense that no element is assumed to be known exactly, the generalized ridge precision estimator seems preferable.

The algorithm for the generalized ridge precision estimator may be used to approximate the generalized graphical lasso precision estimator, which now too shrinks element-wise to a target matrix. This approximation exploits a quadratic (i.e., ridge) approximation to the absolute value (i.e., lasso) function : $|x| \approx |x_0| + \frac{1}{2|x_0|}(x^2 - x_0^2)$ with x_0 an initial value for x. The lasso estimator may then be found by iteratively optimizing the log-likelihood augmented with this approximate lasso penalty using the previous estimate as initial guess (as suggested in Fan and Li 2001).

The "ridge" approximation to the generalized graphical lasso penalty of $\boldsymbol{\Omega}$, using $\boldsymbol{\Omega}^{(0)}$ as initial guess, then becomes

$$\begin{aligned}
\|\boldsymbol{\Lambda}\circ(\boldsymbol{\Omega} - \mathbf{T})\|_1 \approx{}& \|\boldsymbol{\Lambda}\circ(\boldsymbol{\Omega}^{(0)} - \mathbf{T})\|_1 \\
&+ \tfrac{1}{2}\|\sqrt{\check{\boldsymbol{\Lambda}}(\boldsymbol{\Omega}^{(0)})}\circ(\boldsymbol{\Omega} - \mathbf{T})\|_F^2 \\
&- \tfrac{1}{2}\|\sqrt{\check{\boldsymbol{\Lambda}}(\boldsymbol{\Omega}^{(0)})}\circ(\boldsymbol{\Omega}^{(0)} - \mathbf{T})\|_F^2,
\end{aligned}$$

where the modified penalty matrix $\check{\boldsymbol{\Lambda}}$ is element-wisely defined through $[\tilde{\boldsymbol{\Lambda}}(\boldsymbol{\Omega}^{(0)})]_{j_1,j_2} = [(\boldsymbol{\Lambda})_{j_1,j_2}]\ \{|(\boldsymbol{\Omega}^{(0)} - \mathbf{T})_{j_1,j_2}|\}^{-1}$ for $j_1, j_2 = 1,\ldots,p$. When $\mathbf{T} = \mathbf{0}_{pp}$ this reduces to an approximation of the graphical lasso penalty. When defining the generalized graphical lasso estimator as the maximum of

$$\log(|\boldsymbol{\Omega}|) - \text{tr}(\mathbf{S}\boldsymbol{\Omega}) - \|\boldsymbol{\Lambda}\circ(\boldsymbol{\Omega} - \mathbf{T})\|_1, \tag{5}$$

the estimator may be found by iterative application of the Algorithm 1. Replacing the generalized graphical lasso penalty by the "ridge" approximation yields the approximated loss which is proportional to

$$\log(|\boldsymbol{\Omega}|) - \text{tr}(\mathbf{S}\boldsymbol{\Omega}) - \tfrac{1}{2}\left\|\sqrt{\check{\boldsymbol{\Lambda}}(\boldsymbol{\Omega}^{(0)})}\circ(\boldsymbol{\Omega} - \mathbf{T})\right\|_F^2,$$

with $\boldsymbol{\Omega}^{(0)}$ the initial guess for $\boldsymbol{\Omega}$. The function in the last display is the loss function of the generalized ridge estimator. Hence, iteratively applying Algorithm 1 with the previous estimate of $\boldsymbol{\Omega}$ as initial guess converges to the maximum of loss function (5) and, thus, to the generalized graphical lasso estimator.

Algorithm 2 Generalized graphical lasso

Require: $\mathbf{S}$; $\mathbf{T}$; ▷ Data; Target matrix;
$\boldsymbol{\Lambda}$; $\widehat{\boldsymbol{\Omega}}^{(0)}$; ▷ Penalty parameter matrix; Init. estimate;
K; ε. ▷ Max. # iteration; Max. succ. diff.

for $k = 1$ to K **do** ▷ Iterate
∘ Calculate $\check{\boldsymbol{\Lambda}}(\widehat{\boldsymbol{\Omega}}^{(k-1)})$.
∘ Obtain $\widehat{\boldsymbol{\Omega}}^{(k)}$ from algorithm 1
using $\boldsymbol{\Lambda} = \check{\boldsymbol{\Lambda}}(\widehat{\boldsymbol{\Omega}}^{(k-1)})$.
if $\|\widehat{\boldsymbol{\Omega}}^{(k)} - \widehat{\boldsymbol{\Omega}}^{(k-1)}\|_F^2 < \varepsilon$ **then** ▷ Convergence Assessment
`terminate`
end if
end for
return $\widehat{\boldsymbol{\Omega}}^{(k)}$

Of course, other algorithms—than the iterative one presented above (Algorithm 2)—for the numerical evaluation of the generalized graphical lasso may be conceived (see the supplementary materials).

It is now a small step to obtain an approximation of the generalized graphical elastic net estimator of the inverse covariance estimator. This estimator maximizes the log-likelihood augmented with both a ridge and lasso penalty. This linear combination of penalties, referred to as the elastic net penalty, may—as before—be approximated by a ridge penalty through a quadratic approximation of the lasso penalty term

$$\begin{aligned}&\|\boldsymbol{\Lambda}_1 \circ (\boldsymbol{\Omega} - \mathbf{T})\|_1 + \tfrac{1}{2}\|\sqrt{\boldsymbol{\Lambda}_2} \circ (\boldsymbol{\Omega} - \mathbf{T})\|_F^2 \\ &\approx \tfrac{1}{2}\|\sqrt{\boldsymbol{\Lambda}_2 + \check{\boldsymbol{\Lambda}}_1(\boldsymbol{\Omega}^{(0)})} \circ (\boldsymbol{\Omega} - \mathbf{T})\|_F^2,\end{aligned}$$

where $\boldsymbol{\Lambda}_1$ and $\boldsymbol{\Lambda}_2$ temporarily denote the ℓ_1 and ℓ_2 penalty parameters and terms not involving $\boldsymbol{\Omega}$ have been dropped. Algorithm 2 may now be modified correspondingly to yield an approximation of the generalized graphical elastic net estimator of $\boldsymbol{\Omega}$.

4. In Silico Studies

In this section, the quality of the approximation of existing precision estimators by a sequence of ridge precision estimators is studied in silico on the statistical computing platform R (R Core Team 2018). This is followed by a simulation study that compares performance of the generalized ridge estimator to its regular counterpart to identify cases where the former may be preferred over the latter.

4.1. Benchmarking

The proposed algorithms (1) and (2) are benchmarked against those which they generalize. This comprises the following comparisons:

- The generalized ridge algorithm (1) versus the implementation of the original and known chordal support ridge precision estimators (see van Wieringen and Peeters (2016) and Miok, Wilting, and van Wieringen (2017), respectively) both available through the `rags2ridges`-package (Peeters, Bilgrau, and van Wieringen 2017). The former, with either $\boldsymbol{\Lambda} = \lambda \mathbf{1}_{pp}$ or a penalty matrix $\boldsymbol{\Lambda}$ with entries equaling λ or ∞ for nonzero and zero elements, respectively, ought to yield the same estimate as the latter two. In both cases $\mathbf{T} = \mathbf{0}_{pp}$ is used.
- The generalized graphical lasso algorithm (2) versus the graphical lasso. The latter has been implemented in the `glasso`-package (Friedman, Hastie, and Tibshirani 2014). The estimates of both algorithms should yield the same estimate when $\boldsymbol{\Lambda}$ is set equal to $\lambda \mathbf{1}_{pp}$ in the former. Again $\mathbf{T} = \mathbf{0}_{pp}$ is used.

The algorithms are benchmarked in terms of accuracy and computing time. Accuracy is defined as the Frobenius and supremum norm ℓ_∞ of the (vectorized) difference between two corresponding estimates.

Benchmarking is done with the following settings. The data dimension p and the sample size n are set equal to 10, 50, 100 in accordance with a full factorial design. The data are then sampled from a zero-centered multivariate Gaussian with either a banded, blocked, full or "hub" precision matrix $\boldsymbol{\Omega}$ (see supplementary materials for the specifics of their parameterizations and in part constructed using the `Matrix`-package, Bates and Maechler 2018) using the `mvtnorm`-package (Genz et al. 2018). From each data draw all estimates are calculated using a unit penalty parameter throughout, except for known zero precision entries that demanded a large penalty ($\lambda = 10^{10}$) to force an estimate close to zero. The Frobenius and supremum accuracy are evaluated from these estimates. This is repeated one hundred times and summarized by the median. The computing time is evaluated using the `microbenchmark`-package (Mersmann 2015). This takes a single data draw from which the estimators are evaluated one hundred times in random order while measuring their computing time. These times are summarized by their median.

The tables in the supplementary materials contain the accuracy results of the benchmarking. Both the Frobenius and supremum accuracy of the generalized ridge precision algorithm (1) are excellent for the regular ridge precision estimate while very good for its counterpart with known chordal support. The accuracy difference between the two estimates is due to the convergence error in the numerical optimization of the loss function of the known chordal support ridge precision estimate, which is absent in the regular ridge precision estimate as it results from the direct numerical evaluation of an analytic expression. Turning to the accuracies of the generalized graphical lasso algorithm (2) they are good, but less than those for the ridge estimators. In part this too is due to the convergence error of the graphical lasso algorithm itself, but in addition to the error introduced by the quadratic approximation of the lasso penalty function. Finally, the achieved accuracies depend on the chosen convergence tolerances and may be improved upon when setting tighter tolerances for the involved algorithms. These observations are consistent over the various choices of n and p and that of the precision matrix.

The computing time of the proposed algorithms falls behind its competitors (results not shown). This is in line with expectation as the latter have been heavily optimized for these particular cases. But the proposed algorithms have the advantage of being much wider applicable.

4.2. Comparison

Two simulations have been conducted to contrast the generalized ridge precision estimator to its regular counterpart. The regular ridge and generalized ridge precision estimators both shrink to a target matrix but where the former shrinks all elements at equal rate the latter allows for differentiation in these rates. Hence, the simulations focus on how this differentiation may benefit the estimation. In the first simulation this differentiation allows to exploit knowledge on the structure of the precision matrix, while in the second simulation it facilitates incorporation of perfect knowledge on part of the precision matrix.

The simulations are set-up as in Section 4.1, with data drawn from the zero-mean multivariate normal distribution, same

sample sizes, dimensions and number of runs, while the Frobenious loss is used to evaluate the performance of both estimators. The simulations differ in the employed precision, target, and penalty matrix: $\boldsymbol{\Omega}$, $\mathbf{T}$, and $\boldsymbol{\Lambda}$, respectively (with full details on these matrices in the supplementary materials). In the first simulation the $\boldsymbol{\Omega}$ has three bands, while $\mathbf{T}$ matrix has a single band and its nonzero elements are different from their counterparts in $\boldsymbol{\Omega}$. The generalized ridge precision estimator uses a penalty matrix with $(\boldsymbol{\Lambda})_{j,j'} = (|j - j'| + 1)\lambda$ for $j, j' = 1, \ldots, p$. This penalizes bands further from the diagonal more than those close by. As such it encourages a banded precision estimate. In the second simulation $\boldsymbol{\Omega}$ is full and $\mathbf{T}$ comprises only zeros except for the first row and column that are set equal to those of $\boldsymbol{\Omega}$. For the generalized ridge precision estimator $\boldsymbol{\Lambda}$ is parametrized: $(\boldsymbol{\Lambda})_{1,j} = 10^{10} = (\boldsymbol{\Lambda})_{j,1}$ for $j = 1, \ldots, p$, and $(\boldsymbol{\Lambda})_{j,j'} = \lambda$ for all $j, j' = 2, \ldots, p$. By the use of this $\boldsymbol{\Lambda}$ the precision matrix is thus considered to be partially known. The free penalty parameter of the ridge and generalized precision estimators is chosen by 5-fold cross-validation. In the second simulation the contribution of the elements in the first row and column are excluded from the Frobenius loss.

Histograms of the loss differences, $\|\widehat{\boldsymbol{\Omega}}_{\text{gen}}(\boldsymbol{\Lambda}, \mathbf{T}) - \boldsymbol{\Omega}\|_F - \|\widehat{\boldsymbol{\Omega}}_{\text{reg}}(\lambda, \mathbf{T}) - \boldsymbol{\Omega}\|_F$, for all (n, p)-combinations are shown in the supplementary materials. In all histograms, of both simulations, the bulk—if not all—of the loss differences is located in the negative part of the real line. Hence, the loss of the generalized ridge precision estimator is generally superior over that of its regular counterpart. In particular, in the high-dimensional setting the former profits even more the inclusion of imprecise knowledge on the structure of or precise information on part of the precision matrix.

5. Choice of $\boldsymbol{\Lambda}$

The proposed generalized estimator depends on a penalty parameter $\boldsymbol{\Lambda}$ with $\frac{1}{2}p(p+1)$ degrees of freedom. Even for relatively small p this is too prohibitively large to be of practical use, as for each nonredundant entry of $\boldsymbol{\Lambda}$ an informed choice—possibly derived from the data at hand—is required. Theoretically, this flexible penalty parameter may improve on the result of van Wieringen (2017) that shows the existence of a penalty parameter that yields a mean squared error of the regular ridge precision estimator of van Wieringen and Peeters (2016) over that of its maximum likelihood counterpart.

The practical value of the generalized graphical ridge estimator, however, is in the specific cases with a low-dimensionally parametrized $\boldsymbol{\Lambda}$ that it encompasses. Among others these comprise:

- The estimator allows for a variate-specific rather than element-wise penalization. To this end the penalty can be rewritten as $\text{tr}[(\boldsymbol{\Omega} - \mathbf{T})\boldsymbol{\Lambda}_d(\boldsymbol{\Omega} - \mathbf{T})]$ for $\boldsymbol{\Lambda}_d$ diagonal. Such an approach would be appropriate when the random vector $\mathbf{Y}_i$ comprises multiple groups of variates that need to be penalized separately, for instance due to scale differences.
- The diagonal elements of the precision matrix are left unpenalized when the diagonal elements of $\boldsymbol{\Lambda}$ are set equal to—virtually—zero: $0 < \boldsymbol{\Lambda}_{jj} \ll 1$ for all j. This ensures that, for large values of the off-diagonal elements of the penalty parameter $\boldsymbol{\Lambda}$, the marginal variances (i.e., the reciprocal of diagonal elements of the precision) are—almost—left unshrunken, and thus match those of the sample covariance matrix.
- Information on the support of the off-diagonal elements of the precision matrix $\boldsymbol{\Omega}$ may be incorporated in the estimator through very large values of the corresponding elements of the penalty parameter $\boldsymbol{\Lambda}$. This effectively shrinks these elements of $\boldsymbol{\Omega}$ to zero.
- Parts of the precision matrix may be penalized differently w.r.t. a target matrix $\mathbf{T}$. This may be relevant when only a subset of the variates has been observed in a pilot experiment, which has been used to form a quantitative suggestion for a submatrix of $\boldsymbol{\Omega}$ and is included in $\mathbf{T}$.

The above are immediate generalizations of the regular ridge precision estimator of van Wieringen and Peeters (2016) that serve a practical end.

For a low-dimensional parameterization of $\boldsymbol{\Lambda}$ standard penalty parameter selection methods may be used. Cross-validation is a viable choice as it can be applied irrespective of the particulars of the parameterization of $\boldsymbol{\Lambda}$. Cross-validation splits the sample into equally sized subsets that one-by-one are left out, while the remaining subsets are used to estimate $\boldsymbol{\Omega}$ for a given choice of the penalty parameters, which has its performance then evaluated on the left-out subset. This is done for a set of penalty parameter values. The optimal penalty parameter value then yields the best performance.

A grid search of the penalty parameter space quickly becomes infeasible for penalty matrices $\boldsymbol{\Lambda}$ parameterized by more than two parameters. Feng and Simon (2018) then suggest to use gradient ascent, which is outlined for the case at hand. Let the penalty matrix by parametrized by a m-dimensional vector $\boldsymbol{\lambda}$, that is, $\boldsymbol{\Lambda} = \boldsymbol{\Lambda}(\boldsymbol{\lambda})$. The vector $\boldsymbol{\lambda}$ is then chosen to maximize the K-fold cross-validated log-likelihood

$$\begin{aligned}\boldsymbol{\lambda}_{\text{opt}} &= \arg\max_{\boldsymbol{\lambda} \in \mathbb{R}^m_+} \mathcal{L}_{\text{kCV}}(\mathbf{Y}; \boldsymbol{\Lambda}) \\ &:= \arg\max_{\boldsymbol{\lambda} \in \mathbb{R}^m_+} \sum_{k=1}^{K} \log[|\widehat{\boldsymbol{\Omega}}_{-k}(\boldsymbol{\Lambda})|] - \text{tr}[\mathbf{S}_k \widehat{\boldsymbol{\Omega}}_{-k}(\boldsymbol{\Lambda})],\end{aligned}$$

where $\mathbf{S}_k$ is the sample covariance matrix of the kth subset of samples and $\widehat{\boldsymbol{\Omega}}_{-k}(\boldsymbol{\Lambda})$ is the generalized ridge precision estimate derived from all but the kth subset. Gradient ascent, starting from some initial value, approaches the maximizer in step-wise fashion. At each step the direction of the maximum increase of the K-fold cross-validated log-likelihood is determined and the penalty parameter is updated by an increment in this direction. For the uth update of $\boldsymbol{\lambda}_{\text{opt}}$ this amounts to (see Feng and Simon 2018)

$$\begin{aligned}\boldsymbol{\lambda}^{(u+1)}_{\text{opt}} &= \boldsymbol{\lambda}^{(u)}_{\text{opt}} - t^{(u)} \nabla_{\boldsymbol{\lambda}} \mathcal{L}_{\text{kCV}}(\mathbf{Y}; \boldsymbol{\Lambda})\big|_{\boldsymbol{\lambda} = \boldsymbol{\lambda}^{(u)}_{\text{opt}}} \\ &= \boldsymbol{\lambda}^{(u)}_{\text{opt}} - t^{(u)} \sum_{k=1}^{K} \{\nabla_{\boldsymbol{\lambda}} [\widehat{\boldsymbol{\Omega}}_{-k}(\boldsymbol{\Lambda})]\}^{\top} \text{vec}[\widehat{\boldsymbol{\Omega}}_{-k}(\boldsymbol{\Lambda}) - \mathbf{S}_k],\end{aligned}$$

with step size $t^{(u)}$ and (with a slight abuse of notation)

$$\begin{aligned}\nabla_{\boldsymbol{\lambda}}[\widehat{\boldsymbol{\Omega}}_{-k}(\boldsymbol{\Lambda})] &= \{\text{vec}[\tfrac{\partial}{\partial\lambda_1}\widehat{\boldsymbol{\Omega}}_{-k}(\boldsymbol{\Lambda})],\ldots,\text{vec}[\tfrac{\partial}{\partial\lambda_m}\widehat{\boldsymbol{\Omega}}_{-k}(\boldsymbol{\Lambda})]\}\\ &= \{[\widehat{\boldsymbol{\Omega}}_{-k}(\boldsymbol{\Lambda})]^{-1}\otimes[\widehat{\boldsymbol{\Omega}}_{-k}(\boldsymbol{\Lambda})]^{-1} + \text{diag}[\text{vec}(\boldsymbol{\Lambda})]\}^{-1}\\ &\quad\times\big[\nabla_{\boldsymbol{\lambda}}\boldsymbol{\Lambda}\circ\big(\mathbf{1}_m^\top\otimes\{\text{vec}(\mathbf{T})-\text{vec}[\widehat{\boldsymbol{\Omega}}_{-k}(\boldsymbol{\Lambda})]\}\big)\big],\end{aligned}$$

in which $\nabla_{\boldsymbol{\lambda}}\boldsymbol{\Lambda} = [\text{vec}(\frac{\partial}{\partial\lambda_1}\boldsymbol{\Lambda}),\ldots,\text{vec}(\frac{\partial}{\partial\lambda_m}\boldsymbol{\Lambda})]$. Updating is repeated until a stopping criterion is satisfied. For the evaluation of $\nabla_{\boldsymbol{\lambda}}[\widehat{\boldsymbol{\Omega}}_{-k}(\boldsymbol{\Lambda})]$ the inverse of the $p^2\times p^2$ can be avoided by direct evaluation of this gradient through the use of the conjugate gradient algorithm, which makes the gradient ascent approach computationally manageable.

6. Application

The potential of the proposed estimator is illustrated by graphical modeling of the molecular regulatory network in metastasized tumors. With early detection being the most successful strategy against cancer, omics data from primary cancer tumors is generally more widely available than that of metastasized tumors. But as the primary cancer is a precursor stage of metastasis, tumors of both stages are expected to share some resemblance. The larger abundance of molecular information from cancer samples may be exploited to produce a suggestion of the regulatory network active in metastasized tumors, which can subsequently be used in the reconstruction of the regulatory network from the scarcer molecular metastasis data.

The illustration is a reanalysis of part of the data from the The Cancer Genome Atlas (TCGA) lung squamous cell carcinoma study as presented by Cancer Genome Atlas Research Network (2012). This TCGA study comprises 240 lung squamous cell carcinomas that are (among others) characterized genomically by array Comparative Genomic Hybridization (aCGH) and transcriptomically by RNA-sequencing (RNAseq). Preprocessed genomic and transcriptomic data are publicly available and downloaded using the `TCGA2DATA`-package (Wan et al. 2015). Further manipulation of the downloaded data comprised (i) limiting both molecular datasets to overlapping samples and molecular entities; (ii) removal of samples without stage information; (iii) grouping of the samples into two classes, those with lung tumor tissue originating from the lungs and those with lung tumor tissue from beyond (adjacent lymph nodes and other parts of the body but the lung). In the remainder we dub these two groups (primary) cancer and metastasis; (iv) subsetting the omics data to those genes that map to the toll-like receptor signaling pathway (as defined by KEGG, Ogata et al. 1999, using the `KEGG.db`-package, Carlson 2016, and the `biomaRt`-package, Durinck et al. 2009), a pathway known to be implicated in cancer and metastatis (Rakoff-Nahoum and Medzhitov 2009); (v) removal of genes with no expression (i.e., a zero count) in more than 50% of the samples. Effectively, this excludes the IFN (interferon) gene-family, a tumor suppressor genes, known to be often silenced/lost/deleted in lung cancers (Sato, Nakamura, and Tsuchiya 1994); (vi) Gaussianization, an operation that does not affect the conditional independence structure (Liu, Lafferty, and Wasserman 2009), of the data to meet the distributional assumption. The resulting pathway dataset contains the DNA copy number and expression levels of $p = 87$ genes (in total thus $2p = 174$ variates) from 111 cancer and 87 metastasis samples.

The goal of the illustration is operationalized in terms of the presented methodology as follows. The toll-like receptor signaling pathway data in both cancer and metastasis group are assumed to follow a zero-centered multivariate normal distribution with group-specific precision matrices. These precision matrices are estimated by means of the generalized ridge precision estimator using the data of each group and assuming a zero target matrix. In addition, the precision matrix of the metastasis group is also estimated by this estimator with an informative target set equal to the estimated precision matrix of the cancer group (as acquired with a zero target). To adhere to existing molecular biological knowledge, the $2p\times 2p$-dimensional penalty parameter matrix $\boldsymbol{\Lambda}$ is—in each estimation—structured as follows:

- $\text{diag}(\boldsymbol{\Lambda}) = 10^{-10}$. This leaves the diagonal of $\boldsymbol{\Omega}$ (virtually) unpenalized, and thus allows for a good representation of the data by the residual variances.
- $(\boldsymbol{\Lambda})_{j,j'} = \lambda_{\text{ge}}$ for $j,j' = 1,\ldots,p$ and $j\neq j'$. The elements of $\boldsymbol{\Omega}$ relating to gene-gene interactions reflected in the expression data thus enjoy a common penalty λ_{ge}.
- $(\boldsymbol{\Lambda})_{j,j'} = 10^{10}$ for $j,j' = p+1,\ldots,2p$ and $j\neq j'$. Corresponding elements of the precision matrix are thus heavily penalized. This incorporates the assumption that physical interactions among the DNA copy numbers of the pathway's genes are nonexistent.
- $(\boldsymbol{\Lambda})_{j+p,j} = \frac{1}{2}(\lambda_{\text{cn}}+\lambda_{\text{ge}}) = (\boldsymbol{\Lambda})_{j,j+p}$ for $j = 1,\ldots,p$ and λ_{cn} is the DNA copy number specific penalty parameter, which only applies to the DNA copy number-gene expression interaction parameter as the DNA copy numbers are assumed to be (conditionally) independent. The corresponding precision elements, penalized by the average of an expression and DNA copy number penalty parameter, represent the interaction between a gene's DNA copy number and its expression levels.
- $(\boldsymbol{\Lambda})_{j+p,j'} = 10^{10} = (\boldsymbol{\Lambda})_{j,j'+p}$ for $j,j' = 1,\ldots,p$ and $j\neq j'$. Corresponding elements of $\boldsymbol{\Omega}$ are thus encouraged to be zero, which reflects the assumption that a gene's DNA copy number is physically unlikely to have an effect on the expression levels of another gene.

The resulting $\boldsymbol{\Lambda}$ is thus parametrized by two parameters, λ_{ge} and λ_{cn}, that are chosen through optimization of the 5-fold cross-validated log-likelihood for each estimation separately. Finally, note that this penalty matrix is a generalization of that employed by the ordinary ridge precision estimator, in the sense that the former has different penalty parameters for the different omics variates alongside fixed large and small values to encourage zero's in and an unpenalized diagonal of the precision matrix.

We first discuss the resulting cross-validated optimal penalty parameters: $(\lambda^{(c)}_{\text{ge, opt}},\lambda^{(c)}_{\text{cn, opt}}) = (1.495\times10^{-3}, 9.560\times10^{-6})$ and $(\lambda^{(m)}_{\text{ge, opt}},\lambda^{(m)}_{\text{cn, opt}}) = (0.014, 29.507)$ for the cancer and metastasis group, respectively. The optimality of these penalty parameters is confirmed by contourplots of the 5-fold cross-validated log-likelihood (see the supplementary materials). In line with the smaller sample size of the metastasis group its penalty parameters are larger than that of cancer group. Relatively, they are much larger than may be expected on the basis of the sample size difference, in particular that of $\lambda^{(m)}_{\text{cn, opt}}$. This indicates that there is some value in the provided target matrix. The fact that this

excess is more pronounced for $\lambda^{(m)}_{\text{cn, opt}}$ is most likely due to the fact that DNA is a more stable molecule (than mRNA) and the fact that the DNA–mRNA interaction is a well-known and strong regulatory mechanism of gene expression.

With the optimal penalty parameters, the group's precision matrices are estimated. Heatmaps of related partial correlation matrices of these estimates are provided (see supplementary materials). A scatterplot of partial correlations corresponding to mRNA–mRNA interactions reveals that they are much alike between the groups, more than would the precision matrix of the metastasis group have been fitted with a null target (not shown). The same holds for the partial correlations related to the DNA–mRNA, which are more alike due to the corresponding penalty parameter being larger (thus shrinking more to the target derived from the cancer data). The majority of the latter partial correlations are positive, with the positive ones tending to be larger than the negative ones. Biologically, this is in line with the proportionality principle that more DNA facilitates more transcription of the mRNA.

For communication with the medical researcher it is often desirable to sparsify the ridge precision estimator. More specifically, the submatrix of the estimate of $\boldsymbol{\Omega}$ that relates to the interactions among genes' expression levels. For this purpose the local FDR framework introduced by Efron (2004) and translated to precision estimators by Strimmer (2008) is used. The framework assumes a common "null" distribution for the partial correlations corresponding to absent edges. This distribution is estimated by a truncated likelihood approach from the histogram of all estimated partial correlations. With the estimated distribution at hand, the probability of an edge 'being interesting' given the observed partial correlation is calculated. This probability may be endowed with a local FDR interpretation (see Efron 2004). Whenever this probability exceeds 0.80 (following a recommendation by Efron 2004), the corresponding edge is selected. The thus reconstructed networks are presented in Figure 1. The cancer and metastasis network contains 43 and 32 edges, respectively, with 21 overlapping. Hence, the networks are reasonably alike but with clear difference. The most prominent difference is a topological one: the cancer network is more interconnected and contains loops, while the metastasis network forms a tree/chain-like graph. The presence of loops suggests that some type of feedback/forward mechanisms is still active in the cancer group, while it is absent in the metastasis. Biologically, this hints at a (further?) loss of controllability in the metastasized cell, which would be in line with a systems biological paradigm that suggests that as cancer progresses the entropy—reciprocally related to control—of the cellular regulatory network increases (see Teschendorff and Severini 2010; van Wieringen and van der Vaart 2011).

For comparison the analysis was repeated with the regular ridge precision estimator. That is, a target was derived from the cancer data and was subsequently used in the estimation of the precision matrix of the metastasis group. Strikingly, the optimal 5-fold cross-validated penalty parameter for the regular ridge estimator of the metastasis precision matrix was approximately equal to 10^{10}. This large value forces the estimate to be (virtually) equal to the provided (cancer) target. In effect, it does not use the metastasis data for the construction of the precision matrix estimate. As a consequence it yields a worse fit than the generalized ridge precision matrix estimate for the metastasis data. Furthermore, the regular ridge cancer and metastasis precision matrix estimates show no difference. In other words, in

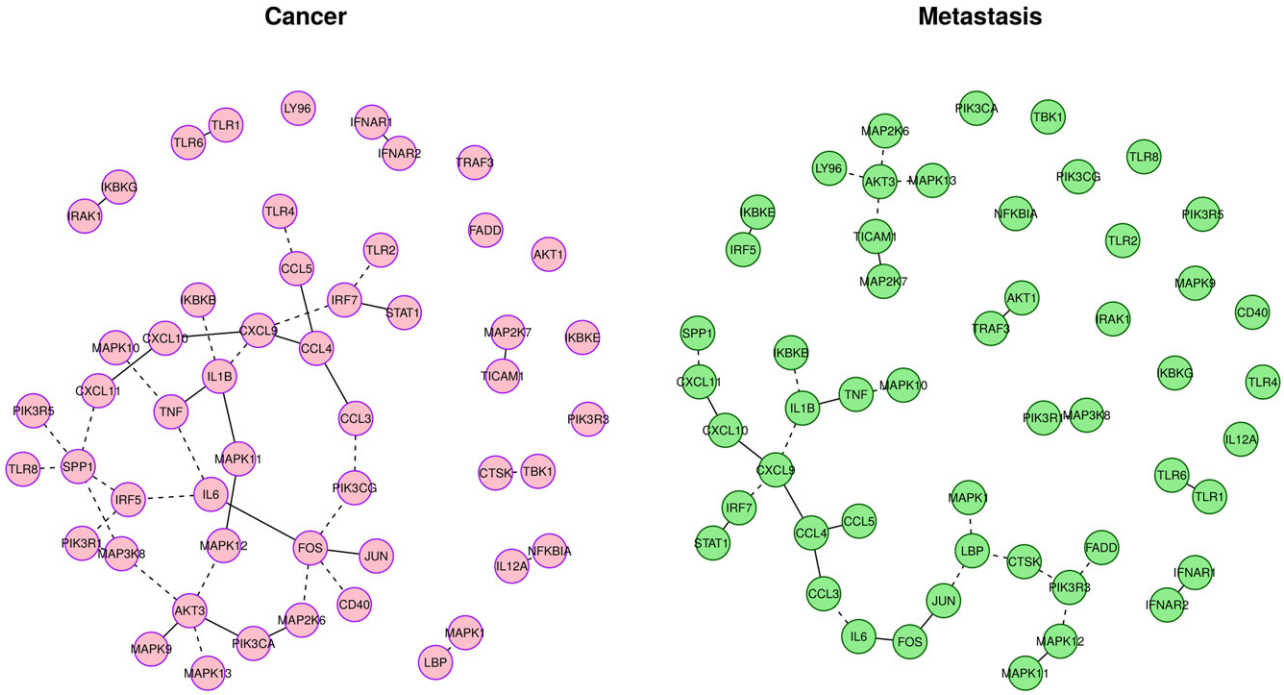

Figure 1. Reconstructed network of the toll-like receptor pathway in cancer (left panel) and metastasis (right panel). For clarity the networks are limited to inferred interactions among genes' expression levels, ignoring those between a gene' expression levels and its DNA copy number as well as unconnected nodes. Dashed and solid edges indicate negative and positive, respectively, signs of the associated partial correlations.

contrast to the flexible generalized ridge precision matrix estimator the stringent penalty structure of its regular counterpart thus rules out differences between the two groups.

This comparison is repeated for the ridge precision estimator with chordal support. The resulting cross-validated optimal penalty parameters for the cancer and metastatis groups are, respectively, $\lambda_{\text{opt}}^{(c)} = 1.707 \times 10^{-3}$ and $\lambda_{\text{opt}}^{(m)} = 3.275 \times 10^{-2}$. The slightly larger optimal penalty parameter of the metastasis group (than that of cancer group) may be due to the smaller sample size but could also indicate that the cancer-derived target matrix contained valuable information for the estimation of the metastasis precision matrix. Heatmaps of the corresponding precision estimates and the correspondingly inferred networks of both groups are included in the supplementary materials. The inferred cancer and metastasis transcriptomic networks comprise 75 and 78 edges, respectively, with an overlap of 55 edges: a substantial overlap but with differences neither clearly interpretable biologically nor from a network topological viewpoint. Without a ground truth there is no way to assess which analysis (generalized or "chordal" ridge) is better. But the ability of the generalized ridge estimator to differentiate in its penalization between the variates from the two molecular levels does provide a flexibility that—in this case—yields a more tangible conclusion for the molecular biologist.

Finally, the various (regular, chordal support, and generalized) pairs of ridge precision estimates are compared in their ability to separate the cancer from the metastasis samples on the basis of their molecular profiles. To this end the quadratic discriminant analysis (QDA) score for each sample is calculated. Class-wise histograms of these scores reveal that the generalized ridge precision estimates achieve the best class separation.

7. Conclusion and Discussion

The proposed generalized ridge precision estimator allows for element-wise penalization with shrinkage to an arbitrary non-random target value. An efficient algorithm for its evaluation is presented. Theoretical properties, for example, positive definiteness and consistency, of the estimator have been shown. It is then pointed out that through iterative application of the proposed estimator the generalized graphical lasso estimator can be approximated. Next, the algorithm for the generalized ridge precision estimator is benchmarked against specific cases which can be deduced from it, indicating a satisfactory computation accuracy. The article closes with an illustrative application of the usage of the presented estimator. In particular, the application shows that the generalized ridge estimator provides ways to incorporate detailed and structured prior knowledge, both quantitative (via the target) and qualitative (by the parameterization of the penalty matrix), into the analysis, which cannot simultaneously be done by existing precision estimators.

The presented work may be extended to adaptively learn the underlying structure of Gaussian graphical model. Here the precision estimator is shrunken to a known target that is set prior to the estimation. The structure of the underlying conditional independence graph reflected in the parameterization of the penalty matrix is fixed throughout the estimation. This may need relaxation. For instance, when variates have a natural ordering due to space or time the local dependence among nearby variates may be approximated by a banded precision matrix. However, neighborhoods of locally depending variates may vary in size. Then, one would like the method to adapt to deviations from the banded structure. For the particular case of ordered variates the nested lasso method of Levina, Rothman, and Zhu (2008) and the related work of Yu and Bien (2017) provide. The presented generalized ridge precision estimator would benefit from the incorporation of similar flexibility for general structures.

Supplementary Materials

Algorithms: `C++` implementation of the proposed algorithms importable into `R` via the `Rcpp`-packages (Eddelbuettel and Francois 2011). (`.cpp`-file).
Algorithms: Alternative generalized graphical lasso algorithm (`.pdf`-file).
Simulations: `R` code (`.r`-files).
Simulations: Description of employed precision matrices (`.pdf`-file).
Simulations: Results (`.pdf`-file).
Illustration: `R` code for the analysis (`.r`-files).
Illustration: Plots (`.pdf`-file).

Acknowledgments

The committed and persistent associate editor and reviewer guided this work toward its current clearer and improved form.

References

Aflakparast, M., de Gunst, M. C. M., and van Wieringen, W. N. (2018), "Reconstruction of Molecular Network Evolution From Cross-Sectional Omics Data," *Biometrical Journal*, 60, 547–583, DOI: 10.1002/bimj.201700102. [935]

Banerjee, O., Ghaoui, L. E., and d'Aspremont, A. (2008), "Model Selection Through Sparse Maximum Likelihood Estimation for Multivariate Gaussian or Binary Data," *Journal of Machine Learning Research*, 9, 485–516. [932]

Bates, D., and Maechler, M. (2018), *Matrix: Sparse and Dense Matrix Classes and Methods*, R Package Version 1.2-14, available at *https://CRAN.R-project.org/package=Matrix*. [937]

Cancer Genome Atlas Research Network (2012), "Comprehensive Genomic Characterization of Squamous Cell Lung Cancers," *Nature*, 489, 519–525. [939]

Carlson, M. (2016), *KEGG.db: A Set of Annotation Maps for KEGG*, R Package Version 3.2.3, available at *https://www.bioconductor.org/packages/release/data/annotation/html/KEGG.db.html*. [939]

Durinck, S., Spellman, P. T., Birney, W., and Huber, W. (2009), "Mapping Identifiers for the Integration of Genomic Datasets With the R/Bioconductor Package biomaRt. R Package Version 2.34.2," *Nature Protocols*, 4, 1184–1191, available at *https://bioconductor.org/packages/release/bioc/html/biomaRt.html*. [939]

Eddelbuettel, D., and Francois, R. (2011), "Rcpp: Seamless R and C++ Integration. R Package Version 0.12.17," *Journal of Statistical Software*, 40, 1—18, available at *https://CRAN.R-project.org/package=Rcpp*. [941]

Efron, B. (2004), "Large-Scale Simultaneous Hypothesis Testing: The Choice of a Null Hypothesis," *Journal of the American Statistical Association*, 99, 96–104. [940]

Fan, J., and Li, R. (2001), "Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties," *Journal of the American Statistical Association*, 96, 1348–1360. [936]

Feng, J., and Simon, N. (2018), "Gradient-Based Regularization Parameter Selection for Problems With Nonsmooth Penalty Functions," *Journal of Computational and Graphical Statistics*, 27, 426–435. [938]

Friedman, J., Hastie, T., and Tibshirani, R. (2008), "Sparse Inverse Covariance Estimation With the Graphical Lasso," *Biostatistics*, 9, 432–441. [932]

——— (2014), *glasso: Graphical Lasso-Estimation of Gaussian Graphical Models*, R Package Version 1.8, available at *https://CRAN.R-project.org/package=glasso*. [937]

Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., and Hothorn, T. (2018), *mvtnorm: Multivariate Normal and t Distributions*, R Package Version 1.0-8, available at *http://CRAN.R-project.org/package=mvtnorm*. [937]

Harville, D. A. (2008), *Matrix Algebra From a Statistician's Perspective*, New York: Springer. [934]

Higham, N. J. (2008), *Functions of Matrices: Theory and Computation*, Philadelphia, PA: SIAM. [933]

Kuismin, M. O., Kemppainen, J. T., and Sillanpää, M. J. (2017), "Precision Matrix Estimation With ROPE," *Journal of Computational and Graphical Statistics*, 26, 682–694. [932,933]

Ledoit, O., and Wolf, M. (2004), "A Well-Conditioned Estimator for Large-Dimensional Covariance Matrices," *Journal of Multivariate Analysis*, 88, 365–411. [935,936]

Levina, E., Rothman, A., and Zhu, J. (2008), "Sparse Estimation of Large Covariance Matrices via a Nested Lasso Penalty," *The Annals of Applied Statistics*, 2, 245–263. [941]

Liu, H., Lafferty, J., and Wasserman, L. (2009), "The Nonparanormal: Semiparametric Estimation of High Dimensional Undirected Graphs," *Journal of Machine Learning Research*, 10, 2295–2328. [939]

Marcus, M., and Sandy, M. (1991), "Hadamard Square Roots," *SIAM Journal on Matrix Analysis and Applications*, 12, 49–69. [933]

Mersmann, O. (2015), *microbenchmark: Accurate Timing Functions*, R Package Version 1.4-2.1, available at *https://CRAN.R-project.org/package=microbenchmark*. [937]

Miok, V., Wilting, S. M., and van Wieringen, W. N. (2017), "Ridge Estimation of the VAR(1) Model and Its Time Series Chain Graph From Multivariate Time-Course Omics Data," *Biometrical Journal*, 59, 172–191. [937]

Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., and Kanehisa, M. (1999), "KEGG: Kyoto Encyclopedia of Genes and Genomes," *Nucleic Acids Research*, 27, 28–34. [939]

Peeters, C. F. W., Bilgrau, A. E., and van Wieringen, W. N. (2017), *rags2ridges: Ridge Estimation of Precision Matrices From High-Dimensional Data*, R Package Version 2.2, available at *https://CRAN.R-project.org/package=rags2ridges*. [937]

Rakoff-Nahoum, S., and Medzhitov, R. (2009), "Toll-Like Receptors and Cancer," *Nature Reviews Cancer*, 9, 57–63. [939]

R Core Team (2018), *R: A Language and Environment for Statistical Computing*, Vienna, Austria: R Foundation for Statistical Computing, available at *https://www.R-project.org/*. [937]

Sato, S., Nakamura, Y., and Tsuchiya, E. (1994), "Difference of Allelotype Between Squamous Cell Carcinoma and Adenocarcinoma of the Lung," *Cancer Research*, 54, 5652–5655. [939]

Strimmer, K. (2008), "fdrtool: A Versatile R Package for Estimating Local and Tail Area-Based False Discovery Rates," *Bioinformatics*, 24, 1461–1462. [940]

Teschendorff, A. E., and Severini, S. (2010), "Increased Entropy of Signal Transduction in the Cancer Metastasis Phenotype," *BMC Systems Biology*, 4, 104. [940]

van der Vaart, A. W. (2000), *Asymptotic Statistics*, New York: Cambridge University Press. [935]

van Wieringen, W. N. (2017), "On the Mean Squared Error of the Ridge Estimator of the Covariance and Precision Matrix," *Statistics and Probablity Letters*, 123, 88–92. [938]

van Wieringen, W. N., and Peeters, C. F. W. (2016), "Ridge Estimation of the Inverse Covariance Matrix From High-Dimensional Data," *Computational Statistics and Data Analysis*, 103, 284–303. [932,933,935,937,938]

van Wieringen, W. N., and van der Vaart, A. W. (2011), "Statistical Analysis of the Cancer Cell's Molecular Entropy Using High-Throughput Data," *Bioinformatics*, 27, 556–563. [940]

Wan, Y. W., Allen, G. I., Anderson, M. L., and Liu, Z. (2015), *TCGA2STAT: Simple TCGA Data Access for Integrated Statistical Analysis in R*, R Package Version 1.2, available at *https://CRAN.R-project.org/package=TCGA2STAT*. [939]

Witten, D. M., Friedman, J. H., and Simon, N. (2011), "New Insights and Faster Computations for the Graphical Lasso," *Journal of Computational and Graphical Statistics*, 20, 892–900. [932]

Yu, G., and Bien, J. (2017), "Learning Local Dependence in Ordered Data," *The Journal of Machine Learning Research*, 18, 1354–1413. [941]