Journal of Multivariate Analysis 178 (2020) 104621

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Updating of the Gaussian graphical model through targeted penalized estimation

Wessel N. van Wieringen [a,b,*], Koen A. Stam [c], Carel F.W. Peeters [a], Mark A. van de Wiel [a,d]

[a] Department of Epidemiology and Biostatistics, Amsterdam Public Health research institute, Amsterdam UMC, location VUmc, P.O. Box 7057, 1007 MB Amsterdam, The Netherlands
[b] Department of Mathematics, Vrije Universiteit Amsterdam, De Boelelaan 1081a, 1081 HV Amsterdam, The Netherlands
[c] Department of Parasitology, Leiden University Medical Center, Leiden, The Netherlands
[d] MRC Biostatistics Unit, University of Cambridge, Cambridge, United Kingdom

ARTICLE INFO

Article history:
Received 11 June 2019
Received in revised form 22 March 2020
Accepted 22 March 2020
Available online 27 March 2020

AMS 2010 subject classifications:
primary 62H12
secondary 62J07
Tertiary 62M05

Keywords:
Conditional independence graph
Inverse covariance
Markov chain
Network
Ridge penalty
Shrinkage

ABSTRACT

Updating of the Gaussian graphical model via shrinkage estimation is studied. This shrinkage is towards a nonzero parameter value representing prior quantitative information. Once new data become available, the previously estimated parameter needs updating. Shrinkage provides the means to this end, using the latter as a shrinkage target to acquire an updated estimate. The process of iteratively updating the Gaussian graphical model in shrinkage fashion is shown to yield an improved fit and an asymptotically unbiased and consistent estimator. The workings of updating via shrinkage are elucidated by linking it to Bayesian updating and through the inheritance by the update of eigen-properties of the previous estimate. The effect of shrinkage on the moments and loss of the estimator are pointed out. Practical issues possibly hampering updating are identified and solutions outlined. The presented updating procedure is illustrated through the reconstruction of a gene–gene interaction network using transcriptomic data.

1. Introduction

The reconstruction of networks from data is a key challenge of our time, as can be witnessed from the number of citations that recently published monographs on graphical models by [38] and [20] have accumulated. Many methods that serve this end have been proposed since, cf., the recently published review by [10]. Virtually all these methods reconstruct such networks *de novo*. That is, they do not make use of information on these networks already available. For molecular networks such information is condensed not only in repositories like KEGG and STRING [27,35], but also in the form of omics data from the model system under comparable conditions available through portals like GEO and TCGA [11,39]. Information from aforementioned sources may enhance the molecular network reconstruction from the data at hand. A thus reconstructed network qualifies itself as novel information on the network. In turn it needs updating when new (experimental) information on the network becomes available in the future. Such iterative reconstruction will progress the knowledge on the network. In this work we explore how this may be achieved.

* Corresponding author at: Department of Epidemiology and Biostatistics, Amsterdam Public Health research institute, Amsterdam UMC, location VUmc, P.O. Box 7057, 1007 MB Amsterdam, The Netherlands.
E-mail address: w.vanwieringen@amsterdamumc.nl (W.N. van Wieringen).

https://doi.org/10.1016/j.jmva.2020.104621
0047-259X/

A network is defined as an undirected graph $\mathcal{G}$ comprising the pair $(\mathcal{V}, \mathcal{E})$. $\mathcal{V}$ is the index set of the nodes – representing the (molecular) entities – of the network. $\mathcal{E} = \mathcal{V} \times \mathcal{V}$ is the edge set, consisting of node pairs that are connected. The graphs under study are undirected, which implies that if node pair $(v_1, v_2) \in \mathcal{E}$ automatically $(v_2, v_1) \in \mathcal{E}$ is included too. An edge is operationalized as a pairwise conditional dependency among the random variables represented by the nodes connected by the edge, given all other random variables represented by the remaining nodes in the network. The absence of an edge is vice versa defined as a conditional independency. The conditional independence graph is to be learned from (say) transcriptomic data on the genes that constitute the network. To this end a Gaussian graphical model (GGM) is assumed: $\mathcal{N}(\mathbf{0}_p, \boldsymbol{\Omega}^{-1})$ with $\boldsymbol{\Omega}$ the inverse covariance matrix $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$. Zeroes and nonzeroes in the inverse covariance matrix $\boldsymbol{\Omega}$ correspond to conditional in- and dependencies, respectively, among the corresponding variate pairs. It thus suffices to estimate the inverse covariance matrix from the data by maximization of the log-likelihood, $\propto \log(|\boldsymbol{\Omega}|) - \text{tr}(\mathbf{S}\boldsymbol{\Omega})$ with sample covariance matrix $\mathbf{S}$ obtained from a sample of size n, yielding the maximum likelihood (ML) estimator $\widehat{\boldsymbol{\Omega}}_{ML} = \mathbf{S}^{-1}$. Rests to infer its support.

The estimation of a Gaussian graphical model is hampered by the collinearity of omics data. Then, the maximum likelihood estimator of the inverse covariance matrix is not (or ill-) defined as $\mathbf{S}$ is (near-)singular. This is overcome by the augmentation of the log-likelihood loss with a penalty. Here we concentrate on the nonzero centered ridge penalty, $\lambda \| \boldsymbol{\Omega} - \mathbf{T}_r \|_F^2$, the sum of the square of the elements of $\boldsymbol{\Omega} - \mathbf{T}_r$. With the subscript r referring to ridge, this $\mathbf{T}_r$ is a user-specified target matrix towards which the precision estimate is shrunken for large values of λ. Maximization of this ridge penalized log-likelihood yields the ridge precision estimator (see [42]):

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

[42] shows that this estimator satisfies the following properties: (i) $\widehat{\boldsymbol{\Omega}}_r(\lambda) \succ 0$, (ii) $\lim_{\lambda \downarrow 0} \widehat{\boldsymbol{\Omega}}_r(\lambda) = \mathbf{S}^{-1}$ (should it exist), (iii) $\lim_{\lambda \to \infty} \widehat{\boldsymbol{\Omega}}_r(\lambda) = \mathbf{T}_r$, and (iv) consistency. Furthermore, for a suitable choice of λ it – using a diagonal, equivariant target $\mathbf{T}_r = \alpha\mathbf{I}_{pp}$ with $\alpha \geq 0$ – also outperforms the ML precision estimator in terms of the mean squared error (MSE) [40]. In that MSE sense, even the inverse of the ridge precision estimator, $\widehat{\boldsymbol{\Sigma}}_r(\lambda) = [\widehat{\boldsymbol{\Omega}}_r(\lambda)]^{-1}$, outperforms the ML covariance estimator. Although the focus of the present work is the ridge precision estimator, an alternative to the ridge precision estimator would be the inverse of the Ledoit–Wolf shrunken covariance matrix [21], $\widehat{\boldsymbol{\Omega}}_{\ell w}(\nu) = [\widehat{\boldsymbol{\Sigma}}_{\ell w}(\nu)]^{-1}$ with $\widehat{\boldsymbol{\Sigma}}_{\ell w}(\nu) = (1-\nu)\mathbf{S} + \nu\mathbf{T}_{\ell w}$ for $\nu \in (0, 1)$. Note that $\mathbf{T}_{\ell w}$ is a covariance target, while $\mathbf{T}_r$ is a precision target. Clearly, this Ledoit–Wolf shrinkage covariance estimator shrinks to $\mathbf{S}$ and $\mathbf{T}_{\ell w}$ as $\nu \downarrow 0$ and $\nu \uparrow 1$, respectively. Moreover, $\widehat{\boldsymbol{\Sigma}}_{\ell w}(\nu)$ too may – for a suitably chosen shrinkage parameter and a diagonal, equivariant target (although it has also been proven for other target choices, [21,31]) – outperform the ML covariance estimator in the mean squared error sense.

The target provides means to include a quantitative suggestion of the precision. The number of choices for $\mathbf{T}_r$ and $\mathbf{T}_{\ell w}$ is in principle unlimited. [16,31] suggest several low-dimensionally parameterized targets. Some archetypical examples (modified for the ridge precision estimator) are:

- The equivariant and rotational invariant target $\mathbf{T}_r = \alpha\mathbf{I}_{pp}$ with $\alpha > 0$.
- A diagonal $\mathbf{T}_r$ with $(\mathbf{T}_r)_{jj} = 1/\mathbf{S}_{jj}$, such that only off-diagonal elements are shrunken.
- A simple model of the covariance and variance by $\mathbf{T}_r = [\sigma^2(1-\rho)\mathbf{I}_{pp} + \sigma^2\rho\mathbf{1}_{pp}]^{-1}$, with $\sigma^2 > 0$ and $\rho \in [-1, 1]$.

[16,31] learn the free parameters of $\mathbf{T}_r$ from the data at hand through, e.g. the minimization of the mean squared error (under assumptions). As an alternative to these low-dimensionally parameterized and data-derived targets, one may also consider a fully unstructured target obtained through other means than the data at hand, e.g.; $\mathbf{T}_r = \boldsymbol{\Omega}_{\text{known}}$ taken from a related and well-studied model system. Of course, the inverse of these $\mathbf{T}_r$ may serve as covariance targets for the Ledoit–Wolf shrinkage covariance estimator $\widehat{\boldsymbol{\Sigma}}_{\ell w}(\nu)$.

Here we describe the problem of updating a Gaussian graphical model after the arrival of novel information. The above introduced shrinkage estimators are employed to accommodate the prior information on the model (in the form of a target that represents the current knowledge/estimate on/of the model at that point in time and formed from the current and preceding data sets) to arrive at an updated estimate from the novel data. Properties of these updating estimators are investigated as well as intuition into their workings is provided. Resolutions for practical complications when updating, batch effects and the simultaneous arrival of multiple sources of novel information, are presented. The paper concludes with an example in which the precision matrix and the therefrom derived gene–gene interaction network of ten pathways are reconstructed from transcriptomic data of ER+ breast cancer samples of a series of five subsequently published data sets. Throughout, proofs are deferred to the appendix, while the supplementary material contains more elaborate versions of the proofs as well as additional results.

1.1. Related work

Estimation of precision matrices from multiple data sets has been considered in the literature. [6,8,29] deal with the simultaneous estimation of multiple precision matrices using so-called 'fused' ridge/lasso penalties, while e.g. [25,28] adopt a Bayesian approach. The methods of [15,43] have a similar thrust but these works are based on parameter communalities between the different precision matrices. On a different note but closely related is the work of [25,26,44], where the focus is on the reconstruction of multiple directed acyclic graphs in order to facilitate causal inference.

Implicitly, these methods assume that the data used for the estimation of each precision matrix are on par, in the sense that no time order is present. This absence of an order motivates the use of a fused penalty that encourages to borrow information across data sets (should the data give rise to do so). In particular, this shrinks the different precision estimates towards each other, but not towards a target matrix that is formed from one of the other data sets. Moreover, this 'across data set'-calibration is done once, instead of repetitively in the envisioned updating context considered here.

2. Updating

When new information becomes available, the previously reconstructed network and GGM needs updating. The latter may then serve as the target in the re-estimation of the network from the new data. This process may be iterated, thus giving rise to an updating scheme. To formalize this, assume an infinite sequence of studies, each producing data from the same normal law $\mathcal{N}(\mathbf{0}_p, \boldsymbol{\Omega}^{-1})$ (variations on and deviations from this scenario are discussed at the end of this section and in Section 4). This yields a sequence of sample covariance matrices $\{\mathbf{S}_k\}_{k=1}^{\infty}$ with corresponding sample sizes $\{n_k\}_{k=1}^{\infty}$. For the ridge precision matrix and Ledoit–Wolf shrinkage covariance matrix estimators the following recursive updating schemes may then be conceived:

$$\begin{cases} \widehat{\boldsymbol{\Omega}}_{r,0}(\lambda_0) &= \mathbf{T}_{r,0} \\ \widehat{\boldsymbol{\Omega}}_{r,k+1}(\lambda_{k+1}) &= \left\{\frac{1}{2}(\mathbf{S}_{k+1} - \lambda_{k+1}\widehat{\boldsymbol{\Omega}}_{r,k}) + [\lambda_{k+1}\mathbf{I}_{pp} + \frac{1}{4}(\mathbf{S}_{k+1} - \lambda_{k+1}\widehat{\boldsymbol{\Omega}}_{r,k})^2]^{1/2}\right\}^{-1}, \qquad k \in \{0, 1, \ldots\} \end{cases} \tag{1}$$

and

$$\begin{cases} \widehat{\boldsymbol{\Sigma}}_{\ell w,0}(\nu_0) &= \mathbf{T}_{\ell w,0} \\ \widehat{\boldsymbol{\Sigma}}_{\ell w,k+1}(\nu_{k+1}) &= (1-\nu_{k+1})\mathbf{S}_{k+1} + \nu_{k+1}\widehat{\boldsymbol{\Sigma}}_{\ell w,k}(\nu_k), \qquad k \in \{0, 1, \ldots\}, \end{cases} \tag{2}$$

respectively. Both schemes are initialized by a (possibly uninformative) target matrix: $\widehat{\boldsymbol{\Omega}}_{r,0}(\lambda_0) = \mathbf{T}_{r,0}$ and $\widehat{\boldsymbol{\Sigma}}_{\ell w,0}(\nu_0) = \mathbf{T}_{\ell w,0}$, where $\{\mathbf{S}_k\}_{k=1}^{\infty}$ is a sequence of independent random variables, the sequences of updated estimators $\{\widehat{\boldsymbol{\Sigma}}_{\ell w,k}(\nu_k)\}_{k=1}^{\infty}$ and $\{\widehat{\boldsymbol{\Omega}}_{r,k}(\lambda_k)\}_{k=1}^{\infty}$ may be viewed as originating from a discrete time, 1st order Markov process as e.g.:

$$\widehat{\boldsymbol{\Omega}}_{r,k+1}(\lambda_{r,k+1}) \,|\, \widehat{\boldsymbol{\Omega}}_{r,k}(\lambda_{r,k}), \ldots, \widehat{\boldsymbol{\Omega}}_{r,0}(\lambda_{r,0}) \sim \widehat{\boldsymbol{\Omega}}_{r,k+1}(\lambda_{r,k+1}) \,|\, \widehat{\boldsymbol{\Omega}}_{r,k}(\lambda_{r,k}).$$

These 1st order Markov processes have continuous state space $\mathcal{S}_{++}$, the space of symmetric, positive definite matrices, and are time-homogeneous as, e.g.:

$$\widehat{\boldsymbol{\Omega}}_{r,k+\kappa+1}(\lambda_{r,k+\kappa+1}) \,|\, \widehat{\boldsymbol{\Omega}}_{r,k+\kappa}(\lambda_{r,k+\kappa}) = \mathbf{U}, \lambda_{r,k+\kappa+1} = \lambda \sim \widehat{\boldsymbol{\Omega}}_{r,k+1}(\lambda_{r,k+1}) \,|\, \widehat{\boldsymbol{\Omega}}_{r,k}(\lambda_{r,k}) = \mathbf{U}, \lambda_{r,k+1} = \lambda$$

for all $\mathbf{U} \in \mathcal{S}_{++}$, $\kappa \in \mathbb{N}$ and $\lambda \in (0, \infty)$. For given $\mathbf{U}$ and λ, the distribution only depends on a sample covariance matrix that is always drawn from the same Wishart distribution. The related sequences $\{\widehat{\boldsymbol{\Omega}}_{\ell w,k}(\nu_k)\}_{k=1}^{\infty}$ and $\{\widehat{\boldsymbol{\Sigma}}_{r,k}(\lambda_k)\}_{k=1}^{\infty}$, defined for all k by $\widehat{\boldsymbol{\Omega}}_{\ell w,k}(\nu_k) = \widehat{\boldsymbol{\Sigma}}^{-1}_{\ell w,k}(\nu_k)$ and $\widehat{\boldsymbol{\Sigma}}_{r,k}(\lambda_k) = \widehat{\boldsymbol{\Omega}}^{-1}_{r,k}(\lambda_k)$ for the Ledoit–Wolf shrinkage covariance and ridge precision estimator, respectively, can also be considered as generated by a discrete time, time-homogeneous 1st order Markov process. We now concentrate on the updating of the covariance matrix, before turning our attention to that of the precision matrix.

Intuitively, it may be expected that the reconstructed network/GGM improves as the number of iterations increases. This is formalized in the next propositions that, relying on concentration inequalities for the sample covariance matrix provided by [19,37], state that the fit (in some sense) improves (with a certain probability) over the iterations. More specifically, the Frobenius and spectral loss of the covariance matrix estimator $\widehat{\boldsymbol{\Sigma}}_{r,k}(\lambda_k)$ minus the sample covariance matrix of the next iteration is likely to be smaller in the next iteration than that of the current. This type of probability is known as a fluctuation inequality (see [5]).

Proposition 1 (*Fluctuation Inequality I, Ridge*). *Let $\mathbf{S}_k$ and $\mathbf{S}_{k+1}$ be sample covariance matrices obtained from samples with size n_k and n_{k+1}, respectively, both drawn from $\mathcal{N}(\mathbf{0}_p, \boldsymbol{\Sigma})$ and let $\widehat{\boldsymbol{\Sigma}}_{r,k}(\lambda_k)$ and $\widehat{\boldsymbol{\Sigma}}_{r,k+1}(\lambda_{k+1})$ be the corresponding updated ridge covariance matrix estimators. Denote by $\mathbf{R}$ the correlation matrix associated with $\boldsymbol{\Sigma}$. Define $\xi_{j,j'} = \max\{[1 - (\mathbf{R})_{j,j'}][(\boldsymbol{\Sigma})_{j,j}(\boldsymbol{\Sigma})_{j',j'}]^{1/2}, [1+(\mathbf{R})_{j,j'}][(\boldsymbol{\Sigma})_{j,j}(\boldsymbol{\Sigma})_{j',j'}]^{1/2}\}$ and $\xi = \max_{j,j'} \xi_{j,j'}$. Then, given the current covariance estimate $\widehat{\boldsymbol{\Sigma}}_{r,k}(\lambda_k)$ with $\lambda_k > 0$, for every $\lambda_{k+1} \in (0, \infty)$, there exists a $\delta(\lambda_{k+1}) > 0$ for which:*

$$\begin{aligned} &P(\|\widehat{\boldsymbol{\Sigma}}_{r,k+1}(\lambda_{k+1}) - \mathbf{S}_{k+1}\|_F^2 < \|\widehat{\boldsymbol{\Sigma}}_{r,k}(\lambda_k) - \mathbf{S}_k\|_F^2) \\ &\quad \geq 1 - \min\left\{1, 2\exp[\log(p^2+p) - 3n_k t^2/(16\xi^2)] + 2\exp[\log(p^2+p) - 3n_{k+1}t^2/(16\xi^2)]\right\}. \end{aligned}$$

with $t = \min\{\delta(\lambda_{k+1}), \frac{1}{2}p^2\xi^2\}$.

The bound for this probability provided by Proposition 1 is crude (as the inequalities used in its proof are not optimal) and could be improved upon, but which would not fundamentally alter the qualitative consequences of the proposition. More importantly, a similar result can be formulated in terms of the spectral norm $\|\mathbf{A}\|_\infty = \lim_{q\to\infty}[\sum_{j=1}^{p} |d_j(\mathbf{A})|^q]^{1/q}$, where $d_j(\mathbf{A})$ is the jth eigenvalue of the $p \times p$ dimensional matrix $\mathbf{A}$.

Proposition 2 (*Fluctuation Inequality II, Ridge*). *Let $\mathbf{S}_k$ and $\mathbf{S}_{k+1}$ be sample covariance matrices obtained from samples with size n_k and n_{k+1}, respectively, both drawn from $\mathcal{N}(\mathbf{0}_p, \boldsymbol{\Sigma})$ and let $\widehat{\boldsymbol{\Sigma}}_{r,k}(\lambda_k)$ and $\widehat{\boldsymbol{\Sigma}}_{r,k+1}(\lambda_{k+1})$ be the corresponding updated ridge covariance matrix estimators. Then, given the current covariance estimate $\widehat{\boldsymbol{\Sigma}}_{r,k}(\lambda_k)$ with $\lambda_k > 0$, for every $\lambda_{k+1} \in (0, \infty)$, there exists a $\delta(\lambda_{k+1}) \in (0, 1)$ such that for all $t \geq 1$:*

$$P(\|\widehat{\boldsymbol{\Sigma}}_{r,k+1}(\lambda_{k+1}) - \mathbf{S}_{k+1}\|_\infty < \|\widehat{\boldsymbol{\Sigma}}_{r,k}(\lambda_k) - \mathbf{S}_k\|_\infty) \geq 1 - 4\exp(-t^2 p),$$

if $n_{k+1} \geq C[2t/\delta(\lambda_{k+1})]^2 p$. Here C depends only on the sub-Gaussian norm of $\mathbf{Y} \sim \mathcal{N}(\mathbf{0}_p, \boldsymbol{\Sigma})$.

The norm equivalence between the spectral and Frobenius norms employed in Propositions 1 and 2, respectively, implies that each contains a bound of the same probability formulated in terms of the other norm. This is not pursued here.

The implication of Propositions 1 and 2 is that, when n_{k+1} is sufficiently large, $P(\|\widehat{\boldsymbol{\Sigma}}_{r,k+1}(\lambda_{k+1}) - \mathbf{S}_{k+1}\| < \|\widehat{\boldsymbol{\Sigma}}_{r,k}(\lambda_k) - \mathbf{S}_k\|)$ will be close to one. Similarly, for the related probability for the Ledoit–Wolf shrinkage covariance estimator. Put differently, the fit of the shrinkage covariance estimators may benefit from updating, i.e. the use of a target constructed from previous data sets. On one hand this suggests a reduction in the potential bias of these estimators (as they are closer – in Frobenius norm – to the sample covariance matrix). On the other, as shrinkage preserves information from past data sets, the shrinkage estimators exhibit less variance. The reconstruction of the underlying conditional independence graph from the shrinkage is likely to profit from the combination of these properties.

Propositions 1 and 2 suggest that the updating procedure may actually converge, as the sequences of estimators are likely to contract in norm towards the kth sample covariance matrix. As the sample covariance matrices $\{\mathbf{S}_k\}_{k=1}^\infty$ are unbiased estimators of the covariance matrix, i.e. $\mathbb{E}(\mathbf{S}_k) = \boldsymbol{\Sigma}$ for all k, one may hope that by the contraction and after enough iterations they will have a positive effect on the estimators in the sense that the estimators converge in expectation to the population covariance matrix. This is stated in the next theorem for the Ledoit–Wolf shrinkage estimator and can be proven straightforwardly by use of the analytic expression of the estimator.

Theorem 1 (*Asymptotic Unbiasedness, Ledoit-Wolf*). *The bias of $\widehat{\boldsymbol{\Sigma}}_{\ell w,k}(\nu_k)$ vanishes as the number of updates increases. Formally, let $\nu_k \in (0, 1)$ for all k, then:* $\lim_{k\to\infty} \mathbb{E}[\widehat{\boldsymbol{\Sigma}}_{\ell w,k}(\nu_k)] = \boldsymbol{\Sigma}$.

A similar result can be formulated for the ridge covariance estimator (Theorem 2). This requires the existence of the stationary density of the Markov process generating the sequence $\{\widehat{\boldsymbol{\Sigma}}_{r,k}(\lambda_k)\}_{k=1}^\infty$, which is warranted (see [32]; Theorem 8.2.14]) by the irreducibility or sufficient mixing and the tightness of the sequence of random variables. The study of the estimating equation under stationarity of the sequence then delivers the asymptotic unbiasedness of the ridge covariance estimator (Theorem 2). A similar line of argument may be applied to $\{\widehat{\boldsymbol{\Sigma}}_{\ell w,k}(\nu_k)\}_{k=1}^\infty$ in order to obtain an alternative proof of Theorem 1.

Theorem 2 (*Asymptotic Unbiasedness, Ridge*). *The bias of $\widehat{\boldsymbol{\Sigma}}_{r,k}(\lambda_k)$ vanishes as the number of updates increases. Formally, let $\lambda_k \in (0, \infty)$ for all k, then:* $\lim_{k\to\infty} \mathbb{E}[\widehat{\boldsymbol{\Sigma}}_{r,k}(\lambda_k)] = \boldsymbol{\Sigma}$.

Effectively, Theorems 1 and 2 state a particular form of asymptotic unbiasedness, in which the sample size increases in steps equal to or larger than one while the already acquired samples are summarized in the form of the target matrix. Although the asymptotic unbiasedness in Theorems 1 and 2 is stated in the k-limit (the number of data sets), the actual sample sizes $\{n_k\}_k^\infty$ are not irrelevant. The latter contribute, together with the choice of the initial target $\mathbf{T}_0$, to the speed of the convergence. Data sets with large sample sizes require little shrinkage. Consequently, the estimate is close to the sample covariance matrix, an unbiased estimator of the covariance matrix. In turn this estimate (or its inverse) serves as the target for the estimate from next data set, thus steering the next updated estimate towards an unbiased estimate of the mean. The convergence speed furthermore depends on the norm of the difference between the current estimate or sample covariance matrix and its target (see Supplementary Material I). Thereby, the size of this norm effectively depends on how close (in norm) subsequent covariance matrices and the true covariance matrix are. Along with the parametric specifics of $\boldsymbol{\Sigma}$, this is determined by the data sets' sample sizes.

A notable peculiarity of Theorems 1 and 2 is the fact that they hold irrespective of the penalty parameter choices. That is, as long as the penalty parameters are nonzero, the updated estimates are shrunken towards the target and profit from the information provided by preceding data sets and accumulated in the target. This is convenient as it is thus irrelevant how the penalty parameters are chosen, although their size matters for the convergence speed. It also expresses that no matter how much shrinkage is applied the data will prevail.

Finally, not explicitly mentioned but Theorems 1 and 2 convey a traditional asymptotic view, i.e. fixed p while $n \to \infty$ (here $k \to \infty$). This suffices for practical purposes where the system under study – e.g. a gene–gene interaction network – is defined from the outset of the updating scheme. Extensions of both theorems into the high-dimensional asymptotic domain, where p also tends to infinity as a function of the sample size, are hampered by the dependency of current estimator on the preceding ones, and ultimately on the finite dimensionally sized target of the estimator from the first data. Those too have to grow with the dimension p for such an asymptotic result. Nonetheless, Theorem 1 can be extended into the high-dimensional asymptotic domain as the influence of the target vanishes as $k \to \infty$. Then, the asymptotic unbiasedness of the updated Ledoit–Wolf covariance estimator hinges only upon the marginal convergence of

the elements of the estimator and is thus not affected by an increase of the dimension. Such an extension for the updated ridge covariance estimator is not immediate and may even be in need of additional assumptions. Such assumptions are expected be analogous to those warranting high-dimensional asymptotic properties of the graphical lasso estimator (see [7,23]). These assumptions are, however, hard to verify for real data, which questions their usefulness for practical purposes [2,3,22]. We therefore refrain from pursuing such results in the remainder.

Next to asymptotic unbiasedness, the updated estimators can – under a particular regime of the penalty parameters – be shown to be consistent. The proof hinges upon a weak law of the weighted average (see [18]), of which both covariance matrix estimators turn out to be (approximate) exponents.

Theorem 3 (*Consistency, Covariance Matrix Estimators*). *Let* $\{\nu_k\}_{k=1}^{\infty}$ *be a sequence such that* $\nu_k \in (0,1)$ *for all* k *and* $\lim_{k\to\infty}(1-\nu_k)/[\sum_{k''=1}^{k}(1-\nu_k)\prod_{k'=1}^{k''-1}\nu_{k'}] = 0$. *Moreover, let* $\{\lambda_k\}_{k=1}^{\infty}$ *be a sequence such that* $\lambda_k \in (0,\infty)$ *for all* k, $\lambda_k^{-1} \gg \lambda_k^{-2}$ *for all* $k > k_0$ *with* $k_0 \in \mathbb{N}$ *sufficiently large, and* $\lim_{k\to\infty}\lambda_{k+1}^{-1}/\{\sum_{k''=1}^{k}\lambda_{k''}^{-1}[\prod_{k'=k''+1}^{k+1}(1-\lambda_{k'}^{-1})]\}$. *Then, the covariance matrix estimators* $\widehat{\boldsymbol{\Sigma}}_{\ell w,k}(\nu_k)$ *and* $\widehat{\boldsymbol{\Sigma}}_{r,k}(\lambda_k)$ *are consistent, i.e.* $\widehat{\boldsymbol{\Sigma}}_{\ell w,k}(\nu_k) \xrightarrow{P} \boldsymbol{\Sigma}$ *and* $\widehat{\boldsymbol{\Sigma}}_{r,k}(\lambda_k) \xrightarrow{P} \boldsymbol{\Sigma}$ *as* $k \to \infty$.

The parameter restrictions in the theorem above, apart from the assurance of convergence of a weighted average, require that more and more is borrowed from the preceding data sets. Furthermore, with additional restrictions on the penalty parameter sequences stronger convergence results may be obtained by virtue of other results on the convergence of weighted averages [13]. In general, i.e. with unrestricted penalty parameters, the estimators are not expected to be consistent for they may then (due to small values of the penalty parameter) depend strongest on the latest sample covariance matrix, which is in practice derived from a finite sample. Hence, the variance of the estimators will then not vanish, which if true, would – by Chebyshev's inequality – yield consistency.

The results above are reported for the covariance matrix estimators $\widehat{\boldsymbol{\Sigma}}_{r,k}(\lambda_r)$ and $\widehat{\boldsymbol{\Sigma}}_{\ell w,k}(\nu_r)$ while interest is in their inverses. The reported results may be expected to carry over to the precision matrices. For instance, the asymptotic unbiasedness of Theorems 1 and 2 will, much like the fact that $\mathbb{E}(\mathbf{S}) = \boldsymbol{\Sigma}$ while $\mathbb{E}(\mathbf{S}^{-1}) = (n-p+1)^{-1}\boldsymbol{\Sigma}^{-1}$, imply that $\mathbb{E}[\widehat{\boldsymbol{\Omega}}_{r,k}(\lambda_r)] = b\boldsymbol{\Omega}$ for some $b > 0$. The precise value of this constant b is irrelevant as it will vanish in the derivation of the partial correlation matrix, the basis for the reconstruction of the conditional independence graph. Indeed, the consistency – and thereby the asymptotic unbiasedness – of the covariance matrix estimators transfers to their inverses, which is a direct consequence of the Continuous Mapping Theorem (see [36]):

Corollary 1 (*Consistency, Precision Matrix Estimators*). *Let* $\{\nu_k\}_{k=1}^{\infty}$ *be a sequence such that* $\nu_k \in (0,1)$ *for all* k *and* $\lim_{k\to\infty}(1-\nu_k)/[\sum_{k''=1}^{k}(1-\nu_k)\prod_{k'=1}^{k''-1}\nu_{k'}] = 0$. *Moreover, let* $\{\lambda_k\}_{k=1}^{\infty}$ *be a sequence such that* $\lambda_k \in (0,\infty)$ *for all* k, $\lambda_k^{-1} \gg \lambda_k^{-2}$ *for all* $k > k_0$ *with* $k_0 \in \mathbb{N}$ *sufficiently large, and* $\lim_{k\to\infty}\lambda_{k+1}^{-1}/\{\sum_{k''=1}^{k}\lambda_{k''}^{-1}[\prod_{k'=k''+1}^{k+1}(1-\lambda_{k'}^{-1})]\}$. *Then, the precision matrix estimators* $\widehat{\boldsymbol{\Omega}}_{\ell w,k}(\nu_k)$ *and* $\widehat{\boldsymbol{\Omega}}_{r,k}(\lambda_k)$ *are consistent, i.e.* $\widehat{\boldsymbol{\Omega}}_{\ell w,k}(\nu_k) \xrightarrow{P} \boldsymbol{\Omega}$ *and* $\widehat{\boldsymbol{\Omega}}_{r,k}(\lambda_k) \xrightarrow{P} \boldsymbol{\Omega}$ *as* $k \to \infty$.

To provide some intuition behind the consistency results consider the following scheme of the Ledoit–Wolf shrinkage parameter $\nu_k = (k+1)^{-1}$ for all k. This scheme reduces the Ledoit–Wolf shrinkage estimator to the pooled sample covariance estimator $\widehat{\boldsymbol{\Sigma}}_{pool,k} = k^{-1}\sum_{k'=1}^{k}\mathbf{S}_{k'}$, a well-known consistent estimator of the covariance matrix. A similar observation can be made for the ridge covariance estimator. Hereto approximate the ridge covariance matrix estimator around '$\lambda_r = \infty$' by the first order negative term of a Laurent series: $\widehat{\boldsymbol{\Sigma}}_{r,k+1}(\lambda_{k+1}) = (1-\lambda_{k+1}^{-1})\widehat{\boldsymbol{\Sigma}}_{r,k}(\lambda_k) + \lambda_{k+1}^{-1}\mathbf{S}_{k+1} + \mathcal{O}(\lambda_{k+1}^{-2})$. Now choose $\lambda_k = (k+1)^{-1}$ for all k and conclude that for appropriate schemes of the penalty/shrinkage parameters both estimators behave as weighted averages. These weights determine how the current estimate is weighed against new data.

In simulation we investigate the effect of the sample size n and dimension p on the convergence of the estimators. Data with fixed sample size $n \in \{10, 25, 50, 100\}$ are drawn $K = 10000$ subsequent times from $\mathcal{N}(\mathbf{0}_p, \boldsymbol{\Omega}^{-1})$ with $p \in \{10, 25, 50, 100\}$. The precision matrix $\boldsymbol{\Omega}$ has unit diagonal and three off-diagonal bands. Its elements are specified through: $(\boldsymbol{\Omega})_{jj} = 1$ for $j \in \{1,\ldots,p\}$, $(\boldsymbol{\Omega})_{j,j+1} = \frac{1}{2} = (\boldsymbol{\Omega}_2)_{j+1,j}$ for $j \in \{1,\ldots,p-1\}$, $(\boldsymbol{\Omega})_{j,j+2} = \frac{1}{5} = (\boldsymbol{\Omega}_2)_{j+2,j}$ for $j \in \{1,\ldots,p-2\}$, $(\boldsymbol{\Omega})_{j,j+3} = \frac{1}{10} = (\boldsymbol{\Omega}_2)_{j+3,j}$ for $j \in \{1,\ldots,p-1\}$, and zero otherwise. The precision matrix is estimated from the K data sets in accordance with updating scheme (1) where the penalty parameter λ_k is chosen through leave-one-out cross-validation (LOOCV). Updating is initiated with three different targets: a zero target $\mathbf{T}_r = \mathbf{0}_{pp}$, a diagonal target $\mathbf{T}_r = \mathbf{I}_{pp}$, and a perfect $\mathbf{T}_r = \boldsymbol{\Omega}$. For each $(n,p,\mathbf{T}_r)$-combination the updating scheme is run twice. For the kth estimate its 'bias', $\sum_{j,j'=1}^{p}[\widehat{\boldsymbol{\Omega}}_{r,k}(\lambda_k)]_{j,j'} - (\boldsymbol{\Omega})_{j,j'}$, and its squared error, $\|\widehat{\boldsymbol{\Omega}}_{r,k}(\lambda_k) - \boldsymbol{\Omega}\|_F^2$, are plotted against k (see Supplementary Material II). The plots reveal that for any $(n,p,\mathbf{T}_r)$-combination the 'bias' and the squared error of the estimates tend to zero as k increases. An increase of the sample size n reduces the variation in these quantities. The biggest effect on the 'bias' and squared loss is attributable to the choice of the initial target $\mathbf{T}_r$. The closer this target is to the true precision matrix, the faster the convergence of the 'bias' and squared loss to zero. Irrespectively, the effect of the choice of the initial target is washed out after enough updates. But with larger dimensions p it takes longer before the effect of the choice of the target $\mathbf{T}_r$ is washed out. Finally, the two replications of each setting do not differ much, suggesting that the convergence of the updating process is stable.

The assumption made in the beginning of this section that each draw originates from the same multivariate normal law is untenable in practice. For instance, the k_0-th draw may stem from a multivariate normal law with a different precision

matrix. Then, when the pooled covariance matrix estimator $\widehat{\boldsymbol{\Sigma}}_{pool,k_0}$ is employed, the 'outlying' draw contributes to the pooled estimator equally to the other draws. Its influence on the estimator will then persist and only slowly vanish as the number of draws increases. The discussed shrinkage estimators offer the possibility to ignore this 'outlying draw' through large values of λ_{k_0} and ν_{k_0}. Such strong penalization overrules the 'outlying' draw in favor of the target (based on the preceding draws). As such, updating by shrinkage estimation has robust potential.

Updating via shrinkage to the previous estimates, thereby using a nonzero target, affects the properties of the ridge precision estimator. This is sketched in Supplementary Material III.

3. Why updating works

Intuition is provided as to why updating improves the knowledge of the sought-for network. In this and the next two sections the subscript k of the estimators is temporarily dropped to reduce notational clutter.

3.1. Target as prior information

Updating is the bread and butter of Bayesian inference (see, e.g. [4]). With each penalized estimation procedure having a Bayesian analogue, it is natural to ask whether updating via targeted penalized estimation can be viewed from a Bayesian perspective. Indeed, the target matrix serves as prior information on the Gaussian graphical model. This is due to the fact that both precision/covariance estimators have – under the assumption of a suitable prior – a Bayesian interpretation. For the shrinkage estimators assume an inverse Wishart prior for $\boldsymbol{\Sigma}$: $\boldsymbol{\Sigma} \sim \mathcal{IW}(c_\sigma \mathbf{T}_{\ell w}, m_\sigma)$. Then, $\mathbb{E}(\boldsymbol{\Sigma} \mid \mathbf{S}) = (c_\sigma \mathbf{T}_{\ell w} + n\mathbf{S})/(m_\sigma + n - p - 1)$, which coincides with $\widehat{\boldsymbol{\Sigma}}_{\ell w}(\nu)$ for suitable choices of c_σ and m_σ. Similarly, for the inverse of the shrinkage covariance estimator assume $\boldsymbol{\Omega} \sim \mathcal{W}(c_\omega^{-1}\mathbf{T}_{\ell w}^{-1}, m_\omega)$. Then: $\mathbb{E}(\boldsymbol{\Omega} \mid \mathbf{S}) = (m_\omega + n)(c_\omega \mathbf{T}_{\ell w} + n\mathbf{S})^{-1}$. As before, for suitable choices of c_ω and m_ω, this posterior mean equals $\widehat{\boldsymbol{\Omega}}_{\ell w}(\nu)$. For the ridge precision estimator it is unknown which prior induces a posterior with a mean that coincides with it. Would instead the maximum a posteriori (MAP) estimator be used, it is immediate from the penalized log-likelihood that the prior has to be normal-like around the target to produce the ridge precision estimator as MAP. Laplace's method may then be used to produce a full normal approximation to the posterior mode (i.e. the ridge precision estimator). The Bernstein–von Mises theorem [36] specifies the conditions that ensure the quality (in some sense) of this approximation. In conclusion, for both the Ledoit-Wolf shrinkage and ridge estimators the target matrix can thus be interpreted as the location of the prior of the Gaussian graphical model parameter. Hence, the proposed updating procedure via targeted penalized estimation may be viewed as a frequentist's analogue of Bayesian updating.

3.2. Eigenspace updating

The use of a target matrix affects both the eigenvalues and vectors of the ridge precision estimator as pointed out by [42]. In particular, the eigenvalues are shrunken to those of the target matrix. To see this consider the eigendecomposition of the matrix $\mathbf{S} - \lambda\mathbf{T}_r = \mathbf{V}_{s-\lambda t}\mathbf{D}_{s-\lambda t}\mathbf{V}_{s-\lambda t}^\top$ with $\mathbf{V}_{s-\lambda t}$ and diagonal $\mathbf{D}_{s-\lambda t}$ the matrices with eigenvectors and eigenvalues, respectively. The eigendecomposition of the ridge precision estimator then is:

$$\widehat{\boldsymbol{\Omega}}_r(\lambda) = \mathbf{V}_{s-\lambda t}[\tfrac{1}{2}\mathbf{D}_{s-\lambda t} + (\lambda\mathbf{I}_{pp} + \tfrac{1}{4}\mathbf{D}_{s-\lambda t}^2)^{1/2}]^{-1}\mathbf{V}_{s-\lambda t}^\top,$$

in which we dropped the subscript r on the right-hand side to avoid notational clutter. For a general target, the eigenvalues of the ridge precision estimator shrink – for large values of the penalty matrix – to those of the target matrix (cf. Proposition 1 of [42]). It is illustrative to consider an equivariant and rotational invariant target $\mathbf{T}_r = \alpha\mathbf{I}_{pp}$. With such a target the eigenspaces of $\widehat{\boldsymbol{\Omega}}_r(\lambda)$ and $\mathbf{S}$ coincide and the eigenvalues of the former are:

$$\mathbf{D}_{\omega(\lambda)} = \{\tfrac{1}{2}(\mathbf{D}_s - \lambda\alpha\mathbf{I}_{pp}) + [\lambda\mathbf{I}_{pp} + \tfrac{1}{4}(\mathbf{D}_s - \lambda\alpha\mathbf{I}_{pp})^2]^{1/2}\}^{-1}.$$

From this it can be seen (e.g. see [42]) that $\lim_{\lambda\to\infty}\mathbf{D}_{\omega(\lambda)} = \alpha\mathbf{I}_{pp}$.

Not mentioned in [42] is how the inclusion of a general target affects the eigenvectors of the ridge precision estimator. With a null or diagonal equivariant target the span of the eigenspace is derived from the data, i.e. $\mathbf{S}$. With a more general target not only the spread along the eigenvectors is shrunken, but also the canonical direction of the spread is changed. This is summarized in Theorem 4 which, relying on standard linear algebra, matrix perturbation theory [33] and the Davis–Kahan $\sin(\Theta)$ theorem [9], states how the eigenvectors of the ridge precision estimator are affected by the penalty parameter. An analogous theorem for the Ledoit–Wolf shrinkage estimator can be formulated. Fig. 1 illustrates the properties of Theorem 4. The left panel shows that the eigenvectors of $\widehat{\boldsymbol{\Omega}}_r(\lambda)$ are a rotation of those of $\mathbf{S}$. The right panel plots the angle between the first eigenvector of $\mathbf{S}$ and that of $\widehat{\boldsymbol{\Omega}}_r(\lambda)$ (i.e., $\angle(\mathbf{v}_{s,1}, \mathbf{v}_{\omega(\lambda),1})$ where $\angle$ denotes the angle) against λ.

Theorem 4. *Let $\mathbf{V}_s$, $\mathbf{V}_t$, $\mathbf{V}_{\omega(\lambda)}$ the matrices with eigenvectors as columns of $\mathbf{S}$, $\mathbf{T}$, and $\widehat{\boldsymbol{\Omega}}(\lambda)$, respectively. Then, the map $\lambda \mapsto \mathbf{V}_{\omega(\lambda)}$*

(i) *is continuous;*
(ii) *has limits $\lim_{\lambda\downarrow 0}\mathbf{V}_{\omega(\lambda)} = \mathbf{V}_s$ and $\lim_{\lambda\to\infty}\mathbf{V}_{\omega(\lambda)} = \mathbf{V}_t$;*

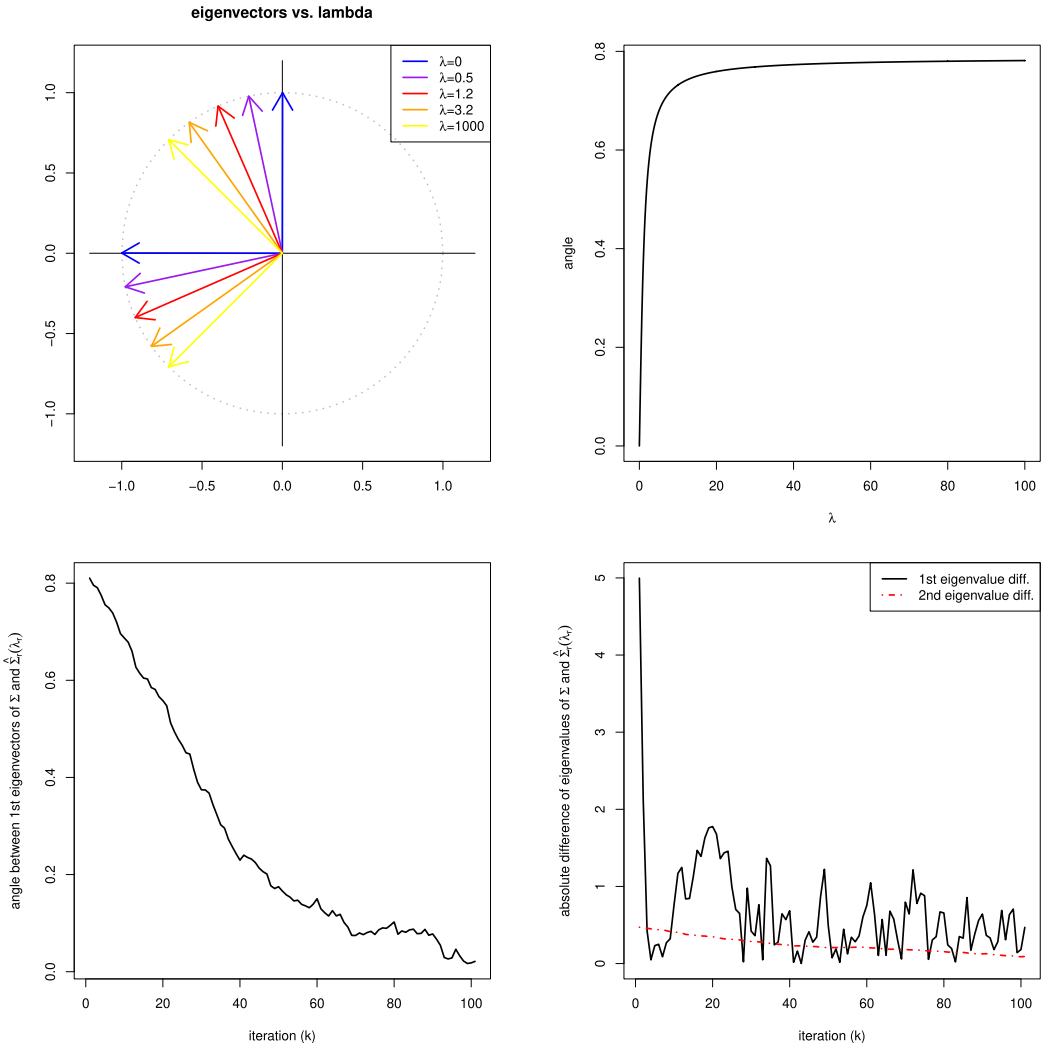

Fig. 1. Top left panel: eigenvector of $\boldsymbol{\Omega}_r(\lambda)$ for various choices of λ. Top tight panel: the angle between the first eigenvector of $\mathbf{S}$ and that of $\boldsymbol{\Omega}_r(\lambda)$ (i.e., $\angle(\mathbf{v}_{s,1}, \mathbf{v}_{\omega(\lambda),1})$) vs. λ. Bottom left panel: angle between the eigenvectors of $\boldsymbol{\Sigma}$ and its kth updated estimate. Bottom right panel: eigenvalue difference between that of $\boldsymbol{\Sigma}$ and its kth updated estimate.

(iii) *can be described by the rotation* $\mathbf{V}_{\omega(\lambda)} = \mathbf{R}_\lambda \mathbf{V}_s$ *with* $\mathbf{R}_\lambda$ *a rotation matrix;*
(iv) *is constant if* $\mathbf{V}_s = \mathbf{V}_t$. *In particular, when in addition the eigenvalues of* $\mathbf{S}$ *and* $\mathbf{T}$ *are reciprocal, we have* $\widehat{\boldsymbol{\Omega}}(\lambda) = \mathbf{S}^{-1}$ *for all* λ *(provided* $\mathbf{S}^{-1}$ *exists).*

With respect to updating the above theorem implies that the angle between corresponding eigenvectors and distance between corresponding eigenvalues of $\boldsymbol{\Sigma}$ and its estimate will vanish. This is illustrated for the updating scheme of the ridge estimator by means of a simple 2×2-dimensional example. The scheme starts with an initial guess $\boldsymbol{\Omega}_0$ unit diagonal and off-diagonal element equaling 0.9. Then, ten samples are drawn from $\mathcal{N}(\mathbf{0}_2, \boldsymbol{\Sigma})$ with $(\boldsymbol{\Sigma})_{11} = 5$, $(\boldsymbol{\Sigma})_{22} = 1$, and $(\boldsymbol{\Sigma})_{12} = 0.1 = (\boldsymbol{\Sigma})_{21}$. From these data the first estimate of the covariance matrix, $\widehat{\boldsymbol{\Sigma}}_{r,1}(\lambda_1)$ with $\lambda_1 = 100$, is obtained. This step is repeated a hundred times with the same penalty parameter choice throughout, each time the latest estimate serves as input for the next one, thus in accordance with updating scheme (1). The results, angle and eigenvalue distance, are plotted in the lower panels of Fig. 1. Convergence of both quantities is observed.

4. Practical complications

In practice issues may arise that hamper updating or require down-stream manipulation of the estimate. Here two such anticipated complications are identified and possible solutions are outlined.

4.1. Batch effect

A target matrix may be constructed from pilot data, but batch effects may cause differences between the pilot and current data. Amongst others these differences may exhibit themselves in the scale. It may then be desirable to scale the target derived from the pilot data derived to match that of the current data. More precise, the target matrix $\mathbf{T}$ to be used for the current data is related to the covariance estimate obtained from the pilot data through a multiplicative constant: $\mathbf{T} = \gamma\, \widehat{\boldsymbol{\Sigma}}^{-1}_{\text{pilot}}$. Effectively, this uses the data once-and-a-little bit, much like an empirical Bayes procedure — which learns few low-dimensional hyperparameters from data. It remains to derive an estimator for γ, for example through likelihood maximization. The log-likelihood, $\log|\gamma^{-1}\widehat{\boldsymbol{\Sigma}}^{-1}_{\text{pilot}}| - \text{tr}(\mathbf{S}\,\gamma^{-1}\widehat{\boldsymbol{\Sigma}}^{-1}_{\text{pilot}})$, is maximized by the estimator $\hat{\gamma} = \text{tr}(\mathbf{S}\widehat{\boldsymbol{\Sigma}}^{-1}_{\text{pilot}})/p$, which can be used to correct the scale of the target.

An alternative batch correction approach would be to Gaussianize the data using the probability integral transform [24]. Apart from the convenient fact that Gaussianization forces the data to adhere to our normality assumption, it replaces this assumption by that of a nonparanormal distribution [24], effectively, a Gaussian copula model. [24] show that the Gaussianization preserves the conditional independence structure underlying the data. After Gaussianization data from different studies are more likely to be modeled by a distribution of the same normal form, as assumed throughout this work.

4.2. Multiple targets

Multiple sources may provide information on the precision matrix. When each source yields its own quantitative suggestion on $\boldsymbol{\Omega}$, multiple targets, denoted $\mathbf{T}_1, \ldots, \mathbf{T}_G$ and each being positive definite, are available. For the Ledoit–Wolf shrinkage covariance estimator [14] show how multiple targets may be dealt with. For the ridge precision estimator these targets can be included in the estimation of $\boldsymbol{\Omega}$ simultaneously through a weighted ridge penalty. This leads to the loss function:

$$\log(|\boldsymbol{\Omega}|) - \text{tr}(\mathbf{S}\boldsymbol{\Omega}) - \lambda \sum_{g=1}^{G} \alpha_g \|\boldsymbol{\Omega} - \mathbf{T}_g\|_2^2,$$

with $\alpha_1 + \cdots + \alpha_G = 1$. The penalty is thus a weighted average of the individual ridge penalties. The weighted average of concave ridge penalties is itself concave and, consequently, so is this penalized log-likelihood, thereby ensuring a unique maximum. Finally, note that the size of the α_g's determines how informative the corresponding target $\mathbf{T}_g$ is for the estimation of the precision matrix.

The weighted ridge penalty induces, when the maximization of the penalized likelihood is reformulated as a constrained estimation problem, a parameter constraint. For illustration purposes the constraint, $\{\boldsymbol{\Omega} \succ 0 : \sum_{k=1}^{K} \alpha_k \|\boldsymbol{\Omega} - \mathbf{T}_g\|_2^2 \leq c(\lambda)\}$ with $c(\lambda)$ some positive function of λ, is visualized in the Supplementary Material IV for the case a 2×2-dimensional precision matrix with two targets and various choices of α. For each choice of α, the constraint is spherical, as to be expected from a ridge-type penalty, with its center at the average of the targets, weighted by the α. The sphere is smallest in diameter for α being uniform, while largest when $\alpha_g = 1$ for any of the $g = 1, \ldots, G$.

An explicit expression of the estimator, given λ and α, can straightforwardly be derived. The estimating equation of $\boldsymbol{\Omega}$, after following a derivation analogous to that presented in [42], is:

$$\boldsymbol{\Omega}^{-1} - \mathbf{S} - \lambda\boldsymbol{\Omega} + \lambda\bar{\mathbf{T}} = \mathbf{0}_{pp},$$

where $\bar{\mathbf{T}} = \sum_{g=1}^{G} \alpha_g \mathbf{T}_g$. The estimator of $\boldsymbol{\Omega}$ then equals the root of this equation, which is (cf. [42]):

$$\widehat{\boldsymbol{\Omega}}(\lambda, \alpha) = \{\tfrac{1}{2}(\mathbf{S} - \lambda\bar{\mathbf{T}}) + [\lambda\mathbf{I}_{pp} + \tfrac{1}{4}(\mathbf{S} - \lambda\bar{\mathbf{T}})^2]^{1/2}\}^{-1}, \tag{3}$$

where $\alpha = (\alpha_1, \ldots, \alpha_G)$. The properties of the ridge precision estimator as formulated in the introduction then carry over to the estimator above:

Theorem 5. *Let $\widehat{\boldsymbol{\Omega}}(\lambda, \alpha)$ be defined as in Display (3). Then*

(i) $\widehat{\boldsymbol{\Omega}}(\lambda, \alpha) \succ 0$;
(ii) $\lim_{\lambda \downarrow 0} \widehat{\boldsymbol{\Omega}}(\lambda, \alpha) = \mathbf{S}^{-1}$ *(provided* $\mathbf{S} \succ 0$*)*;
(iii) $\lim_{\lambda \to \infty} \widehat{\boldsymbol{\Omega}}(\lambda, \alpha) = \bar{\mathbf{T}}$;
(iv) $\widehat{\boldsymbol{\Omega}}(\lambda, \alpha)$ *is a consistent estimator of* $\boldsymbol{\Omega}$ *if* $\lambda_n \xrightarrow{P} 0$ *for* $n \to \infty$;
(v) *for a suitable choice of* λ, *the precision estimator* $\widehat{\boldsymbol{\Omega}}(\lambda, \alpha)$ *outperforms the ML precision estimator in terms of the mean squared error.*

An informed choice of the λ and α is made by cross-validation. Cross-validation is feasible as practice suggests that the expected number of available targets is unlikely to exceed three, thus limiting the number of parameters to be determined by cross-validation. Cross-validation divides the samples into a pre-specified number of non-overlapping and exhaustive subsets. Each subset is left out once, while the remaining subsets are then used to estimate the parameters of interest for

given λ and α. The performance of the thus obtained parameter estimates is evaluated through the log-likelihood of the left out subset. The average of these performances reflects the quality of the employed λ and α. The λ and α that yield the maximum cross-validated log-likelihood are considered optimal. The Supplementary Material IV contains a simulation that reveals that ridge estimation with multiple targets and cross-validated parameters λ and α can indeed distinguish between various targets.

4.3. Edge selection

For display and communications purposes the conditional independence graph underlying the estimated Gaussian graphical model needs to be inferred. To select edges [42] invoke the work of [12] and [34] to arrive at a probabilistically motivated threshold on the estimated partial correlations. That work is also applied here. It assumes that in the underlying conditional independence network most edges are absent. The fraction of absent edges is denoted by π_0. In the precision matrix the absent/present edges correspond to a zero/nonzero diagonal element. This also holds for the corresponding partial correlations. The marginal distribution of the latter is then assumed to follow a mixture of the form: $\pi_0 f_0(r) + (1-\pi_0) f_1(r)$, where $f_0(r)$ and $f_1(r)$ are the densities of the absent and present edges, respectively. By the assumption that most of the edges are absent, most partial correlations follow the $f_0(\cdot)$-law. Using the (say) 80% percentage of the partial correlations closest to zero, the implementation of [34] estimates $f_0(\cdot)$ by means of a truncated likelihood approach. With $f_0(\cdot)$ available, one can now calculate the probability of an edge being absent giving the observed partial correlation. This probability can be endowed with a local FDR (False Discovery Rate) interpretation (see [12]). A cut-off on this probability is then to be chosen and used as the selection criterion.

5. Conclusion and discussion

We studied the problem of learning a Gaussian graphical model from an experiment that is carried out repeatedly. When data from the next experiment become available, the current estimate of the model parameter needs updating. We proposed to re-estimate the parameter by means of shrinkage estimators that strike a balance between the latest estimate of the parameter and the new experimental data. The resulting estimators yield with high probability an improvement in fit for subsequent experiments. Moreover, the sequence of estimators was shown to be asymptotically unbiased and consistent. Intuition behind the workings of updating was provided by linking it to Bayesian updating and studying the relation between the eigenspaces of the current and updated estimator. Next, the effect of updating on moments and loss of the estimators was discussed. Practical complications that may be encountered when updating were identified and possible solutions outlined. Finally, the proposed updating procedure was illustrated on pathway data from five consecutively published breast cancer studies.

The employed shrinkage-type estimators, endowed with a post-estimation selection procedure, may be replaced by a lasso equivalent (as presented in [41]). This equivalent yields sparse estimates, sparse not in the sense of the vast majority of parameters being zero but many elements of the updated parameter estimate not deviating from the corresponding elements of the current one. Selection by this lasso estimator thus refers to the fact that the data give rise to update a particular parameter element to a novel value.

Follow-up work will concentrate on a choice of the penalty/shrinkage parameters. Cross-validation (employed here) may be improved upon – perhaps via constraints – to yield a $\{\lambda_k\}_{k=1}^{\infty}$ and $\{\nu_k\}_{k=1}^{\infty}$ scheme that ensures fast convergence. Moreover, as for certain such schemes both estimators exhibit the behavior of a weighted average, the weighing scheme may robustify the estimators against an outlying sample.

Another issue to be addressed in future research is the possibility of heterogeneity at the sample level. To deal with such heterogeneity one commonly assumes a mixture model, here a mixture of Gaussian graphical models [1]. This mixture model, assumed to be static, is to be updated over data sets, for which the penalized estimation procedure of [1] can be used. This brings about several issues that need resolving: (i) it needs to be decided on the outset (or via some model selection procedure) how many components are to be included in the mixture; (ii) the label switching problem becomes prominent and an efficient algorithm to address this is required; (iii) samples from each component of the mixture may be underrepresented in some data sets and the EM algorithm needs to deal with small mixing proportions without risking overfitting; (iv) finally, of course, it remains to translate the theoretical results for a single Gaussian graphical model to a mixture of such models.

Finally, the domain of application of the proposed methodology extends beyond that of omics-data, in particular to epidemiology. Examples of large and long-running epidemiological studies are the LASA-study (https://www.lasa-vu.nl/index.htm), the Hoorn-study (https://hoornstudies.com/), and the NaKo-study (https://nako.de/informationen-auf-englisch/). Weekly or monthly updates, i.e. interim-analysis, result in much larger K than that of the presented application.

6. Proofs

Prior to the proof of Proposition 1 several prerequisites are stated.

Lemma 1. *Assume* $\mathbf{T}_r^{-1} \neq \mathbf{S}$. *Then,* $\|\widehat{\boldsymbol{\Sigma}}_r(\lambda) - \mathbf{S}\|_F^2$ *is strictly increasing in* λ.

Proof. The proof proceeds by showing that the derivatives of the squared norm with respect to λ is strictly positive. Write the squared Frobenius norm as a trace and expand its argument: $\|\widehat{\boldsymbol{\Sigma}}_r(\lambda)-\mathbf{S}\|_F^2 = \text{tr}[\widehat{\boldsymbol{\Sigma}}_r^2(\lambda)] - 2\text{tr}[\widehat{\boldsymbol{\Sigma}}_r(\lambda)\mathbf{S}] + \text{tr}(\mathbf{S}^2)$. The derivative w.r.t. λ of the right-hand side is:

$$\frac{d}{d\lambda}\|\widehat{\boldsymbol{\Sigma}}_r(\lambda)-\mathbf{S}\|_F^2 = 2\text{tr}\Big\{\big[\widehat{\boldsymbol{\Sigma}}_r(\lambda)-\mathbf{S}\big]\frac{d}{d\lambda}\widehat{\boldsymbol{\Sigma}}_r(\lambda)\Big\}, \tag{4}$$

where: $\frac{d}{d\lambda}\widehat{\boldsymbol{\Sigma}}(\lambda) = \frac{1}{2}[\lambda\mathbf{I}_{pp} + \frac{1}{4}(\mathbf{S}-\lambda\mathbf{T}_r)^2]^{-1/2}\widehat{\boldsymbol{\Sigma}}_r(\lambda)[\widehat{\boldsymbol{\Sigma}}_r^{-1}(\lambda) - \mathbf{T}_r]$. Substitute the latter into Eq. (4) and use that $\widehat{\boldsymbol{\Sigma}}_r(\lambda)$ satisfies the estimating equation, i.e. $\widehat{\boldsymbol{\Sigma}}_r(\lambda) - \mathbf{S} = \lambda[\widehat{\boldsymbol{\Sigma}}_r^{-1}(\lambda) - \mathbf{T}_r]$, to arrive at:

$$\frac{d}{d\lambda}\|\widehat{\boldsymbol{\Sigma}}_r(\lambda)-\mathbf{S}\|_F^2 = \lambda\text{tr}\Big\{[\lambda\mathbf{I}_{pp} + \tfrac{1}{4}(\mathbf{S}-\lambda\mathbf{T}_r)^2]^{-1/2}\widehat{\boldsymbol{\Sigma}}_r(\lambda)[\widehat{\boldsymbol{\Sigma}}_r^{-1}(\lambda) - \mathbf{T}_r]^2\Big\}.$$

This trace is positive, whenever each term in the trace is positive definite. Only the last term, $[\widehat{\boldsymbol{\Sigma}}_r^{-1}(\lambda) - \mathbf{T}_r]^2$, may be semi-positive definite, but only if $\widehat{\boldsymbol{\Sigma}}_r^{-1}(\lambda) = \mathbf{T}_r$. This occurs either when $\mathbf{S} = \mathbf{T}_r^{-1}$ (which is excluded by the conditions of the lemma) or in the limit $\lambda \to \infty$. □

A direct consequence of this lemma is stated next.

Corollary 2. *Consider the sequence of sample covariance matrices $\{\mathbf{S}_k\}_{k=1}^{\infty}$ formed from different draws of the same $\mathcal{N}(\mathbf{0}_p, \boldsymbol{\Sigma})$-law and let $\{\widehat{\boldsymbol{\Sigma}}_{r,k}(\lambda_k)\}_{k=1}^{\infty}$ be the sequences of corresponding updated ridge covariance estimators, respectively. Then, given current covariance estimate $\widehat{\boldsymbol{\Sigma}}_{r,k}(\lambda_k)$, the updated ones improve the fit: $\|\widehat{\boldsymbol{\Sigma}}_{r,k+1}(\lambda_{k+1}) - \mathbf{S}_{k+1}\|_F^2 < \|\widehat{\boldsymbol{\Sigma}}_{r,k}(\lambda_k) - \mathbf{S}_{k+1}\|_F^2$, for all $\lambda_{k+1} \in [0, \infty)$.*

The following corollary follows from Lemma 12 in the supplementary material of [19].

Corollary 3. *Let $\mathbf{S}_1$ and $\mathbf{S}_2$ be sample covariance matrices obtained from samples with size n_k and n_{k+1}, respectively, drawn from $\mathcal{N}(\mathbf{0}_p, \boldsymbol{\Sigma})$. Denote by $\mathbf{R}$ the correlation matrix associated with $\boldsymbol{\Sigma}$. Define $\xi_{j,j'} = \max\{[1-(\mathbf{R})_{j,j'}][(\boldsymbol{\Sigma})_{j,j}(\boldsymbol{\Sigma})_{j',j'}]^{1/2}, [1+(\mathbf{R})_{j,j'}][(\boldsymbol{\Sigma})_{j,j}(\boldsymbol{\Sigma})_{j',j'}]^{1/2}\}$. Then, for all $t \in [0, \frac{1}{2}p^2\xi_{j,j'}^2)$:*

$$P(\|\mathbf{S}_1 - \boldsymbol{\Sigma}\|_F^2 + \|\mathbf{S}_2 - \boldsymbol{\Sigma}\|_F^2 \geq t) \leq \min\Big\{1, 4\sum_{j=1,j'=j}^{p}\big\{\exp[-3n_k t^2/(16\xi_{j,j'}^2)] + \exp[-3n_{k+1}t^2/(16\xi_{j,j'}^2)]\big\}\Big\}.$$

In particular, when ξ is defined as $\max_{j,j'}\xi_{j,j'}$, for all $t \in [0, \frac{1}{2}p^2\xi^2)$:

$$\begin{aligned}&P(\|\mathbf{S}_1 - \boldsymbol{\Sigma}\|_F^2 + \|\mathbf{S}_2 - \boldsymbol{\Sigma}\|_F^2 \geq t)\\ &\quad\leq \min\Big\{1, 2\big\{\exp[\log(p^2+p) - 3n_k t^2/(16\xi^2)] + \exp[\log(p^2+p) - 3n_{k+1}t^2/(16\xi^2)]\big\}\Big\}.\end{aligned}$$

Proof. The proof bounds the probability of the complementary event:

$$\begin{aligned}P\big(\|\mathbf{S}_k - \boldsymbol{\Sigma}\|_F^2 + \|\mathbf{S}_{k+1} - \boldsymbol{\Sigma}\|_F^2 < t\big) \;&\geq\; P\big(\|\mathbf{S}_k - \boldsymbol{\Sigma}\|_F^2 < \tfrac{1}{2}t, \|\mathbf{S}_{k+1} - \boldsymbol{\Sigma}\|_F^2 < \tfrac{1}{2}t\big)\\
&\geq P\big(|(\mathbf{S}_k)_{j,j'} \quad - (\boldsymbol{\Sigma})_{j,j'}| < \tilde{t}, j \in \{1,\ldots,p\}, j' \in \{j,\ldots,p\},\\
&\quad\; |(\mathbf{S}_{k+1})_{j,j'} - (\boldsymbol{\Sigma})_{j,j'}| < \tilde{t}, j \in \{1,\ldots,p\}, j' \in \{j,\ldots,p\},\big)\\
&\geq \max\Big\{0, \sum_{j=1,j'=j}^{p} P\big(|(\mathbf{S}_k)_{j,j'} \quad - (\boldsymbol{\Sigma})_{j,j'}| < \tilde{t}\big)\\
&\qquad + \sum_{j=1,j'=j}^{p} P\big(|(\mathbf{S}_{k+1})_{j,j'} - (\boldsymbol{\Sigma})_{j,j'}| < \tilde{t}\big) - [2 \times \tfrac{1}{2}(p^2+p) - 1]\Big\}\\
&= \max\Big\{0, 1 - \sum_{j=1,j'=j}^{p}[P\big(|(\mathbf{S}_k)_{j,j'} \quad - (\boldsymbol{\Sigma})_{j,j'}| \geq \tilde{t}\big) + P\big(|(\mathbf{S}_{k+1})_{j,j'} - (\boldsymbol{\Sigma})_{j,j'}| \geq \tilde{t}\big)]\Big\}\\
&\geq \max\Big\{0, 1 - \sum_{j=1,j'=j}^{p} 4\{\exp[-3n_k\tilde{t}^2/(16\xi_{j,j'}^2)] + \exp[-3n_{k+1}\tilde{t}^2/(16\xi_{j,j'}^2)]\}\Big\}\end{aligned}$$

where the j' runs from j to p in order to exclude duplicate events stemming from the symmetry of $\mathbf{S}_{k+1}$ and $\boldsymbol{\Sigma}$, which do not add to the probability, $\tilde{t} = 2^{-1/2}p^{-1}t^{1/2}$, Fréchet's inequality to the probability of the conjunction of $2 \times \frac{1}{2}(p^2+p)$ has been applied, and, finally, Lemma 12 of the supplementary material of [19] with $\tilde{t} \in [0, \frac{1}{2}\xi_{j,j'})$. □

We are now ready to prove Proposition 1.

Proof of Proposition 1. From Corollary 2 it follows that for any $\lambda_{k+1} \in (0, \infty)$ there exists a $\delta(\lambda_{k+1}) > 0$ such that:

$$\begin{aligned}\|\widehat{\boldsymbol{\Sigma}}_{r,k+1}(\lambda_{k+1}) - \mathbf{S}_{k+1}\|_F^2 + \delta(\lambda_{k+1}) &= \|\widehat{\boldsymbol{\Sigma}}_{r,k}(\lambda_k) - \mathbf{S}_{k+1}\|_F^2 \\ &\leq \|\widehat{\boldsymbol{\Sigma}}_{r,k}(\lambda_k) - \mathbf{S}_k\|_F^2 + \|\mathbf{S}_k - \boldsymbol{\Sigma}\|_F^2 + \|\mathbf{S}_{k+1} - \boldsymbol{\Sigma}\|_F^2,\end{aligned}$$

where the triangle inequality has been applied repeatedly. Hence, if $\|\mathbf{S}_k - \boldsymbol{\Sigma}\|_F^2 + \|\mathbf{S}_{k+1} - \boldsymbol{\Sigma}\|_F^2 < \delta(\lambda_{k+1})$, then $\|\widehat{\boldsymbol{\Sigma}}_{r,k+1}(\lambda_{k+1}) - \mathbf{S}_{k+1}\|_F^2 < \|\widehat{\boldsymbol{\Sigma}}_{r,k}(\lambda_k) - \mathbf{S}_k\|_F^2$. The probability of this happening can, using Corollary 3, be bounded as:

$$\begin{aligned}&P\big(\|\mathbf{S}_k - \boldsymbol{\Sigma}\|_F^2 + \|\mathbf{S}_{k+1} - \boldsymbol{\Sigma}\|_F^2 < \delta(\lambda_{k+1})\big) = 1 - P\big(\|\mathbf{S}_k - \boldsymbol{\Sigma}\|_F^2 + \|\mathbf{S}_{k+1} - \boldsymbol{\Sigma}\|_F^2 \geq \delta(\lambda_{k+1})\big) \\ &\quad \geq 1 - \min\Big\{1, 2\big\{\exp[\log(p^2+p) - 3n_k t^2/(16\xi^2)] + \exp[\log(p^2+p) - 3n_{k+1} t^2/(16\xi^2)]\big\}\Big\},\end{aligned}$$

with $t = \min\{\delta(\lambda_{k+1}), \frac{1}{2}p^2\xi^2\}$ and ξ defined as in Corollary 3. □

Prior to the proof of Proposition 2 a prerequisite is stated.

Lemma 2. *Assume $\mathbf{T}_r^{-1} \neq \mathbf{S}$. Then, for $q \in \mathbb{N}$, $\|\widehat{\boldsymbol{\Sigma}}_r(\lambda) - \mathbf{S}\|_q$ is strictly increasing in λ. In particular, the monotony of the difference persists into the spectral norm ($\|\cdot\|_\infty$).*

Proof. Analogous to Lemma 1, the proof proceeds by showing that the derivative of the norm with respect to λ is strictly positive. Write the Schatten q-norm as a trace: $\|\widehat{\boldsymbol{\Sigma}}_r(\lambda) - \mathbf{S}\|_q = \big[\mathrm{tr}\big(\{[\widehat{\boldsymbol{\Sigma}}_r^2(\lambda) - \mathbf{S}]^2\}^{q/2}\big)\big]^{1/q}$ (where $[\cdot]^q$ is written as $\{[\cdot]^2\}^{q/2}$ is to emphasize and ensure the positiveness of the quantity under study). The derivative w.r.t. λ of the right-hand side is:

$$\frac{d}{d\lambda}\|\widehat{\boldsymbol{\Sigma}}_r(\lambda) - \mathbf{S}\|_q = (\|\widehat{\boldsymbol{\Sigma}}_r(\lambda) - \mathbf{S}\|_q)^{1-q}\mathrm{tr}\Big(\{\big[\widehat{\boldsymbol{\Sigma}}_r(\lambda) - \mathbf{S}\big]^2\}^{q/2-1}\big[\widehat{\boldsymbol{\Sigma}}_r(\lambda) - \mathbf{S}\big]\frac{d}{d\lambda}\widehat{\boldsymbol{\Sigma}}_r(\lambda)\Big), \tag{5}$$

where $\frac{d}{d\lambda}\widehat{\boldsymbol{\Sigma}}(\lambda)$ is as in Lemma 1). Substitute the latter into Eq. (5) and use that $\widehat{\boldsymbol{\Sigma}}_r(\lambda)$ satisfies the estimating equation, i.e. $\widehat{\boldsymbol{\Sigma}}_r(\lambda) - \mathbf{S} = \lambda[\widehat{\boldsymbol{\Sigma}}_r^{-1}(\lambda) - \mathbf{T}_r]$, to arrive at:

$$\frac{d}{d\lambda}\|\widehat{\boldsymbol{\Sigma}}_r(\lambda) - \mathbf{S}\|_q = (\|\widehat{\boldsymbol{\Sigma}}_r(\lambda) - \mathbf{S}\|_q)^{1-q}\lambda^{-1}\mathrm{tr}\big([\lambda\mathbf{I}_{pp} + \tfrac{1}{4}(\mathbf{S} - \lambda\mathbf{T}_r)^2]^{-1/2}\widehat{\boldsymbol{\Sigma}}_r(\lambda)\{[\widehat{\boldsymbol{\Sigma}}_r(\lambda) - \mathbf{S}]^2\}^{q/2}\big).$$

This trace is positive, whenever each term in the trace is positive definite. Only the last term, $[\widehat{\boldsymbol{\Sigma}}_r(\lambda) - \mathbf{S}]^2$, may be semi-positive definite, but only if $\widehat{\boldsymbol{\Sigma}}_r(\lambda) = \mathbf{S}$. This occurs either when $\mathbf{S} = \mathbf{T}_r^{-1}$ (which is excluded by the conditions of the lemma) or in the limit $\lambda \downarrow 0$. By l'Hopital's rule the derivative is still positive in this limit.

Extra work is needed to prove the monotony of the spectral norm ($q = \infty$). Hereto we study the sequence $\{\|\widehat{\boldsymbol{\Sigma}}_r(\lambda) - \mathbf{S}\|_q\}_{q=0}^\infty$ and that of its derivative with respect to λ. Both sequences are bounded as $\|\widehat{\boldsymbol{\Sigma}}_r(\lambda) - \mathbf{S}\|_q \leq \|\widehat{\boldsymbol{\Sigma}}_r(\lambda)\|_q$ by Corollary 4.1.3 of [17]. For the sequence of derivatives, we have (by the nonnegativity of the trace of a product of two semi-positive definite matrices):

$$\frac{d}{d\lambda}\|\widehat{\boldsymbol{\Sigma}}_r(\lambda) - \mathbf{S}\|_q \leq \lambda^{-1}d_{\max}(\|\widehat{\boldsymbol{\Sigma}}_r(\lambda) - \mathbf{S}\|_q)^{1-q}\mathrm{tr}\big(\{[\widehat{\boldsymbol{\Sigma}}_r(\lambda) - \mathbf{S}]^2\}^{q/2}\big) = \lambda^{-1}d_{\max}\|\widehat{\boldsymbol{\Sigma}}_r(\lambda) - \mathbf{S}\|_q \leq \lambda^{-1}d_{\max}\|\widehat{\boldsymbol{\Sigma}}_r(\lambda)\|_q,$$

where $d_{\max}$ is the largest eigenvalue of $[\lambda\mathbf{I}_{pp} + \frac{1}{4}(\mathbf{S} - \lambda\mathbf{T}_r)^2]^{-1/2}\widehat{\boldsymbol{\Sigma}}_r(\lambda)$, which, by the positive definiteness of this matrix product, is positive. The limit of $\{\|\widehat{\boldsymbol{\Sigma}}_r(\lambda) - \mathbf{S}\|_q\}_{q=0}^\infty$ is the spectral radius, that is, the largest (in the absolute sense) eigenvalue as follows from Gelfand's formula. By the smoothness and boundedness of the map $\lambda \mapsto \widehat{\boldsymbol{\Sigma}}_r(\lambda) - \mathbf{S}$ and the definition of the spectral radius the point-wise limit is itself smooth. Due to their boundedness these sequences (of continuous functions) converge point-wise. In particular, by Dini's theorem [30] they converge uniformly on any compact interval of λ. We may now invoke Theorem 7.17 of [30] to conclude:

$$\frac{d}{d\lambda}\|\widehat{\boldsymbol{\Sigma}}_r(\lambda) - \mathbf{S}\|_\infty = \lim_{q\to\infty}\frac{d}{d\lambda}\|\widehat{\boldsymbol{\Sigma}}_r(\lambda) - \mathbf{S}\|_q \geq \lambda^{-1}d_{\min}\|\widehat{\boldsymbol{\Sigma}}_r(\lambda) - \mathbf{S}\|_\infty > 0,$$

where the inequality originates from the same argument (here used in the opposite direction) as that in the preceding display and $d_{\min}$ denotes the smallest eigenvalue of $[\lambda\mathbf{I}_{pp} + \frac{1}{4}(\mathbf{S} - \lambda\mathbf{T}_r)^2]^{-1/2}\widehat{\boldsymbol{\Sigma}}_r(\lambda)$, which, by the positive definiteness of this matrix product, is positive. □

We are now ready to prove Proposition 2.

Proof of Proposition 2. From Lemma 2 it follows that for every $\lambda_{k+1} \in (0, \infty)$ there exists a $\delta(\lambda_{k+1}) > 0$ such that:

$$\|\widehat{\boldsymbol{\Sigma}}_{r,k+1}(\lambda_{k+1}) - \mathbf{S}_{k+1}\|_\infty + \delta(\lambda_{k+1}) = \|\widehat{\boldsymbol{\Sigma}}_{r,k}(\lambda_k) - \mathbf{S}_{k+1}\|_\infty \leq \|\widehat{\boldsymbol{\Sigma}}_{r,k}(\lambda_k) - \mathbf{S}_k\|_\infty + \|\mathbf{S}_k - \boldsymbol{\Sigma}\|_\infty + \|\mathbf{S}_{k+1} - \boldsymbol{\Sigma}\|_\infty,$$

where the triangle inequality has been applied repeatedly. Hence, if $\|\mathbf{S}_k - \boldsymbol{\Sigma}\|_\infty + \|\mathbf{S}_{k+1} - \boldsymbol{\Sigma}\|_\infty < \delta(\lambda_{k+1})$, then $\|\widehat{\boldsymbol{\Sigma}}_{r,k+1}(\lambda_{k+1}) - \mathbf{S}_{k+1}\|_\infty < \|\widehat{\boldsymbol{\Sigma}}_{r,k}(\lambda_k) - \mathbf{S}_k\|_\infty$. The probability of this happening can, using Corollary 5.50, [37], be bounded:

$$\begin{aligned} P\big(\|\mathbf{S}_k - \boldsymbol{\Sigma}\|_\infty + \|\mathbf{S}_{k+1} - \boldsymbol{\Sigma}\|_\infty < \delta(\lambda_{k+1})\big) &\geq P\big(\|\mathbf{S}_k - \boldsymbol{\Sigma}\|_\infty < \tfrac{1}{2}\delta(\lambda_{k+1}), \|\mathbf{S}_{k+1} - \boldsymbol{\Sigma}\|_\infty < \tfrac{1}{2}\delta(\lambda_{k+1})\big) \\ &\geq \max\big\{0, P\big(\|\mathbf{S}_k - \boldsymbol{\Sigma}\|_\infty \leq \tfrac{1}{2}\delta(\lambda_{k+1})\big) + P\big(\|\mathbf{S}_{k+1} - \boldsymbol{\Sigma}\|_\infty \leq \tfrac{1}{2}\delta(\lambda_{k+1})\big) - 1\big\} \;\geq\; \max\{0, 1 - 4\exp(-t^2 p)\}, \end{aligned}$$

for $n_{k+1} \geq C[2t/\delta(\lambda_{k+1})]^2 p$. □

Proof of Theorem 1. Define $\widehat{\boldsymbol{\Sigma}}_k(\nu_k) = (1-\nu_k)\mathbf{S}_k + \nu_k \mathbf{T}_{\ell w,k}$. Let $\mathbf{T}_{\ell w,k} = \widehat{\boldsymbol{\Sigma}}_{k-1}(\nu_{k-1})$. Then:

$$\widehat{\boldsymbol{\Sigma}}_{\ell w,k}(\nu_k) = \sum_{\kappa=1}^{k}\Big[\prod_{\ell=\kappa+1}^{k}\nu_\ell\Big](1-\nu_\kappa)\mathbf{S}_\kappa + \Big[\prod_{\kappa=1}^{k}\nu_\kappa\Big]\mathbf{T}_{\ell w,1}.$$

This is a weighted average of all sample covariance matrices, placing more weight on the more recent ones, and the initial target matrix. In the $k \to \infty$ limit the second summand of right-hand side of the preceding display will vanish for all $\nu_k \in (0,1)$. That is: $\lim_{k\to\infty} \widehat{\boldsymbol{\Sigma}}_{\ell w,k}(\nu_k) = \lim_{k\to\infty} \sum_{\kappa=1}^{k}\big[\prod_{\ell=\kappa+1}^{k}\nu_\ell\big](1-\nu_\kappa)\mathbf{S}_\kappa$. Each of these sample covariances is an unbiased estimator of $\boldsymbol{\Sigma}$. Hence, so is the weighted average: $\lim_{k\to\infty}\mathbb{E}[\widehat{\boldsymbol{\Sigma}}_{\ell w,k}(\nu_k)] = \boldsymbol{\Sigma}$. □

Proof of Theorem 2. The proof proceeds by showing the existence of a stationary density of the Markov process defined by the updating of the ridge covariance estimators. Then, using stationarity, the claimed result is easily seen from the estimating equation.

The conditions for the existence of a stationary density of a discrete time, time-homogeneous Markov process with a continuous state space are specified in Theorem 8.2.14 of [32]. By this theorem it suffices to show for the process at hand that (i) it is irreducible, i.e. it satisfies the mixing condition, (ii) it exhibits geometric drift to the center, and (iii) the sequence of its marginal densities is uniformly integrable. Conditions (i) and (ii) are checked next, as the uniform integrability follows from a general argument laid out in [32] and that is applicable here.

The mixing condition requires to show that any $\mathbf{U} \in \mathcal{S}_{++}$ reachable from $\mathbf{U}' \in \mathcal{S}_{++}$ with positive probability. From the analytic expression of the estimator,

$$\widehat{\boldsymbol{\Sigma}}_{r,k+1}(\lambda_{k+1}) = \tfrac{1}{2}[\mathbf{S}_{k+1} - \lambda_{k+1}\widehat{\boldsymbol{\Sigma}}_{r,k}^{-1}(\lambda_k)] + \{\lambda_{k+1}\mathbf{I}_{pp} + \tfrac{1}{4}[\mathbf{S}_{k+1} - \lambda_{k+1}\widehat{\boldsymbol{\Sigma}}_{r,k}^{-1}(\lambda_k)]^2\}^{1/2},$$

it is immediate that any $\mathbf{S}_{k+1}$ of the form $\mathbf{S}_{k+1} = \lambda_{k+1}\widehat{\boldsymbol{\Sigma}}_{r,k}^{-1}(\lambda_k) + \mathbf{U}''$ will remove the influence of the preceding (time-wise) estimator. The problem now reduces to showing that an $\mathbf{S}_{k+1}$ of this form may assume any value in $\mathcal{S}_{++}$ with positive probability. From the estimating equation, $\mathbf{S}_{k+1} = \widehat{\boldsymbol{\Sigma}}_{r,k+1}(\lambda_{k+1}) - \lambda_{k+1}\widehat{\boldsymbol{\Sigma}}_{r,k+1}^{-1}(\lambda_{k+1}) + \lambda_{k+1}\widehat{\boldsymbol{\Sigma}}_{r,k}^{-1}(\lambda_k)$, it is clear that $\mathbf{U}''$ shares its eigenspace with that of $\widehat{\boldsymbol{\Sigma}}_{r,k+1}(\lambda_{k+1})$. Let $\mathbf{D}_{\sigma,k+1}$ denote a diagonal matrix containing the eigenvalues of $\widehat{\boldsymbol{\Sigma}}_{r,k+1}(\lambda_{k+1})$. That of $\mathbf{U}''$ then needs to equal $\mathbf{D}_{\sigma,k+1} - \lambda_{k+1}\mathbf{D}_{\sigma,k+1}^{-1}$ to warrant that the updated estimator indeed equals $\widehat{\boldsymbol{\Sigma}}_{r,k+1}(\lambda_{k+1})$. Rests to verify that the required $\mathbf{S}_{k+1}$ is symmetric and positive definite, which then, by the fact that it follows a Wishart distribution, has positive probability. The symmetry of $\mathbf{S}_{k+1}$ is immediate from its construction. Its positive definiteness is warranted if $\mathbf{D}_{\sigma,k+1} - \lambda_{k+1}\mathbf{D}_{\sigma,k+1}^{-1} \succ 0$, which happens when $\min_j\{\mathrm{diag}(\mathbf{D}_{\sigma,k+1}^2)\} > \lambda_{k+1} > 0$. As the penalty parameter λ_{k+1} is chosen in data-driven fashion, it may be thought of as following some distribution: $\lambda_{k+1} \sim f_\lambda(\cdot)$ with positive probability on $\mathbb{R}_{>0}$. Hence, $P[\min_j\{\mathrm{diag}(\mathbf{D}_{\sigma,k+1}^2)\} > \lambda_{k+1}] > 0$.

For the process' geometric drift to the center let $K[\mathbf{U}', \mathbf{U}]$ denote be the Markov kernel of the process, i.e. the density of $\mathbf{U}$ given that the previous kth observation equals $\mathbf{U}'$, such that $\int_{\mathcal{S}_{++}} K[\mathbf{U}', \mathbf{U}]\, d\mathbf{U} = 1$. Then, let $f_{\mathcal{W}}$ be the density of the Wishart distribution and bound the conditional expectation of the estimator as:

$$\begin{aligned} \int_{\mathcal{S}_{++}} \|\mathbf{U}\|_q\, K[\mathbf{U}', \mathbf{U}]\, d\mathbf{U} &\leq \int_{\mathcal{S}_{++}} \|\mathbf{U} - \mathbf{S}_{k+1}\|_q\, K[\mathbf{U}', \mathbf{U}]\, d\mathbf{U} + \int_{\mathcal{S}_{++}} \|\mathbf{S}_{k+1}\|_q\, |\mathbf{I}_{pp} + \lambda_{k+1}\mathbf{S}_{k+1}^{-2}|\, f_{\mathcal{W}}(\mathbf{S}_{k+1})\, d\mathbf{S}_{k+1} \\ &= \int_{\mathcal{S}_{++}} \alpha\|\mathbf{U}' - \mathbf{S}_{k+1}\|_q\, K[\mathbf{U}', \mathbf{U}]\, d\mathbf{U} + \int_{\mathcal{S}_{++}} \|\mathbf{S}_{k+1}\|_q\, |\mathbf{I}_{pp} + \lambda_{k+1}\mathbf{S}_{k+1}^{-2}|\, f_{\mathcal{W}}(\mathbf{S}_{k+1})\, d\mathbf{S}_{k+1} \\ &\leq \alpha\|\mathbf{U}'\|_q \int_{\mathcal{S}_{++}} K[\mathbf{U}', \mathbf{U}]\, d\mathbf{U} + \alpha \int_{\mathcal{S}_{++}} \|\mathbf{S}_{k+1}\|_q K[\mathbf{U}', \mathbf{U}]\, d\mathbf{U} \\ &\quad + \int_{\mathcal{S}_{++}} \|\mathbf{S}_{k+1}\|_q\, |\mathbf{I}_{pp} + \lambda_{k+1}\mathbf{S}_{k+1}^{-2}|\, f_{\mathcal{W}}(\mathbf{S}_{k+1})\, d\mathbf{S}_{k+1} \\ &\leq \alpha\|\mathbf{U}'\|_q + 2\int_{\mathcal{S}_{++}} \|\mathbf{S}_{k+1}\|_q\, |\mathbf{I}_{pp} + \lambda_{k+1}\mathbf{S}_{k+1}^{-2}|\, f_{\mathcal{W}}(\mathbf{S}_{k+1})\, d\mathbf{S}_{k+1} \end{aligned}$$

where $\alpha \in (0,1)$, Lemma 2 has been inferred, and the Jacobian determinant is derived from the reformulated estimating equation of the ridge precision estimator: $\mathbf{S}_{k+1} = \boldsymbol{\Sigma}_{k+1} - \lambda_{k+1}\boldsymbol{\Sigma}_{k+1}^{-1} + \lambda_{k+1}\boldsymbol{\Sigma}_k$. From which the tightness of the sequence now follows.

To conclude the proof, use the fact that by Theorem 8.2.14 of [32] the process converges to a stationary density. For large enough k the process may be assumed to have reached stationarity. Then, consider the estimating equation:

$\widehat{\boldsymbol{\Sigma}}_{k+1}(\lambda_{k+1}) - \mathbf{S}_{k+1} = \lambda_{k+1}[\widehat{\boldsymbol{\Sigma}}_{k+1}(\lambda_{k+1}) - \widehat{\boldsymbol{\Sigma}}_{k}(\lambda_{k})]$. Take the expectation with respect to the stationary distribution, and note that the right-hand side in the preceding display cancels. Hence, $\mathbb{E}[\widehat{\boldsymbol{\Sigma}}_{k+1}(\lambda_{k+1})] = \mathbb{E}(\mathbf{S}_{k+1}) = \boldsymbol{\Sigma}$ for large enough k. □

Proof of Theorem 3. The proof invokes Theorem 1 of [18] on the weak law of weighted averages. It is left to verify, under the specified assumptions, that the ridge and Ledoit–Wolf shrinkage covariance matrix estimator sequences satisfy the conditions of Theorem 1 of [18]. First define the pooled covariance matrix estimator $\widehat{\boldsymbol{\Sigma}}_{pool,k} = k^{-1}\sum_{k'=1}^{k}\mathbf{S}_{k'}$. This average of sample covariance matrices is an unbiased and consistent (in k) estimator of $\boldsymbol{\Sigma}$. The sequence of Ledoit–Wolf shrinkage covariance estimators $\{\widehat{\boldsymbol{\Sigma}}_{\ell w,k}(\nu_k)\}_{k=1}^{\infty}$ is itself a sequence of weighted averages of the sample covariance matrices as:

$$\widehat{\boldsymbol{\Sigma}}_{\ell w,k}(\nu_k) = \sum_{\kappa=1}^{k}\Big[\prod_{\ell=\kappa+1}^{k}\nu_\ell\Big](1-\nu_\kappa)\mathbf{S}_\kappa + \Big[\prod_{\kappa=1}^{k}\nu_\kappa\Big]\mathbf{T}_{\ell w,1}.$$

By the condition on $\{\nu_k\}_{k=1}^{\infty}$ the weights of this weighted average satisfy the condition of Theorem 1 of [18], and convergence in probability follows (by Theorem 1 of [18]) from that of the pooled covariance matrix estimator.

For the ridge covariance estimator assume, without loss of generality, that the sequence $\{\widehat{\boldsymbol{\Sigma}}_{r,k+1}(\lambda_{k+1})\}_{k=1}^{\infty}$ is initiated by the stationary density. Hence, the sequence is stationary and unbiased from the start, irrespective of the choice of the penalty parameters. Now approximate the ridge covariance matrix estimator around '$\lambda_k = \infty$' by the first order negative term of a Laurent series:

$$\begin{aligned}\widehat{\boldsymbol{\Sigma}}_{r,k+1}(\lambda_{k+1}) &= (1-\lambda_{k+1}^{-1})\widehat{\boldsymbol{\Sigma}}_{r,k}(\lambda_k) + \lambda_{k+1}^{-1}\mathbf{S}_{k+1} + \mathcal{O}(\lambda_{k+1}^{-2})\\ &= (1-\lambda_{k+1}^{-1})[(1-\lambda_k^{-1})\widehat{\boldsymbol{\Sigma}}_{r,k-1}(\lambda_{k-1}) + \lambda_k^{-1}\mathbf{S}_k] + \lambda_{k+1}^{-1}\mathbf{S}_{k+1} + \mathcal{O}(\lambda_k^{-2}) + \mathcal{O}(\lambda_{k+1}^{-2})\\ &= \sum_{k''=1}^{k}\lambda_{k''}^{-1}\Big[\prod_{k'=k''+1}^{k+1}(1-\lambda_{k'}^{-1})\Big]\mathbf{S}_{k'} + \lambda_{k+1}^{-1}\mathbf{S}_{k+1} + \sum_{\kappa=1}^{k+1}\mathcal{O}(\lambda_\kappa^{-2}),\end{aligned}$$

in which we have used (or chosen such) that $\lambda_k^{-1} \gg \lambda_k^{-2}$ for all k. By the conditions on the penalty parameter, and thereby the weights in the last expression of the preceding display, and the consistency of the pooled covariance matrix estimator, application of Theorem 1 of [18] concludes the proof. □

Proof of Theorem 4.

(i) The eigenvectors of $\widehat{\boldsymbol{\Omega}}(\lambda)$ coincide with those of $\mathbf{S} - \lambda\mathbf{T}$ (cf. [42]). Matrix perturbation theory [33] then provides that, for $\delta > 0$ small enough, $\mathbf{V}_{\omega(\lambda+\delta)} \approx \mathbf{V}_{\omega(\lambda)} + \delta g(\lambda, \mathbf{S}, \mathbf{T})$ with $g()$ a smooth function that does not involve δ. Put together this warrants the continuity of the defined map from the penalty parameter to the eigenvectors of the ridge covariance and precision matrix.

(ii) Proposition 1 of [42] states that $\lim_{\lambda\downarrow 0}\widehat{\boldsymbol{\Omega}}(\lambda) = \mathbf{S}^{-1}$ (should it exist) and $\lim_{\lambda\to\infty}\widehat{\boldsymbol{\Omega}}(\lambda) = \mathbf{T}$. In combination with the continuity shown in part (i) the statement is now evident.

(iii) The existence of a rotation matrix follows directly from the fact that any orthonormal basis of $\mathbb{R}^p$ is a rotation of any other orthonormal basis of that space. Rests to show that $\lambda \mapsto \mathbf{R}_\lambda$ is continuous. This is warranted by a variant of the Davis–Kahan $\sin(\Theta)$ theorem [9] that states that the principal angles between two sets of eigenvectors from two matrices can be bounded by a constant times the difference of these matrices. This constant depends on the distance of contiguous eigenvalues of one of these matrices, but not on their difference. Part (i) of the oroposition then warrants, for any $\varepsilon > 0$, the existence of a $\delta > 0$ such that the difference between $\mathbf{R}_\lambda$ and $\mathbf{R}_{\lambda+\delta}$ is smaller than ε.

(iv) When $\mathbf{V}_s = \mathbf{V}_t$ it is immediate that $\mathbf{S} - \lambda\mathbf{T} = \mathbf{V}_s(\mathbf{D}_s - \lambda\mathbf{D}_t)\mathbf{V}_s^\top$. Thus, as the eigenvectors of $\widehat{\boldsymbol{\Omega}}(\lambda)$ coincide with those of $\mathbf{S} - \lambda\mathbf{T}$, then $\mathbf{V}_{\omega(\lambda)} = \mathbf{V}_s$, which is independent of λ. If additionally $\mathbf{D}_s = \mathbf{D}_t^{-1}$, the eigenvalues of $[\widehat{\boldsymbol{\Omega}}(\lambda)]^{-1}$ equal: $\mathbf{D}_{\omega(\lambda)}^{-1} = \lambda^{1/2}[\widetilde{\mathbf{D}} + (\mathbf{I}_{pp} + \widetilde{\mathbf{D}}^2)^{1/2}]$, where $\widetilde{\mathbf{D}} = \frac{1}{2}(\lambda^{-1/2}\mathbf{D}_s - \lambda^{1/2}\mathbf{D}_s^{-1})$. Using ready algebra applied to the diagonal elements of $\mathbf{D}_{\omega(\lambda)}^{-1}$ it can now be seen that $\mathbf{D}_{\omega(\lambda)}^{-1}$ simplifies to $\mathbf{D}_s$. □

Proof of Theorem 5. Note that: $\lambda\sum_{g=1}^{G}\alpha_g\|\boldsymbol{\Omega} - \mathbf{T}_g\|_F^2 \propto \lambda\|\boldsymbol{\Omega} - \bar{\mathbf{T}}\|_F^2$. Parts (i), (ii), (iii), (iv) and (v) are now immediate from the corresponding statements on the 'regular' ridge precision estimator given in [42] and [40]. □

CRediT authorship contribution statement

Wessel N. van Wieringen: Conceptualization, Formal analysis, Methodology, Software, Writing - original draft. **Koen A. Stam:** Data curation. **Carel F.W. Peeters:** Conceptualization, Writing - review & editing. **Mark A. van de Wiel:** Writing - review & editing.

Acknowledgments

We thank the encouraging Editor and referees for their comments that led to this improved version of our work.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jmva.2020.104621. The supplementary material contains (i) the proofs of and related material to the theoretical results (`pdf`-file) (SM I), (ii) simulation results (SM II, III, IV), and (iii) an application illustrating the methodology (SM V). Finally, `R`-scripts (`.r`-files) of the data analysis are provided.

References

[1] M. Aflakparast, M.C.M. de Gunst, W.N. van Wieringen, Reconstruction of molecular network evolution from cross-sectional omics data, Biom. J. 60 (2018) 547–563.
[2] A. Anandkumar, V.Y.'F. Tan, F. Huang, A.S. Willsky, High-dimensional structure estimation in ising models: Local separation criterion, Ann. Statist. 40 (2012) 1346–1375.
[3] J. Bento, A. Montanari, Which graphical models are difficult to learn? Adv. Neural Inf. Process. Syst. (2009) 1303–1311.
[4] J.O. Berger, Statistical Decision Theory and Bayesian Analysis, Springer Science & Business Media, 2013.
[5] P.J. Bickel, M.K. Wichura, Convergence criteria for multiparameter stochastic processes and some applications, Ann. Math. Stat. 42 (1971) 1656–1670.
[6] A.E. Bilgrau, C.F.W. Peeters, P.S. Eriksen, M. Bøgsted, W.N. Van Wieringen, Targeted fused ridge estimation of inverse covariance matrices from multiple high-dimensional data classes, J. Mach. Learn. Res. 21 (2020) 1–52.
[7] S. Chen, D.M. Witten, A. Shojaie, Selection and estimation for mixed graphical models, Biometrika 102 (2015) 47–64.
[8] P. Danaher, P. Wang, D.M. Witten, The joint graphical lasso for inverse covariance estimation across multiple classes, J. R. Stat. Soc. Ser. B Stat. Methodol. 76 (2014) 373–397.
[9] C. Davis, W.M. Kahan, The rotation of eigenvectors by a perturbation. III, SIAM J. Numer. Anal. 7 (1970) 1–46.
[10] M. Drton, M.H. Maathuis, Structure learning in graphical modeling, Annu. Rev. Stat. Appl. 4 (2017) 365–393.
[11] R. Edgar, M. Domrachev, A.E. Lash, Gene Expression Omnibus: NCBI gene expression and hybridization array data repository, Nucleic Acids Res. 30 (2002) 207–210.
[12] B. Efron, Large-scale simultaneous hypothesis testing: the choice of a null hypothesis, J. Amer. Statist. Assoc. 99 (2004) 96–104.
[13] N. Etamadi, Convergence of weighted averages of random variables revisited, Proc. Amer. Math. Soc. 134 (2006) 2739–2744.
[14] H. Gray, G.G.R. Leday, C.A. Vallejos, S. Richardson, Shrinkage estimation of large covariance matrices using multiple shrinkage targets, submitted for publication, 2018, https://arxiv.org/abs/1809.08024.
[15] J. Guo, E. Levina, G. Michailidis, J. Zhu, Joint estimation of multiple graphical models, Biometrika 98 (2011) 1–15.
[16] A. Hannart, P. Naveau, Estimating high dimensional covariance matrices: A new look at the Gaussian conjugate framework, J. Multivariate Anal. 131 (2014) 149–162.
[17] A. Horn, C.R. Johnson, Matrix Analysis, Cambridge University Press, 2009.
[18] B. Jamison, S. Orey, W. Pruitt, Convergence of weighted averages of independent random variables, Z. Wahrscheinlichkeitstheor. Verwandte Geb. 4 (1965) 40–44.
[19] M. Kolar, H. Liu, Marginal regression for multitask learning, in: Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics, vol. 22, 2012, pp. 647–655.
[20] D. Koller, N. Friedman, Probabilistic Graphical Models: Principles and Techniques, MIT Press, 2009.
[21] O. Ledoit, M. Wolf, A well-conditioned estimator for large-dimensional covariance matrices, J. Multivariate Anal. 88 (2004) 365–411.
[22] J. Lee, T. Hastie, Learning the structure of mixed graphical models, J. Comput. Graph. Statist. 24 (2013) 230–253.
[23] J.D. Lee, Y. Sun, J. Taylor, On model selection consistency of M-estimators, Electron. J. Stat. 9 (2015).
[24] H. Liu, J. Lafferty, L. Wasserman, The nonparanormal: Semiparametric estimation of high dimensional undirected graphs, J. Mach. Learn. Res. 10 (2009) 2295–2328.
[25] Y. Ni, P. Müller, Y. Zhu, Y. Ji, Heterogeneous reciprocal graphical models, Biometrics 74 (2018) 606–615.
[26] C.J. Oates, J.Q. Smith, S. Mukherjee, Estimating causal structure using conditional dag models, J. Mach. Learn. Res. 17 (2016) 1880–1903.
[27] H. Ogata, S. Goto, K. Sato, W. Fujibuchi, H. Bono, M. Kanehisa, Kegg: Kyoto encyclopedia of genes and genomes, Nucleic Acids Res. 27 (1999) 29–34.
[28] C. Peterson, F.C. Stingo, M. Vannucci, Bayesian Inference of multiple Gaussian graphical models, J. Amer. Statist. Assoc. 110 (2015) 159–174.
[29] B.S. Price, C.J. Geyer, A.J. Rothman, Ridge fusion in statistical learning, J. Comput. Graph. Statist. 24 (2015) 439–454.
[30] W. Rudin, Principles of Mathematical Analysis, third ed., McGraw-hill, New York, 1964.
[31] J. Schäfer, K. Strimmer, A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics, Stat. Appl. Genet. Mol. Biol. 4 (2005) 32.
[32] J. Stachurski, Economic Dynamics: Theory and Computation, MIT Press, Cambridge, MA, 2009.
[33] G.W. Stewart, J.G. Sun, Matrix Perturbation Theory, Academic Press, Boston, 1990.
[34] K. Strimmer, A unified approach to false discovery rate estimation, BMC Bioinformatics 9 (2008) 303.
[35] D. Szklarczyk, A. Franceschini, M. Kuhn, et al., The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored, Nucleic Acids Res. 39 (2010) D561–D568.
[36] A.W. van der Vaart, Asymptotic Statistics, Cambridge University Press, 1998.
[37] R. Vershynin, Introduction to the non-asymptotic analysis of random matrices, in: Y. Eldar, G. Kutyniok (Eds.), Chapter 5 of: Compressed Sensing, Theory and Applications, Cambridge University Press, 2012.
[38] M.J. Wainwright, M.I. Jordan, Graphical models, exponential families, and variational inference, Found. Trends Mach. Learn. (2008).
[39] J.N. Weinstein, E.A. Collisson, G.B. Mills, et al., The cancer genome atlas pan-cancer analysis project, Nature Genet. 45 (2013) 1113–1120.
[40] W.N. van Wieringen, On the mean squared error of the ridge estimator of the covariance and precision matrix, Stat. Probab. Lett. 123 (2017) 88–92.
[41] W.N. van Wieringen, The generalized ridge estimator of the inverse covariance matrix, J. Comput. Graph. Statist. 28 (2019) 932–942.
[42] W.N. va. Wieringen, C.F.W. Peeters, Ridge estimation of the inverse covariance matrix from high-dimensional data, Comput. Statist. Data Anal. 103 (2016) 284–303.
[43] W.N. van Wieringen, C.F.W. Peeters, R.X. d. Menezes, M.A. van de Wiel, Testing for pathway (in) activation by using Gaussian graphical models, J. R. Stat. Soc. Ser. C. Appl. Stat. 67 (2018) 1419–1436.
[44] M. Yajima, D. Telesca, Y. Ji, P. Müller, Detecting differential patterns of interaction in molecular pathways, Biostatistics 16 (2015) 240–251.