

Supplementary Material to:
Updating of the Gaussian graphical model through targeted
penalized estimation

Wessel N. van Wieringen^{1,2,*}, Koen A. Stam³, Carel F.W. Peeters¹, Mark A. van de Wiel^{1,4}

¹ Department of Epidemiology and Biostatistics, Amsterdam Public Health research institute,
Amsterdam UMC, location VUmc, Amsterdam, The Netherlands

² Department of Mathematics, VU University Amsterdam, Amsterdam, The Netherlands

³ Department of Parasitology, Leiden University Medical Center, Leiden, The Netherlands

⁴ MRC Biostatistics Unit, University of Cambridge, Cambridge, United Kingdom

*Corresponding author. Email: w.vanwieringen@amsterdamumc.nl

SM I: Proofs of and related results

SM Ia: Proposition 1

Remark:

In this Supplementary Material Proposition 1 of the main text is renamed to Proposition S2, due to the preliminary results (lemma's, propositions and corollaries) that are presented here and are omitted in the main text.

Lemma S1.

Assume $\mathbf{T}_r^{-1} \neq \mathbf{S} \neq \mathbf{T}_{\ell w}$. Then:

- i) $\|\widehat{\boldsymbol{\Sigma}}_{\ell w}(\nu) - \mathbf{S}\|_F^2$ is strictly increasing in ν , and
- ii) $\|\widehat{\boldsymbol{\Sigma}}_r(\lambda) - \mathbf{S}\|_F^2$ is strictly increasing in λ .

Proof. For both lemma entries the proof proceeds by showing that the derivatives of the squared norm with respect to λ and ν are strictly positive.

- i) Substitute the expression for the estimator in the squared Frobenius norm to arrive at:

$$\|\widehat{\boldsymbol{\Sigma}}_{\ell w}(\nu) - \mathbf{S}\|_F^2 = \|(1 - \nu)\mathbf{S} + \nu\mathbf{T}_{\ell w} - \mathbf{S}\|_F^2 = \nu^2\|\mathbf{S} - \mathbf{T}_{\ell w}\|_F^2.$$

Clearly, its derivative with respect to ν is strictly positive (provided $\mathbf{S} \neq \mathbf{T}_{\ell w}$).

- ii) 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:

$$\frac{d}{d\lambda}\|\widehat{\boldsymbol{\Sigma}}_r(\lambda) - \mathbf{S}\|_F^2 = 2\text{tr}\left\{\left[\widehat{\boldsymbol{\Sigma}}_r(\lambda) - \mathbf{S}\right]\frac{d}{d\lambda}\widehat{\boldsymbol{\Sigma}}_r(\lambda)\right\}, \quad (1)$$

where:

$$\begin{aligned} \frac{d}{d\lambda}\widehat{\boldsymbol{\Sigma}}(\lambda) &= -\frac{1}{2}\mathbf{T}_r + \frac{1}{2}[\lambda\mathbf{I}_{pp} + \frac{1}{4}(\mathbf{S} - \lambda\mathbf{T}_r)^2]^{-1/2}[\mathbf{I}_{pp} + \frac{1}{2}(\lambda\mathbf{T}_r - \mathbf{S})\mathbf{T}_r] \\ &= \frac{1}{2}[\lambda\mathbf{I}_{pp} + \frac{1}{4}(\mathbf{S} - \lambda\mathbf{T}_r)^2]^{-1/2}\widehat{\boldsymbol{\Sigma}}_r(\lambda)[\widehat{\boldsymbol{\Sigma}}_r^{-1}(\lambda) - \mathbf{T}_r]. \end{aligned}$$

Substitute the latter into Equation (1) 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:

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

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 $\lambda \rightarrow \infty$. \square

An analogous lemma could be formulated stating that $\|\widehat{\boldsymbol{\Sigma}}_{\ell w}(\nu) - \mathbf{T}_{\ell w}\|_F^2$ and $\|\widehat{\boldsymbol{\Sigma}}_r(\lambda) - \mathbf{T}_r^{-1}\|_F^2$ are strictly decreasing in ν and λ , respectively.

Corollary S1.

For $\nu \in [0, 1]$ and $\lambda \in [0, \infty)$:

- i) $\|\widehat{\boldsymbol{\Sigma}}_{\ell w}(\nu) - \mathbf{S}\|_F^2 < \|\mathbf{T}_{\ell w} - \mathbf{S}\|_F^2$.
- ii) $\|\widehat{\boldsymbol{\Sigma}}_r(\lambda) - \mathbf{S}\|_F^2 < \|\mathbf{T}_r^{-1} - \mathbf{S}\|_F^2$.

Proof. This is immediate from Lemma S1 when taking the limits of λ to 0 and ∞ on the left- and right-hand side, respectively. Similarly, for the second part, take the limit of ν to 0 and 1, respectively. \square

Corollary S1 thus states that the fit of the shrinkage covariance estimators will always improve – in terms of the Frobenius loss – in comparison to its target. In the context of updating the target is the current estimate of the covariance matrix and Corollary S1 implies the next corollary.

Corollary S2.

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 $\{\widehat{\boldsymbol{\Sigma}}_{\ell w, k}(\nu_k)\}_{k=1}^\infty$ and $\{\widehat{\boldsymbol{\Sigma}}_{r, k}(\lambda_k)\}_{k=1}^\infty$ be the sequences of corresponding updated Ledoit-Wolf and ridge covariance estimators, respectively. Then, given current covariance estimates $\widehat{\boldsymbol{\Sigma}}_{\ell w, k}(\nu_k)$ and $\widehat{\boldsymbol{\Sigma}}_{r, k}(\lambda_k)$, the updated ones improve the fit:

- i) $\|\widehat{\boldsymbol{\Sigma}}_{\ell w, k+1}(\nu_{k+1}) - \mathbf{S}_{k+1}\|_F^2 < \|\widehat{\boldsymbol{\Sigma}}_{\ell w, k}(\nu_k) - \mathbf{S}_{k+1}\|_F^2$,
 - ii) $\|\widehat{\boldsymbol{\Sigma}}_{r, k+1}(\lambda_{k+1}) - \mathbf{S}_{k+1}\|_F^2 < \|\widehat{\boldsymbol{\Sigma}}_{r, k}(\lambda_k) - \mathbf{S}_{k+1}\|_F^2$,
- for all $\nu_{k+1} \in [0, 1)$ and $\lambda_{k+1} \in [0, \infty)$.

That is, when new data arrives the fit of the target to the sample covariance matrix formed from the new data is – in accordance with Corollary S2 – outperformed by the updated shrinkage covariance estimators using the newly formed sample covariance matrix and the current target. As the sample covariance matrix is an unbiased of the covariance matrix, one may hope that the updated shrinkage covariance estimators will – after enough updates – become unbiased estimators of the covariance matrix, irrespective of the initial target matrix. This is first explored in the following intermezzo.

+ Intermezzo

To shape intuition the behavior of the updated shrinkage covariance estimators is studied in the unrealistic case – which will happen with negligible probability – where the sample covariance matrix calculated from each newly arrived data set is the same, i.e. $\mathbf{S}_1 = \mathbf{S}_2 = \dots = \mathbf{S}_k = \dots$. This assumption removes the between-data-set variation and greatly simplifies the study into the effect of updating on the quality of the estimators.

Corollary S3.

Assume $\mathbf{S}_k = \mathbf{S}$ for all $k \in \mathbb{N}$. Then, for $k \in \mathbb{N}$, $\nu_k \in [0, 1)$ and $\lambda_k \in [0, \infty)$:

- i) $\|\widehat{\boldsymbol{\Sigma}}_{\ell w, k+1}(\nu_{k+1}) - \mathbf{S}\|_F^2 < \|\widehat{\boldsymbol{\Sigma}}_{\ell w, k}(\nu_k) - \mathbf{S}\|_F^2$,
- ii) $\|\widehat{\boldsymbol{\Sigma}}_{r, k+1}(\lambda_{k+1}) - \mathbf{S}\|_F^2 < \|\widehat{\boldsymbol{\Sigma}}_{r, k}(\lambda_k) - \mathbf{S}\|_F^2$.

Proof. In Corollary S1 note that $\mathbf{T}_{\ell w, k+1} = \widehat{\boldsymbol{\Sigma}}_k(\nu_k)$ and $\mathbf{T}_{r, k+1}^{-1} = \widehat{\boldsymbol{\Sigma}}_k(\lambda_k)$. \square

Corollary S3 says that, under the assumption of $\mathbf{S}_k = \mathbf{S}$ for all k , that the losses of subsequent updated shrinkage estimators forms a strict monotonically decreasing sequence.

The strict monotonicity of the loss sequences suggests it has a limit, confer Proposition S1.

Proposition S1.

Assume $\mathbf{S}_k = \mathbf{S}$ for all $k \in \mathbb{N}$. Then, for $k \in \mathbb{N}$, $\nu_k \in [0, 1)$ and $\lambda_k \in \mathbb{R}_{\geq 0}$:

- i) $\lim_{k \rightarrow \infty} \widehat{\boldsymbol{\Sigma}}_{\ell w, k}(\nu_k) = \mathbf{S}$,
- ii) $\lim_{k \rightarrow \infty} \widehat{\boldsymbol{\Sigma}}_{r, k}(\lambda_k) = \mathbf{S}$.

Proof. Corollary S3 implies that the sequences $\{\|\widehat{\boldsymbol{\Sigma}}_{\ell w, k}(\nu_k) - \mathbf{S}\|_F^2\}_{k=1}^\infty$ and $\{\|\widehat{\boldsymbol{\Sigma}}_{r, k}(\lambda_k) - \mathbf{S}\|_F^2\}_{k=1}^\infty$ are strict monotone decreasing. As the sequences are also bounded, they must converge. Rests to find their limits:

i) Due to the convergence, we have $\widehat{\Sigma}_{\ell w, k}(\nu_k) = \widehat{\Sigma}_{\ell w, k+1}(\nu_{k+1})$ as $k \rightarrow \infty$. In this limit:

$$\begin{aligned}\widehat{\Sigma}_{\ell w, k+1}(\nu_{k+1}) &= (1 - \nu_{k+1})\mathbf{S} + \nu_{k+1}\mathbf{T}_{\ell w, k+1} = (1 - \nu_{k+1})\mathbf{S} + \nu_{k+1}\widehat{\Sigma}_{\ell w, k}(\nu_k) \\ &= (1 - \nu_{k+1})\mathbf{S} + \nu_{k+1}\widehat{\Sigma}_{\ell w, k+1}(\nu_{k+1}),\end{aligned}$$

which implies $\widehat{\Sigma}_{\ell w, k+1}(\nu_{k+1}) = \mathbf{S}$.

ii) Due to the convergence, we have $\widehat{\Sigma}_{r, k}(\lambda_k) = \widehat{\Sigma}_{r, k+1}(\lambda_{k+1})$ as $k \rightarrow \infty$. Substitute this limit in the estimating equation:

$$\begin{aligned}\widehat{\Sigma}_{r, k}(\lambda_k) - \mathbf{S} &= \lambda_k[\widehat{\Sigma}_{r, k+1}^{-1}(\lambda_{k+1}) - \mathbf{T}_{r, k+1}] = \lambda_k[\widehat{\Sigma}_{r, k+1}^{-1}(\lambda_{k+1}) - \mathbf{T}_{r, k+1}] \\ &= \lambda_k[\widehat{\Sigma}_{r, k+1}^{-1}(\lambda_{k+1}) - \widehat{\Sigma}_{r, k}^{-1}(\lambda_k)] = \lambda_k[\widehat{\Sigma}_{r, k+1}^{-1}(\lambda_{k+1}) - \widehat{\Sigma}_{r, k+1}^{-1}(\lambda_{k+1})] = \mathbf{0}_{pp},\end{aligned}$$

from which the claim follows. \square

Proposition S1 tells us that construction of the target matrix \mathbf{T} from \mathbf{S} followed by repeated re-estimation of the covariance matrix with the previous estimate used as target eventually returns the original input data \mathbf{S} . We, unlike Baron Munchhausen, cannot pull ourselves from the swamp by our own hair.

Nonetheless, as the updated shrinkage covariance estimators converge to the sample covariance matrix, they are asymptotically unbiased estimators.

Corollary S4.

Assume $\mathbf{S}_k = \mathbf{S}$ for all $k \in \mathbb{N}$. Then:

- i) $\lim_{k \rightarrow \infty} \mathbb{E}[\widehat{\Sigma}_{\ell w, k}(\nu_k)] = \Sigma$.
- ii) $\lim_{k \rightarrow \infty} \mathbb{E}[\widehat{\Sigma}_{r, k}(\lambda_k)] = \Sigma$.

Proof. In Proposition S1 take the expectation on both sides. \square

+-----+
+ End of intermezzo
+-----+

Hence, without between-data-set variation, updating produces asymptotically unbiased estimator, irrespective of the choice of the initial target \mathbf{T}_0 . We now return to the case of data set specific sample covariance matrices, letting between-data-set variation back-in and study the asymptotic behavior of the updated shrinkage covariance estimators.

First, as a prerequisite, a result on the element-wise deviation of the sample covariance matrix from the population matrix is quoted from Kolar and Liu (2012).

Lemma S2. (*Lemma 12 in the Supplementary Material of Kolar and Liu, 2012*)

Let $\mathbf{Y} \in \mathbb{R}^{n \times p}$ be a random matrix whose rows are independent and identically distributed as $\mathcal{N}(\mathbf{0}_p, \Sigma)$. Denote by \mathbf{R} the correlation matrix associated with Σ . The sample covariance matrix \mathbf{S} is defined as $\mathbf{S} = \frac{1}{n}\mathbf{Y}^\top \mathbf{Y}$. Let $\xi_{j, j'} = \max\{[1 - (\mathbf{R})_{j, j'}][(\Sigma)_{j, j}(\Sigma)_{j', j'}]^{1/2}, [1 + (\mathbf{R})_{j, j'}][(\Sigma)_{j, j}(\Sigma)_{j', j'}]^{1/2}\}$. Then, for all $t \in [0, \xi_{j, j'}/2]$:

$$P(|(\mathbf{S})_{j, j'} - (\Sigma)_{j, j'}| \geq t) \leq 4 \exp[-3nt^2/(16\xi_{j, j'}^2)].$$

With little extra work, the following corollary follows from this lemma.

Corollary S5.

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 $\mathcal{N}(\mathbf{0}_p, \Sigma)$. Denote by \mathbf{R} the correlation matrix associated with Σ . Define $\xi_{j, j'} = \max\{[1 - (\mathbf{R})_{j, j'}][(\Sigma)_{j, j}(\Sigma)_{j', j'}]^{1/2}, [1 + (\mathbf{R})_{j, j'}][(\Sigma)_{j, j}(\Sigma)_{j', j'}]^{1/2}\}$. Then, for all $t \in [0, \frac{1}{2}p^2\xi_{j, j'}^2]$:

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

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

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

Proof. Write the probability as one minus that of its complement:

$$P(\|\mathbf{S}_1 - \boldsymbol{\Sigma}\|_F^2 + \|\mathbf{S}_2 - \boldsymbol{\Sigma}\|_F^2 \geq t) = 1 - P(\|\mathbf{S}_1 - \boldsymbol{\Sigma}\|_F^2 + \|\mathbf{S}_2 - \boldsymbol{\Sigma}\|_F^2 < t). \quad (2)$$

The probability of its complement is bounded from below as follows:

$$\begin{aligned} & P(\|\mathbf{S}_k - \boldsymbol{\Sigma}\|_F^2 + \|\mathbf{S}_{k+1} - \boldsymbol{\Sigma}\|_F^2 < t) \\ & \geq P(\|\mathbf{S}_k - \boldsymbol{\Sigma}\|_F^2 < \frac{1}{2}t, \|\mathbf{S}_{k+1} - \boldsymbol{\Sigma}\|_F^2 < \frac{1}{2}t) \\ & \geq P(|(\mathbf{S}_k)_{j,j'} - (\boldsymbol{\Sigma})_{j,j'}| < 2^{-1/2}p^{-1}t^{1/2} \text{ for all } j = 1, \dots, p, j' = j, \dots, p, \\ & \quad |(\mathbf{S}_{k+1})_{j,j'} - (\boldsymbol{\Sigma})_{j,j'}| < 2^{-1/2}p^{-1}t^{1/2} \text{ for all } j = 1, \dots, p, j' = j, \dots, p), \end{aligned}$$

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. Write $\tilde{t} = 2^{-1/2}p^{-1}t^{1/2}$ and apply Fréchet's inequality to the probability of the conjunction of $2 \times \frac{1}{2}(p^2 + p)$ events in the last line of the preceding display:

$$\begin{aligned} & \stackrel{\text{(continued)}}{\geq} \max \left\{ 0, \sum_{j=1, j'=j}^p P(|(\mathbf{S}_k)_{j,j'} - (\boldsymbol{\Sigma})_{j,j'}| < \tilde{t}) \right. \\ & \quad \left. + \sum_{j=1, j'=j}^p P(|(\mathbf{S}_{k+1})_{j,j'} - (\boldsymbol{\Sigma})_{j,j'}| < \tilde{t}) - [2 \frac{1}{2}(p^2 + p) - 1] \right\} \\ & = \max \left\{ 0, 1 - (p^2 + p) + \sum_{j=1, j'=j}^p P(|(\mathbf{S}_k)_{j,j'} - (\boldsymbol{\Sigma})_{j,j'}| < \tilde{t}) \right. \\ & \quad \left. + \sum_{j=1, j'=j}^p P(|(\mathbf{S}_{k+1})_{j,j'} - (\boldsymbol{\Sigma})_{j,j'}| < \tilde{t}) \right\} \\ & = \max \left\{ 0, 1 - (p^2 + p) + \sum_{j=1, j'=j}^p [1 - P(|(\mathbf{S}_k)_{j,j'} - (\boldsymbol{\Sigma})_{j,j'}| \geq \tilde{t})] \right. \\ & \quad \left. + \sum_{j=1, j'=j}^p [1 - P(|(\mathbf{S}_{k+1})_{j,j'} - (\boldsymbol{\Sigma})_{j,j'}| \geq \tilde{t})] \right\} \\ & = \max \left\{ 0, 1 - \sum_{j=1, j'=j}^p [P(|(\mathbf{S}_k)_{j,j'} - (\boldsymbol{\Sigma})_{j,j'}| \geq \tilde{t}) + P(|(\mathbf{S}_{k+1})_{j,j'} - (\boldsymbol{\Sigma})_{j,j'}| \geq \tilde{t})] \right\} \\ & \geq \max \left\{ 0, 1 - \sum_{j=1, j'=j}^p 4 \{ \exp[-3n_k \tilde{t}^2 / (16\xi_{j,j'}^2)] + \exp[-3n_{k+1} \tilde{t}^2 / (16\xi_{j,j'}^2)] \} \right\} \end{aligned}$$

in which Lemma S2 have been used and where $\tilde{t} \in [0, \frac{1}{2}\xi_{j,j'}]$. Substitute the obtained bound in Display (2):

$$\begin{aligned} & P(\|\mathbf{S}_1 - \boldsymbol{\Sigma}\|_F^2 + \|\mathbf{S}_2 - \boldsymbol{\Sigma}\|_F^2 \geq t) \\ & = 1 - P(\|\mathbf{S}_1 - \boldsymbol{\Sigma}\|_F^2 + \|\mathbf{S}_2 - \boldsymbol{\Sigma}\|_F^2 < t) \\ & \leq 1 - \max \left\{ 0, 1 - \sum_{j=1, j'=j}^p 4 \{ \exp[-3n_k \tilde{t}^2 / (16\xi_{j,j'}^2)] + \exp[-3n_{k+1} \tilde{t}^2 / (16\xi_{j,j'}^2)] \} \right\} \\ & = \min \left\{ 1, \sum_{j=1, j'=j}^p 4 \{ \exp[-3n_k \tilde{t}^2 / (16\xi_{j,j'}^2)] + \exp[-3n_{k+1} \tilde{t}^2 / (16\xi_{j,j'}^2)] \} \right\}, \end{aligned}$$

which is as claimed. To arrive at the second inequality stated in the corollary, replace all $\xi_{j,j'}$ by ξ , note that $\exp(-c/\xi_{j,j'}) \leq \exp(-c/\xi)$ as $\xi \geq \xi_{j,j'}$, sum over j and j' , and, finally, write $p(p+1) = \exp[\log(p^2 + p)]$. \square

Proposition S2. (*Fluctuation probability 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:

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

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

Proof. From Corollary S2 it follows that for any $\lambda_{k+1} \in (0, \infty)$ there exists a $\delta(\lambda_{k+1}) > 0$ such that:

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

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 be bounded from below as:

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

with $t = \min\{\delta(\lambda_{k+1}), \frac{1}{2}p^2 \min_{j,j'} \xi_{j,j'}^2\}$ (in which we have used Corollary S5). \square

Proposition S3. (*Fluctuation probability I, Ledoit-Wolf*)

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 $\widehat{\boldsymbol{\Sigma}}_{\ell w,k}(\nu_k)$ and $\widehat{\boldsymbol{\Sigma}}_{\ell w,k+1}(\nu_{k+1})$ be the corresponding updated Ledoit-Wolf 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 $\widehat{\boldsymbol{\Sigma}}_{\ell w,k}(\nu_k)$ with $\nu_k > 0$, for every $\nu_{k+1} \in (0, 1)$, there exists a $\delta(\nu_{k+1}) > 0$ for which:

$$\begin{aligned} P(\|\widehat{\boldsymbol{\Sigma}}_{\ell w,k+1}(\nu_{k+1}) - \mathbf{S}_{k+1}\|_F^2 < \|\widehat{\boldsymbol{\Sigma}}_{\ell w,k}(\nu_k) - \mathbf{S}_k\|_F^2) \\ \geq 1 - \min\{1, 2 \exp[\log(p^2 + p) - 3n_k t^2 / (16\xi^2)] + 2 \exp[\log(p^2 + p) - 3n_{k+1} t^2 / (16\xi^2)]\}. \end{aligned}$$

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

Proof. The proof is analogous to that of the ridge covariance estimator (cf. Proposition S2). \square

An important implication of Proposition S2 (and similarly of Proposition S3) is that, when n_k and n_{k+1} are sufficiently large, $P(\|\widehat{\boldsymbol{\Sigma}}_{r,k+1}(\lambda_{k+1}) - \mathbf{S}_{k+1}\|_F^2 < \|\widehat{\boldsymbol{\Sigma}}_{r,k}(\lambda_k) - \mathbf{S}_k\|_F^2)$ will be close to one. Furthermore, the bound for this probability provided by e.g. Proposition S2 is crude (as the inequalities used in its proof are not optimal) and could be improved upon.

SM Ib: Proposition 2

Remark:

In this Supplementary Material Proposition 2 of the main text is renamed to Proposition S4, due to the preliminary results (lemma's, propositions and corollaries) that are presented here and are omitted in the main text.

Definition S1. (Schatten matrix norm)

The Schatten matrix q -norm with $q > 0$ of a $p \times p$ -dimensional, symmetric matrix \mathbf{A} , denoted $\|\mathbf{A}\|_q$, is:

$$\|\mathbf{A}\|_q = \left[\sum_{j=1}^p |d_j(\mathbf{A})|^q \right]^{1/q},$$

where $d_j(\mathbf{A})$ is the j -th eigenvalue of \mathbf{A} . In particular, $\|\mathbf{A}\|_\infty = \lim_{q \rightarrow \infty} \|\mathbf{A}\|_q = \max_j |d_j(\mathbf{A})|$, which is called the spectral norm.

Lemma S3.

Assume $\mathbf{T}_r^{-1} \neq \mathbf{S} \neq \mathbf{T}_{\ell w}$. Then, for $q \in \mathbb{N}$:

- i) $\|\widehat{\Sigma}_{\ell w}(\nu) - \mathbf{S}\|_q$ is strictly increasing in ν , and
- ii) $\|\widehat{\Sigma}_r(\lambda) - \mathbf{S}\|_q$ is strictly increasing in λ .

In particular, the monotony of the difference persists into the spectral norm ($\|\cdot\|_\infty$).

Proof. Analogous to Lemma S1, the proof proceeds by showing that the derivative of the norm with respect to λ is strictly positive.

- i) Substitute the expression for the estimator in the Schatten q -norm and obtain:

$$\|\widehat{\Sigma}_{\ell w}(\lambda) - \mathbf{S}\|_q = \|(1-\nu)\mathbf{S} + \nu\mathbf{T}_{\ell w} - \mathbf{S}\|_q = \nu^2 \|\mathbf{S} - \mathbf{T}_{\ell w}\|_q.$$

Clearly, its derivative with respect to ν is strictly positive (provided $\mathbf{S} \neq \mathbf{T}_{\ell w}$), even in the $q \rightarrow \infty$ limit.

- ii) Write the Schatten q -norm as a trace: $\|\widehat{\Sigma}_r(\lambda) - \mathbf{S}\|_q = [\text{tr}(\{[\widehat{\Sigma}_r^2(\lambda) - \mathbf{S}]^2\}^{q/2})]^{1/q}$ (where $[\cdot]^q$ is written as $\{[\cdot]^2\}^{q/2}$ to emphasize and ensure the positiveness of the quantity under study). The derivative w.r.t. λ of the right-hand side is:

$$\begin{aligned} \frac{d}{d\lambda} \|\widehat{\Sigma}_r(\lambda) - \mathbf{S}\|_q &= (\|\widehat{\Sigma}_r(\lambda) - \mathbf{S}\|_q)^{1-q} \\ &\quad \times \text{tr} \left(\{[\widehat{\Sigma}_r(\lambda) - \mathbf{S}]^2\}^{q/2-1} [\widehat{\Sigma}_r(\lambda) - \mathbf{S}] \frac{d}{d\lambda} \widehat{\Sigma}_r(\lambda) \right), \end{aligned} \quad (3)$$

where (as in Lemma S1):

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

Substitute the latter into Equation (3) and use that $\widehat{\Sigma}_r(\lambda)$ satisfies the estimating equation, i.e. $\widehat{\Sigma}_r(\lambda) - \mathbf{S} = \lambda [\widehat{\Sigma}_r^{-1}(\lambda) - \mathbf{T}_r]$, to arrive at:

$$\begin{aligned} \frac{d}{d\lambda} \|\widehat{\Sigma}_r(\lambda) - \mathbf{S}\|_q &= (\|\widehat{\Sigma}_r(\lambda) - \mathbf{S}\|_q)^{1-q} \\ &\quad \times \lambda^{-1} \text{tr} \left([\lambda \mathbf{I}_{pp} + \frac{1}{4} (\mathbf{S} - \lambda \mathbf{T}_r)^2]^{-1/2} \widehat{\Sigma}_r(\lambda) \{[\widehat{\Sigma}_r(\lambda) - \mathbf{S}]^2\}^{q/2} \right). \end{aligned}$$

This trace is positive, whenever each term in the trace is positive definite. Only the last term, $[\widehat{\Sigma}_r(\lambda) - \mathbf{S}]^2$, may be semi-positive definite, but only if $\widehat{\Sigma}_r(\lambda) = \mathbf{S}$. 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 $\{\|\widehat{\Sigma}_r(\lambda) - \mathbf{S}\|_q\}_{q=0}^\infty$ and that of its derivative with respect to λ . Both sequences are bounded as $\|\widehat{\Sigma}_r(\lambda) - \mathbf{S}\|_q \leq \|\widehat{\Sigma}_r(\lambda)\|_q$ by Corollary 4.1.3 of Horn and Johnson (1990). For the sequence of derivatives, we have (by the nonnegativity of the trace of a product of two semi-positive definite matrices):

$$\begin{aligned} \frac{d}{d\lambda} \|\widehat{\Sigma}_r(\lambda) - \mathbf{S}\|_q &\leq \lambda^{-1} d_{\max} (\|\widehat{\Sigma}_r(\lambda) - \mathbf{S}\|_q)^{1-q} \text{tr}(\{[\widehat{\Sigma}_r(\lambda) - \mathbf{S}]^2\}^{q/2}) \\ &= \lambda^{-1} d_{\max} \|\widehat{\Sigma}_r(\lambda) - \mathbf{S}\|_q \leq \lambda^{-1} d_{\max} \|\widehat{\Sigma}_r(\lambda)\|_q, \end{aligned}$$

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 c.q. 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 Rudin (1964) they converge uniformly on any compact interval of λ . We may now invoke Theorem 7.17 of Rudin (1964) to conclude:

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

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. \square

Towards a different bound, use to following corollary, which requires the definition of the sub-Gaussian norm:

Definition S2. A random vector \mathbf{Y} in \mathbb{R} is sub-Gaussian if the one-dimensional marginals $\langle \mathbf{Y}, \mathbf{y} \rangle$ are sub-Gaussian random variables for all $\mathbf{y} \in \mathbb{R}^p$. The sub-Gaussian norm of \mathbf{Y} is defined as:

$$\|\mathbf{Y}\|_{\psi_2} = \sup_{\{\mathbf{y} \in \mathbb{R}^p : \|\mathbf{y}\|=1\}} |\langle \mathbf{Y}, \mathbf{y} \rangle|.$$

Corollary S6. (Corollary 5.50, Vershynin, 2012)

Consider a sub-Gaussian distribution in \mathbb{R}^p with covariance matrix $\boldsymbol{\Sigma}$, and let $\varepsilon \in (0, 1)$, $t \geq 1$. Then, with probability at least $1 - 2 \exp(-t^2 p)$ one has:

$$\|\mathbf{S} - \boldsymbol{\Sigma}\|_\infty \leq \varepsilon \quad \text{if } n \geq C(t/\varepsilon)^2 p.$$

Here $C = C_k$ depends only on the sub-Gaussian norm $K = \|\mathbf{Y}\|_{\psi_2}$ of a random vector taken from this distribution.

We are now ready to prove the main result:

Proposition S4. (Fluctuation probability 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 $\mathbf{Y} \sim \mathcal{N}(\mathbf{0}_p, \boldsymbol{\Sigma})$.

Proof. From Lemma S3 it follows that for every $\lambda_{k+1} \in (0, \infty)$ there exists a $\delta(\lambda_{k+1}) > 0$ such that:

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

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, Vershynin (2012), be bounded:

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

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

Corollary S7. (*Fluctuation probability II, Ledoit-Wolf*) 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 $\widehat{\boldsymbol{\Sigma}}_{\ell w, k}(\nu_k)$ and $\widehat{\boldsymbol{\Sigma}}_{\ell w, k+1}(\nu_{k+1})$ be the corresponding updated ridge covariance matrix estimators. Then, given the current covariance estimate $\widehat{\boldsymbol{\Sigma}}_{\ell w, k}(\nu_k)$ with $\nu_k > 0$, for every $\nu_{k+1} \in (0, 1)$, there exists a $\delta(\nu_{k+1}) \in (0, 1)$ such that for all $t \geq 1$:

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

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

Proof. The proof is analogous to that of the ridge covariance estimator (cf., Proposition S4). \square

SM Ic: Theorem 1

Theorem S1.

The bias of $\widehat{\boldsymbol{\Sigma}}_{\ell w, k}(\nu_k)$, as defined by the updating scheme (2, of the main text), vanishes as the number of updates increases. Formally, let $\nu_k \in (0, 1)$ for all k , then: $\lim_{k \rightarrow \infty} \mathbb{E}[\widehat{\boldsymbol{\Sigma}}_{\ell w, k}(\nu_k)] = \boldsymbol{\Sigma}$.

Proof. 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:

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

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 $k \rightarrow \infty$ limit the second summand of right-hand side of the preceding display will vanish for all $\nu_k \in (0, 1)$. That is:

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

Each of these sample covariances is an unbiased estimator of $\boldsymbol{\Sigma}$. Hence, so is the weighted average: $\lim_{k \rightarrow \infty} \mathbb{E}[\widehat{\boldsymbol{\Sigma}}_{\ell w, k}(\nu_k)] = \boldsymbol{\Sigma}$. \square

SM Id: Theorem 2

Theorem S2.

The bias of $\widehat{\boldsymbol{\Sigma}}_{r, k}(\lambda_k)$, as defined by the updating scheme (1, of the main text), vanishes as the number of updates increases. Formally, let $\lambda_k \in (0, \infty)$ for all k , then: $\lim_{k \rightarrow \infty} \mathbb{E}[\widehat{\boldsymbol{\Sigma}}_{r, k}(\lambda_k)] = \boldsymbol{\Sigma}$.

Proof. 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 Stachurski (2009). 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 Stachurski (2009) 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,

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

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 $\min_j\{\text{diag}(\mathbf{D}_{\sigma,k+1}^2)\} > \lambda_{k+1} > 0$. 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, $P[\min_j\{\text{diag}(\mathbf{D}_{\sigma,k+1}^2)\} > \lambda_{k+1}] > 0$.

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 k -th 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:

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

where $\alpha \in (0, 1)$, Corollary S3 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 Stachurski (2009) 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 . \square

SM Ie: Theorem 3

Theorem S3. (Consistency)

Let $\{\nu_k\}_{k=1}^\infty$ be a sequence such that $\nu_k \in (0, 1)$ for all k and $\lim_{k \rightarrow \infty} (1 - \nu_k) / [\sum_{k''=1}^k (1 - \nu_{k'}) \prod_{k'=1}^{k''-1} \nu_{k'}] = 0$. 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 $\lim_{k \rightarrow \infty} \lambda_{k+1}^{-1} / \{\sum_{k''=1}^k \lambda_{k''}^{-1} [\prod_{k'=k''+1}^{k+1} (1 - \lambda_{k'}^{-1})]\}$. Then, the covariance matrix estimators $\widehat{\Sigma}_{\ell w, k}(\nu_k)$ and $\widehat{\Sigma}_{r, k}(\lambda_k)$ are consistent, i.e. $\widehat{\Sigma}_{\ell w, k}(\nu_k) \xrightarrow{P} \Sigma$ and $\widehat{\Sigma}_{r, k}(\lambda_k) \xrightarrow{P} \Sigma$ as $k \rightarrow \infty$.

Proof. The proof invokes Theorem 1 of Jamison, Orey, and Pruitt (1965) 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 Jamison, Orey, and Pruitt (1965). First define the pooled covariance matrix estimator $\widehat{\Sigma}_{pool, k} = k^{-1} \sum_{k'=1}^k \mathbf{S}_{k'}$. This average of sample covariance matrices is an unbiased and consistent (in k) estimator of Σ . The sequence of Ledoit-Wolf shrinkage covariance estimators $\{\widehat{\Sigma}_{\ell w, k}(\nu_k)\}_{k=1}^\infty$ is itself a sequence of weighted averages of the sample covariance matrices as:

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

By the condition on $\{\nu_k\}_{k=1}^\infty$ the weights of this weighted average satisfy the condition of Theorem 1 of Jamison, Orey, and Pruitt (1965), and convergence in probability follows (by Theorem 1 of Jamison, Orey, and Pruitt, 1965) from that of the pooled covariance matrix estimator.

For the ridge covariance estimator assume, without loss of generality, that the sequence $\{\widehat{\Sigma}_{r, k+1}(\lambda_{k+1})\}_{k=1}^\infty$ 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:

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

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, Theorem 1 of Jamison, Orey, and Pruitt (1965) warrants the consistency of the ridge covariance matrix estimator. \square

SM If: Theorem 4

Theorem S4.

Let \mathbf{V}_s , \mathbf{V}_t , $\mathbf{V}_{\omega(\lambda)}$ the matrices with eigenvectors as columns of \mathbf{S} , \mathbf{T} , and $\widehat{\Omega}(\lambda)$. 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 $\lim_{\lambda \rightarrow \infty} \mathbf{V}_{\omega(\lambda)} = \mathbf{V}_t$,
- iii) can be described by the rotation $\mathbf{V}_{\omega(\lambda)} = \mathbf{R}_\lambda \mathbf{V}_s$ with \mathbf{R}_λ a rotation matrix, and
- 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 (and thus $\mathbf{S} = \mathbf{T}^{-1}$), we have $\widehat{\Omega}(\lambda) = \mathbf{S}^{-1}$ for all λ (provided \mathbf{S}^{-1} exists).

Proof.

- i)* The eigenvectors of $\widehat{\mathbf{\Omega}}(\lambda)$ coincide with those of $\mathbf{S} - \lambda\mathbf{T}$ (cf. van Wieringen and Peeters, 2016). Matrix perturbation theory (Stewart and Sun, 1990) 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 $g(\cdot)$ 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 van Wieringen and Peeters (2016) states that $\lim_{\lambda \downarrow 0} \widehat{\mathbf{\Omega}}(\lambda) = \mathbf{S}^{-1}$ (should it exist) and $\lim_{\lambda \rightarrow \infty} \widehat{\mathbf{\Omega}}(\lambda) = \mathbf{T}$. 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 (Davis and Kahan, 1970; Yu, Wang, and Samworth, 2015) 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 Proposition then warrants, for any $\varepsilon > 0$, the existence of a $\delta > 0$ such that the difference between \mathbf{R}_λ 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 $\widehat{\mathbf{\Omega}}(\lambda)$ 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 $[\widehat{\mathbf{\Omega}}(\lambda)]^{-1}$ equal:

$$\mathbf{D}_{\omega(\lambda)}^{-1} = \lambda^{1/2}[\widetilde{\mathbf{D}} + (\mathbf{I}_{pp} + \widetilde{\mathbf{D}}^2)^{1/2}], \quad (4)$$

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 equation (4) it can now be seen that $\mathbf{D}_{\omega(\lambda)}^{-1}$ simplifies to \mathbf{D}_s . □

SM Ig: Theorem 5

An explicit expression of the ridge precision estimator with multiple targets, given λ and $\boldsymbol{\alpha}$, can straightforwardly be derived. The estimating equation of $\mathbf{\Omega}$, after following a derivation analogous to that presented in van Wieringen and Peeters (2016), is:

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

where $\bar{\mathbf{T}} = \sum_{g=1}^G \alpha_g \mathbf{T}_g$. The estimator of $\mathbf{\Omega}$ then equals the root of this equation, which is (cf. van Wieringen and Peeters, 2016):

$$\widehat{\mathbf{\Omega}}(\lambda, \boldsymbol{\alpha}) = \left\{ \frac{1}{2}(\mathbf{S} - \lambda\bar{\mathbf{T}}) + [\lambda\mathbf{I}_{pp} + \frac{1}{4}(\mathbf{S} - \lambda\bar{\mathbf{T}})^2]^{1/2} \right\}^{-1}, \quad (5)$$

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

Theorem S5.

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

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

Proof. 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 van Wieringen and Peeters (2016) and van Wieringen (2017). \square

SM II: Simulation I

Banded Ω , $p = 10$, bias

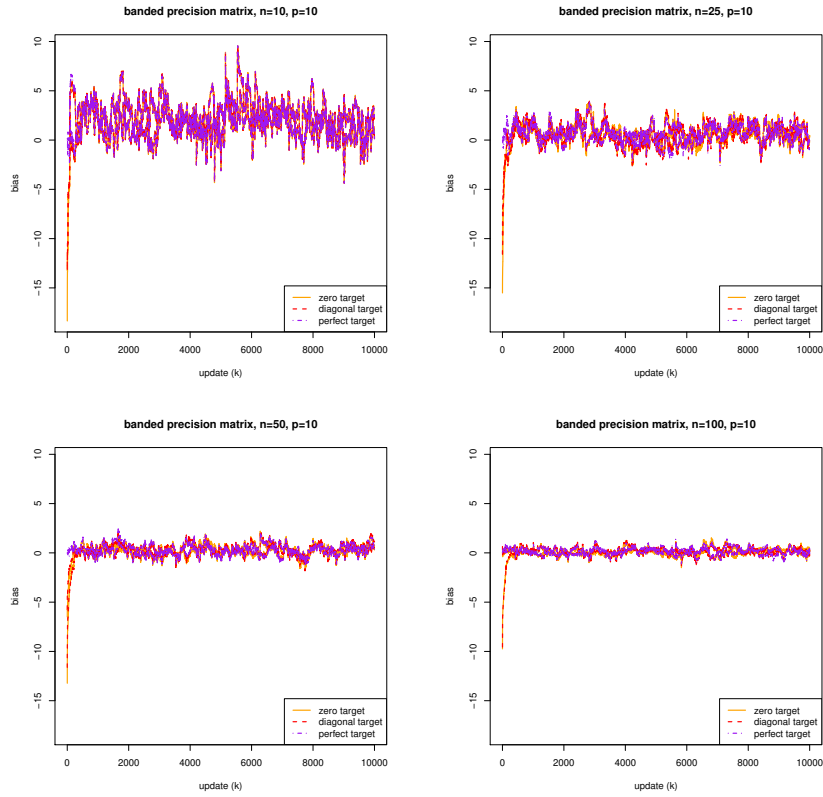


Figure 1: The bias of the updated ridge precision estimate (with various targets) vs. k (the update). The updated ridge precision estimator is initiated with *i*) a zero target $\mathbf{T}_r = \mathbf{0}_{pp}$, *ii*) a diagonal target $\mathbf{T}_r = \mathbf{I}_{pp}$, and a perfect target $\mathbf{T}_r = \Omega$. Each panel shows – per target – two sequences of the bias of the updated ridge precision estimate. Top panels: $n = 10$ and $n = 25$. Bottom panels: $n = 50$ and $n = 100$.

Banded Ω , $p = 10$, squared Frobenius loss

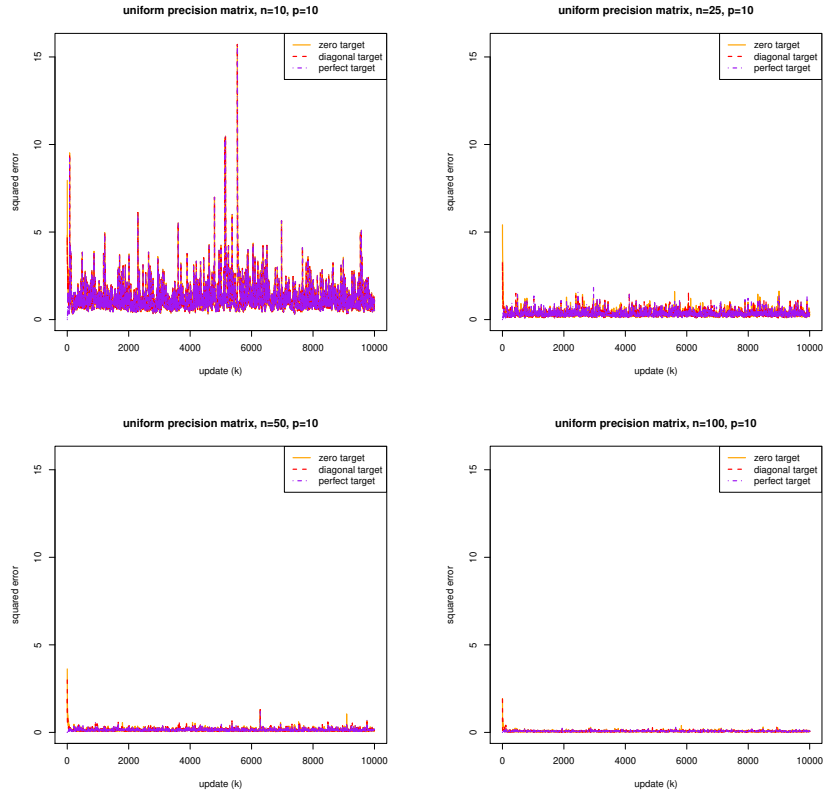


Figure 2: The squared Frobenius loss of the updated ridge precision estimate (with various targets) vs. k (the update). The updated ridge precision estimator is initiated with *i*) a zero target $\mathbf{T}_r = \mathbf{0}_{pp}$, *ii*) a diagonal target $\mathbf{T}_r = \mathbf{I}_{pp}$, and a perfect target $\mathbf{T}_r = \Omega$. Each panel shows – per target – two sequences of the bias of the updated ridge precision estimate. Top panels: $n = 10$ and $n = 25$. Bottom panels: $n = 50$ and $n = 100$.

Banded Ω , $p = 25$, bias

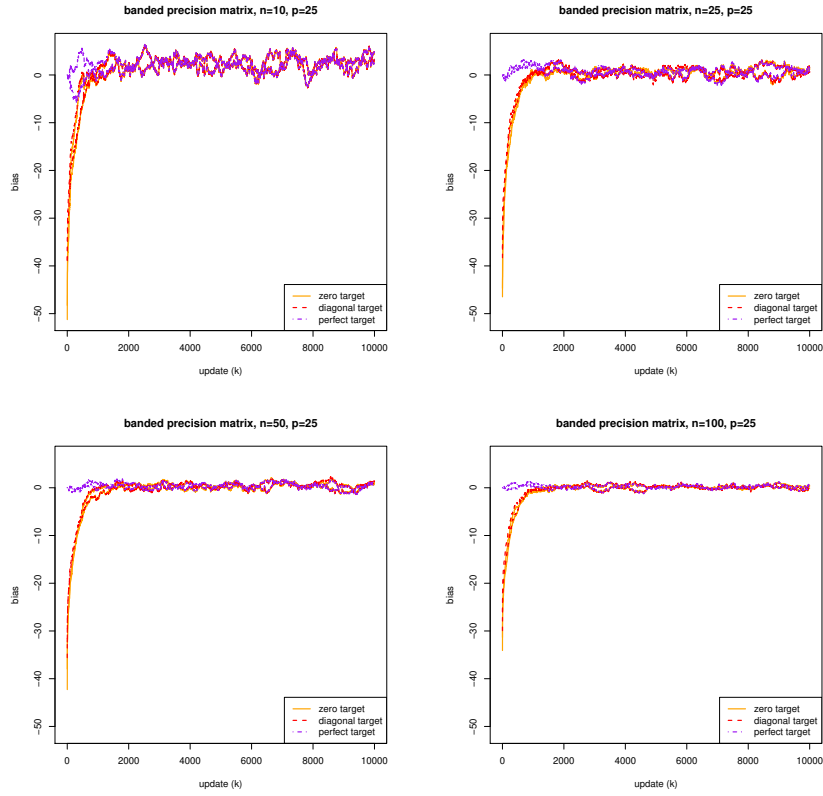


Figure 3: The bias of the updated ridge precision estimate (with various targets) vs. k (the update). The updated ridge precision estimator is initiated with *i*) a zero target $\mathbf{T}_r = \mathbf{0}_{pp}$, *ii*) a diagonal target $\mathbf{T}_r = \mathbf{I}_{pp}$, and a perfect target $\mathbf{T}_r = \Omega$. Each panel shows – per target – two sequences of the bias of the updated ridge precision estimate. Top panels: $n = 10$ and $n = 25$. Bottom panels: $n = 50$ and $n = 100$.

Banded Ω , $p = 25$, squared Frobenius loss

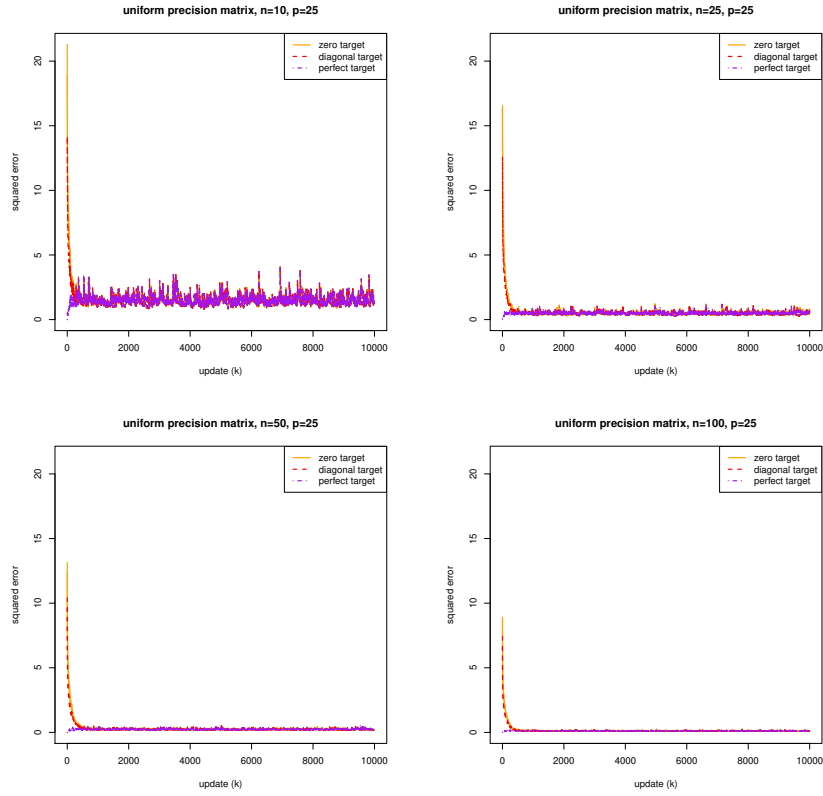


Figure 4: The squared Frobenius loss of the updated ridge precision estimate (with various targets) vs. k (the update). The updated ridge precision estimator is initiated with *i*) a zero target $\mathbf{T}_r = \mathbf{0}_{pp}$, *ii*) a diagonal target $\mathbf{T}_r = \mathbf{I}_{pp}$, and a perfect target $\mathbf{T}_r = \Omega$. Each panel shows – per target – two sequences of the bias of the updated ridge precision estimate. Top panels: $n = 10$ and $n = 25$. Bottom panels: $n = 50$ and $n = 100$.

Banded Ω , $p = 50$, bias

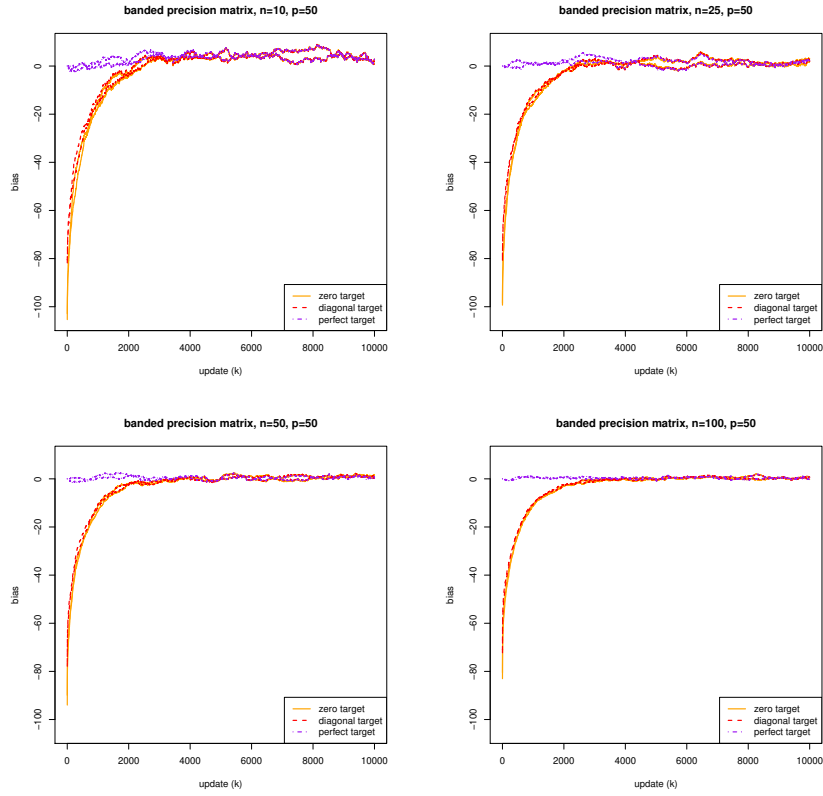


Figure 5: The bias of the updated ridge precision estimate (with various targets) vs. k (the update). The updated ridge precision estimator is initiated with *i*) a zero target $\mathbf{T}_r = \mathbf{0}_{pp}$, *ii*) a diagonal target $\mathbf{T}_r = \mathbf{I}_{pp}$, and a perfect target $\mathbf{T}_r = \Omega$. Each panel shows – per target – two sequences of the bias of the updated ridge precision estimate. Top panels: $n = 10$ and $n = 25$. Bottom panels: $n = 50$ and $n = 100$.

Banded Ω , $p = 50$, squared Frobenius loss

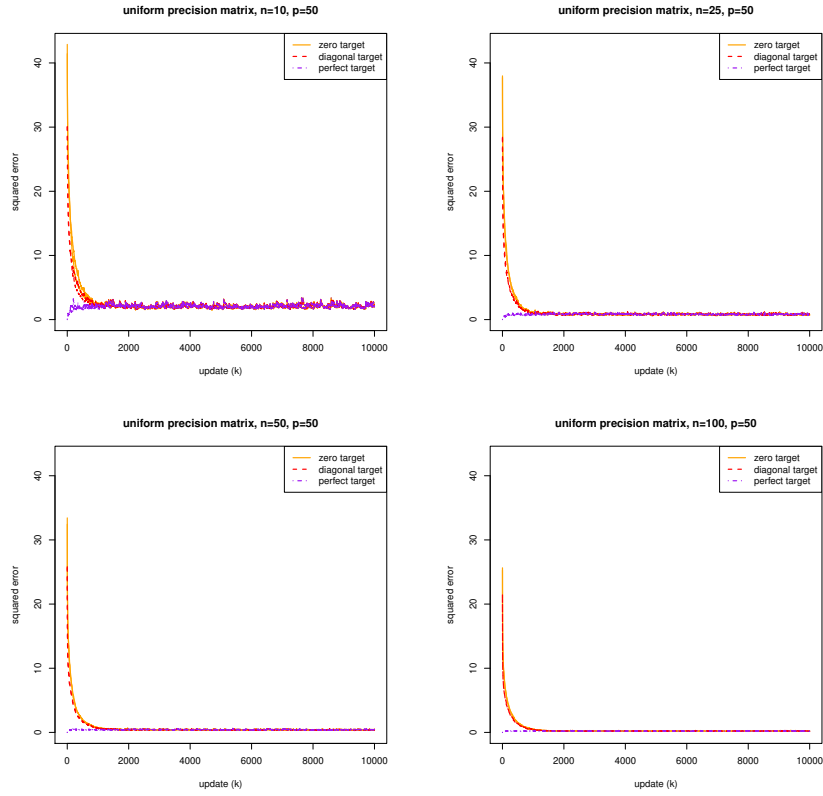


Figure 6: The squared Frobenius loss of the updated ridge precision estimate (with various targets) vs. k (the update). The updated ridge precision estimator is initiated with *i*) a zero target $\mathbf{T}_r = \mathbf{0}_{pp}$, *ii*) a diagonal target $\mathbf{T}_r = \mathbf{I}_{pp}$, and a perfect target $\mathbf{T}_r = \Omega$. Each panel shows – per target – two sequences of the bias of the updated ridge precision estimate. Top panels: $n = 10$ and $n = 25$. Bottom panels: $n = 50$ and $n = 100$.

Banded Ω , $p = 100$, bias

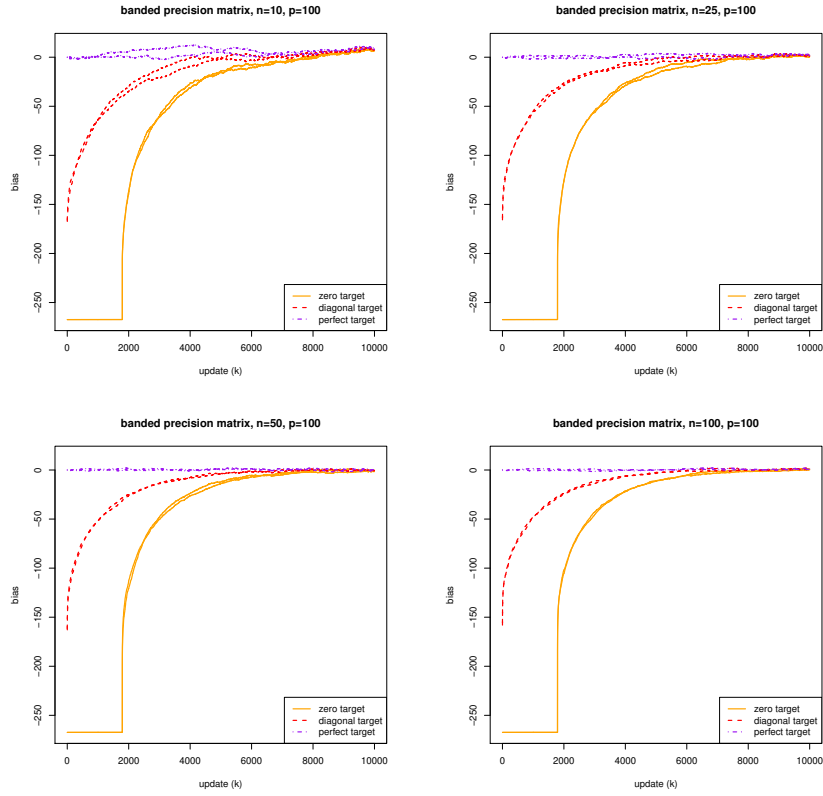


Figure 7: The bias of the updated ridge precision estimate (with various targets) vs. k (the update). The updated ridge precision estimator is initiated with *i*) a zero target $\mathbf{T}_r = \mathbf{0}_{pp}$, *ii*) a diagonal target $\mathbf{T}_r = \mathbf{I}_{pp}$, and a perfect target $\mathbf{T}_r = \Omega$. Each panel shows – per target – two sequences of the bias of the updated ridge precision estimate. Top panels: $n = 10$ and $n = 25$. Bottom panels: $n = 50$ and $n = 100$.

Banded Ω , $p = 100$, squared Frobenius loss

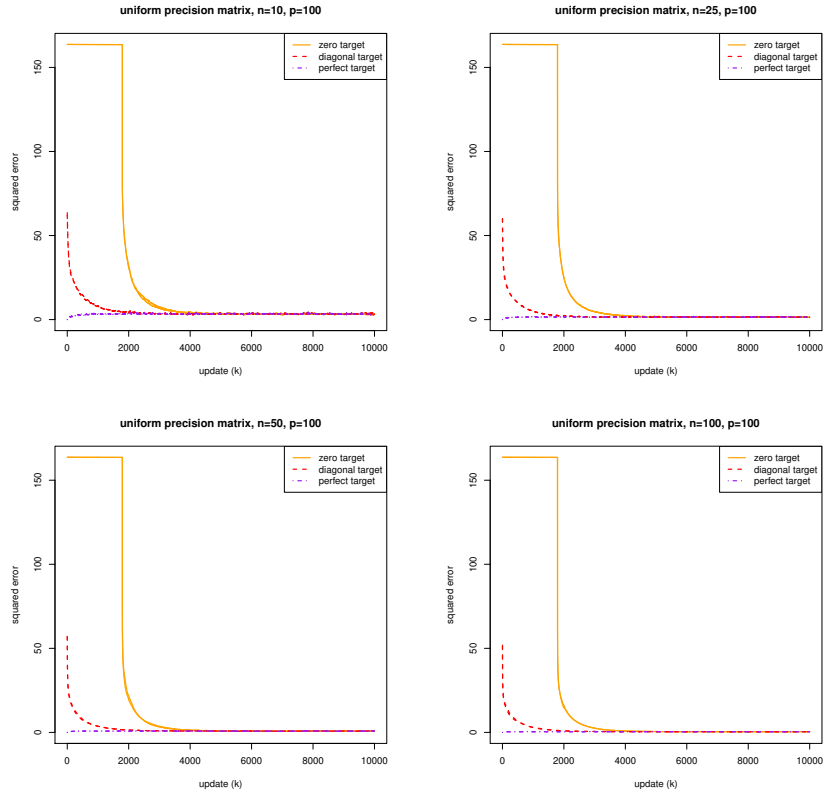


Figure 8: The squared Frobenius loss of the updated ridge precision estimate (with various targets) vs. k (the update). The updated ridge precision estimator is initiated with *i*) a zero target $\mathbf{T}_r = \mathbf{0}_{pp}$, *ii*) a diagonal target $\mathbf{T}_r = \mathbf{I}_{pp}$, and a perfect target $\mathbf{T}_r = \Omega$. Each panel shows – per target – two sequences of the bias of the updated ridge precision estimate. Top panels: $n = 10$ and $n = 25$. Bottom panels: $n = 50$ and $n = 100$.

SM III: Effect of updating

Here the various ways in which the target affects the properties of the ridge precision estimator are described. Similar conclusions can be drawn with respect to the Ledoit-Wolf shrinkage covariance matrix and its inverse and are omitted here.

Moments

The use of a non-zero target also affects the moments of the ridge precision estimator. For instance, a penalized estimate is generally biased, but the bias is influenced by the target (cf. van Wieringen and Peeters, 2016). A spot-on target may reduce the bias, whereas an off target achieves the opposite. Similarly, the variance of a penalized estimate generally vanishes when a larger penalty parameter is employed, and the target influences the speed with which this occurs. When put together the target also determines how much the mean squared error (MSE) of the ridge precision estimator improves that of its maximum likelihood counterpart (cf., van Wieringen, 2017). These claims can be deduced from a matrix series expansion of the ridge precision estimator. This is illustrated for the one dimensional case in Figure 9.

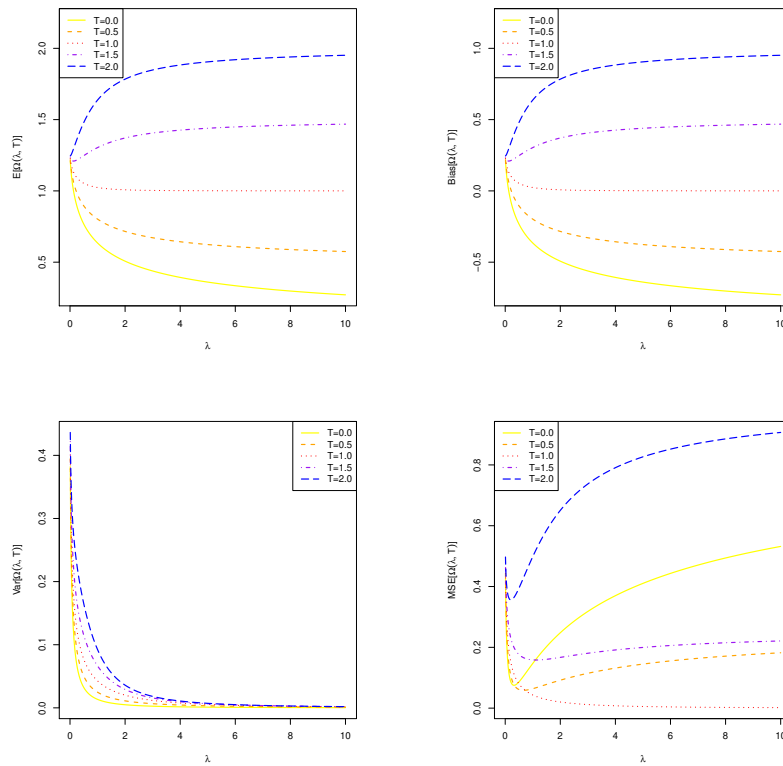


Figure 9: The expectation (top left), bias (top right), variance (bottom left) and mean squared error (bottom right) of the targeted ridge precision matrix (with various choices for the target) vs. the penalty parameter.

Loss

The loss, be it Frobenius or quadratic, may either benefit or suffer from the employed target, depending on whether it is correct or wrong. This is immediate from the penalty that it introduces

bias for any target but the correct one. In general, targets closer (in some sense) to the correct one yield lower losses than those less resembling the target (cf. the simulations reported in van Wieringen and Peeters, 2016; Bilgrau et al., 2015).

Edge selection

The target also has an effect on edge selection. van Wieringen and Peeters (2016) invoke the work of Efron (2004) and Strimmer (2008) to select 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 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 Strimmer (2008) 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 Efron, 2004). A cut-off on this probability is then to be chosen and used as the selection criterion.

To provide some intuition how this procedure is affected by the choice of the target assume – for the sake of the argument – that *a)* testing amounts to simple thresholding of the partial correlation estimate, *b)* the threshold is chosen prior to testing and independent of the choice of the target, *c)* the population precision matrix has a unit diagonal, and *d)* a ridge precision estimate $\hat{\Omega}(\lambda)$ with a unit diagonal. Effectively, the assumptions boil down to a ‘testing with known variance’ setting, and we are only bothered whether the (biased) mean estimates deviate from zero. Now note that a perfect target cannot increase the bias. In fact, then the estimate gets less biased for larger values of λ and, consequently, the partial correlation estimates get closer to their true values. This will benefit selection through thresholding as it will yield less false positives and negatives. Of course, an imperfect target may have the opposite effect on the edge selection.

SM IV: Multiple targets

Through simulation we assess whether ridge estimation with multiple targets and cross-validated parameters λ and $\boldsymbol{\alpha}$ may actually distinguish between various targets. That is, would the proposed procedure yield the largest penalty parameter λ_{α_g} for the target closest to the truth and thus shrink the estimator most towards the ‘best’ target? To this end a test scenario, comprising either two or three targets and various sample and dimension sizes, is constructed. The employed precision and target matrices employed are:

- The banded precision matrix $\boldsymbol{\Omega}$ is as in the simulation of Section 2 of the main document.
- The first target matrix \mathbf{T}_1 has unit diagonal and a single off-diagonal band. Its elements are specified through: $(\boldsymbol{\Omega})_{jj} = 1$ for $j = 1, \dots, p$, $(\boldsymbol{\Omega})_{j,j+1} = \frac{1}{5} = (\boldsymbol{\Omega}_2)_{j+1,j}$ for $j = 1, \dots, p-1$, and zero otherwise.
- The second target matrix \mathbf{T}_2 has a unit diagonal with a uniform partial correlation structure: $(\mathbf{T}_2)_{jj} = 1$ for $j = 1, \dots, p$ and $(\mathbf{T}_2)_{j_1,j_2} = \frac{1}{5}$ for $(j_1, j_2) \in \{j_1 \neq j_2 : j_1, j_2 = 1, \dots, p\}$.
- The third target matrix \mathbf{T}_3 is uninformative: $\mathbf{T}_3 = \mathbf{I}_{pp}$.

Intuitively, the first target, \mathbf{T}_1 , is best (which could be operationalized as being closest to $\boldsymbol{\Omega}$ in the Frobenius norm). For the sample size and dimension we choose $n, p \in \{10, 25, 50, 100\}$ and consider all sixteen possible combinations. For each (n, p) -choice data are sampled from the multivariate normal $\mathcal{N}(\mathbf{0}_p, \boldsymbol{\Omega}^{-1})$. From these data optimal penalty parameters, denoted λ_{opt} and $\boldsymbol{\alpha}_{\text{opt}}$, are determined through leave-one-out cross-validation. The latter is done using all three targets, but also with only two targets using either \mathbf{T}_1 and \mathbf{T}_2 or \mathbf{T}_1 and \mathbf{T}_3 . The above is repeated ten thousand times.

The determined $\boldsymbol{\alpha}_{\text{opt}}$ are summarized element-wise as histograms in the SM IV. The following can be deduced from these histograms:

- In the three-target case the $\alpha_{\text{opt},1}$ are larger than the other elements of $\boldsymbol{\alpha}_{\text{opt}}$. This indicates that \mathbf{T}_1 is most often the preferred target.
- When the sample size n is increased while the dimension p stays fixed, the LOOCV procedure yields the dominance of the $\alpha_{\text{opt},1}$ over $\alpha_{\text{opt},2}$ and $\alpha_{\text{opt},3}$ becomes more pronounced. Hence, more data helps to delineate the best target.
- When the dimension p is increased while the sample size stays fixed, the $\alpha_{\text{opt},1}$ generally dominate the other elements of $\boldsymbol{\alpha}_{\text{opt}}$ but to a lesser degree. Hence, it becomes harder to identify the best target when the problem becomes more high-dimensional.
- In the two-target case, the penalty parameter of the \mathbf{T}_1 generally dominates that of the other target, be it \mathbf{T}_2 or \mathbf{T}_3 . Moreover, the observations above with respect to the increase of sample size and dimension remain valid.

The above concentrates on the $\boldsymbol{\alpha}_{\text{opt}}$. The λ_{opt} determines the size and thereby relevance of the observed difference in the $\boldsymbol{\alpha}_{\text{opt}}$. This was investigated by pairwise scatterplots (not shown) of the $\lambda_{\text{opt}} \boldsymbol{\alpha}_{\text{opt},g}$. Although the scatterplots slightly attenuate the points made above on the basis of the histograms, the main message is unaffected.

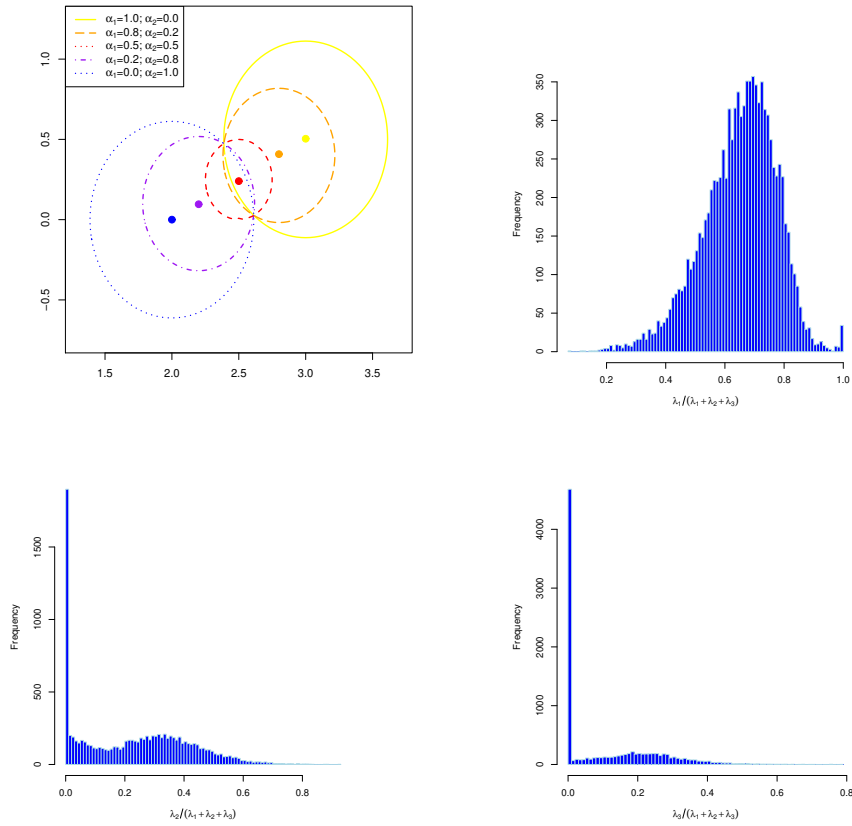


Figure 10: Topleft panel: Constraint on the entries of α induced by the mixture ridge penalty, for various choices of α . The solid dots represent the center of each constraint. Remaining panels, clockwise: Histograms of the elements of the optimal α as determined by leave-one-out cross-validation.

SM IVa: Three-target case, $p = 10$

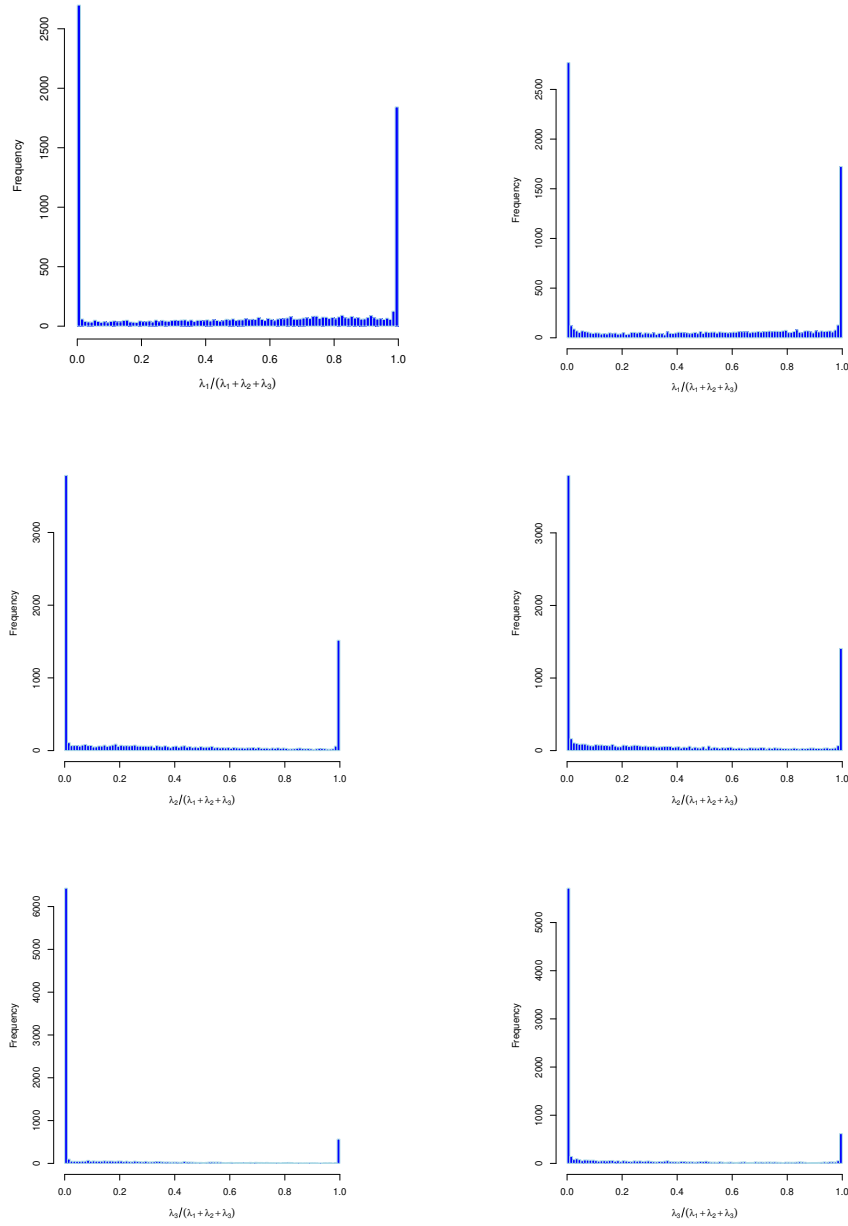


Figure 11: Three-target case with $p = 10$. Histograms of the elements of the optimal α as determined by leave-one-out cross-validation. Left and right column correspond to samples $n = 10$ and $n = 25$, respectively. The rows, from top to bottom, correspond to the first to last elements of α_{opt} , representing the weights of each target's contribution to the penalty.

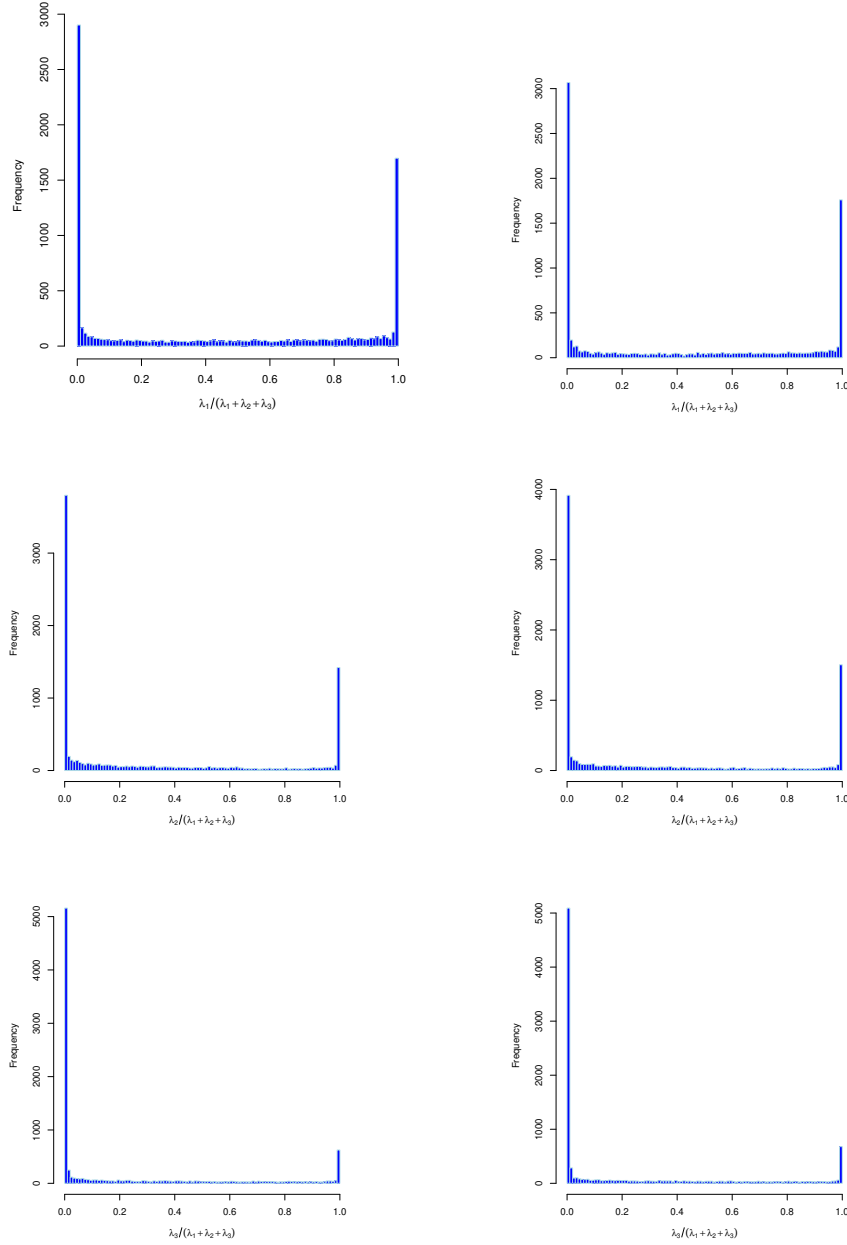


Figure 12: Three-target case with $p = 10$. Histograms of the elements of the optimal α as determined by leave-one-out cross-validation. Left and right column correspond to samples $n = 50$ and $n = 100$, respectively. The rows, from top to bottom, correspond to the first to last elements of α_{opt} , representing the weights of each target's contribution to the penalty.

SM IVb: Three-target case, $p = 25$

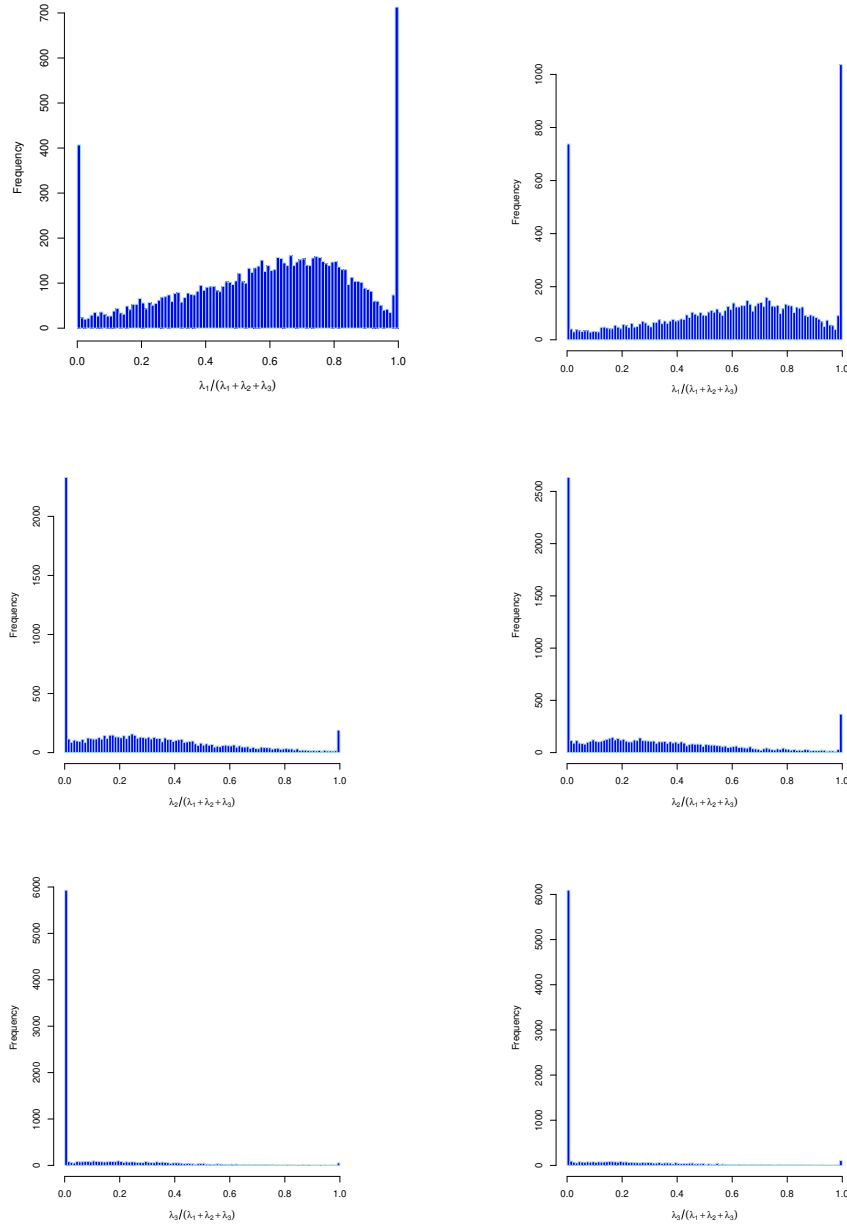


Figure 13: Three-target case with $p = 25$. Histograms of the elements of the optimal α as determined by leave-one-out cross-validation. Left and right column correspond to samples $n = 10$ and $n = 25$, respectively. The rows, from top to bottom, correspond to the first to last elements of α_{opt} , representing the weights of each target's contribution to the penalty.

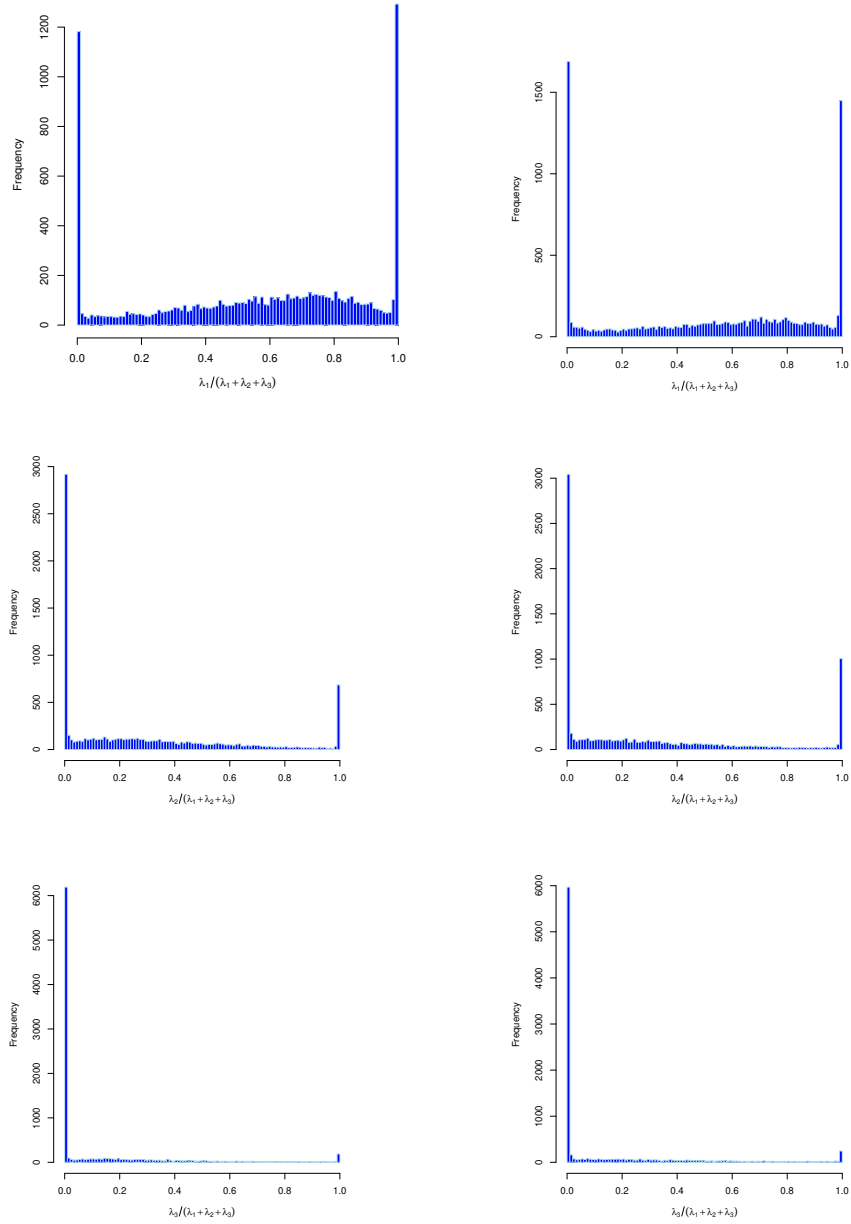


Figure 14: Three-target case with $p = 25$. Histograms of the elements of the optimal α as determined by leave-one-out cross-validation. Left and right column correspond to samples $n = 50$ and $n = 100$, respectively. The rows, from top to bottom, correspond to the first to last elements of α_{opt} , representing the weights of each target's contribution to the penalty.

SM IVc: Three-target case, $p = 50$

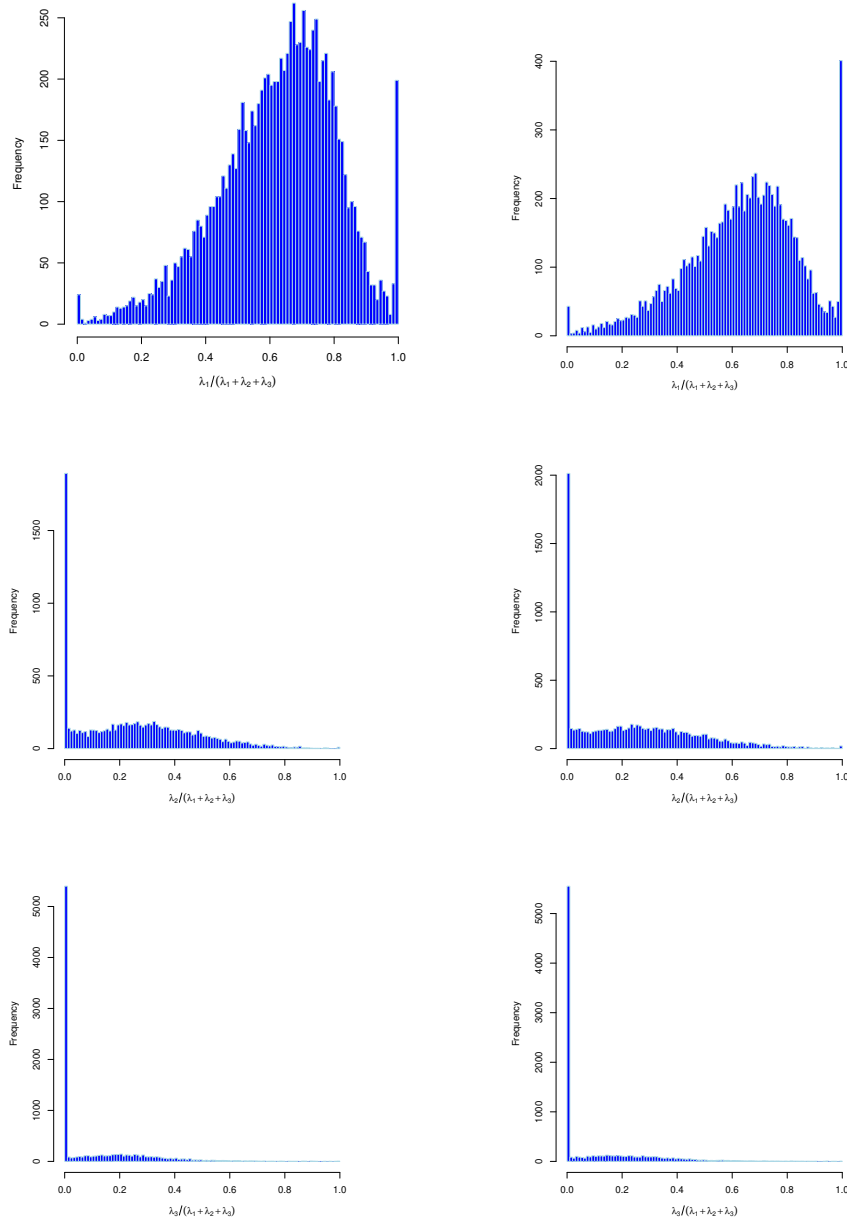


Figure 15: Three-target case with $p = 50$. Histograms of the elements of the optimal α as determined by leave-one-out cross-validation. Left and right column correspond to samples $n = 10$ and $n = 25$, respectively. The rows, from top to bottom, correspond to the first to last elements of α_{opt} , representing the weights of each target's contribution to the penalty.

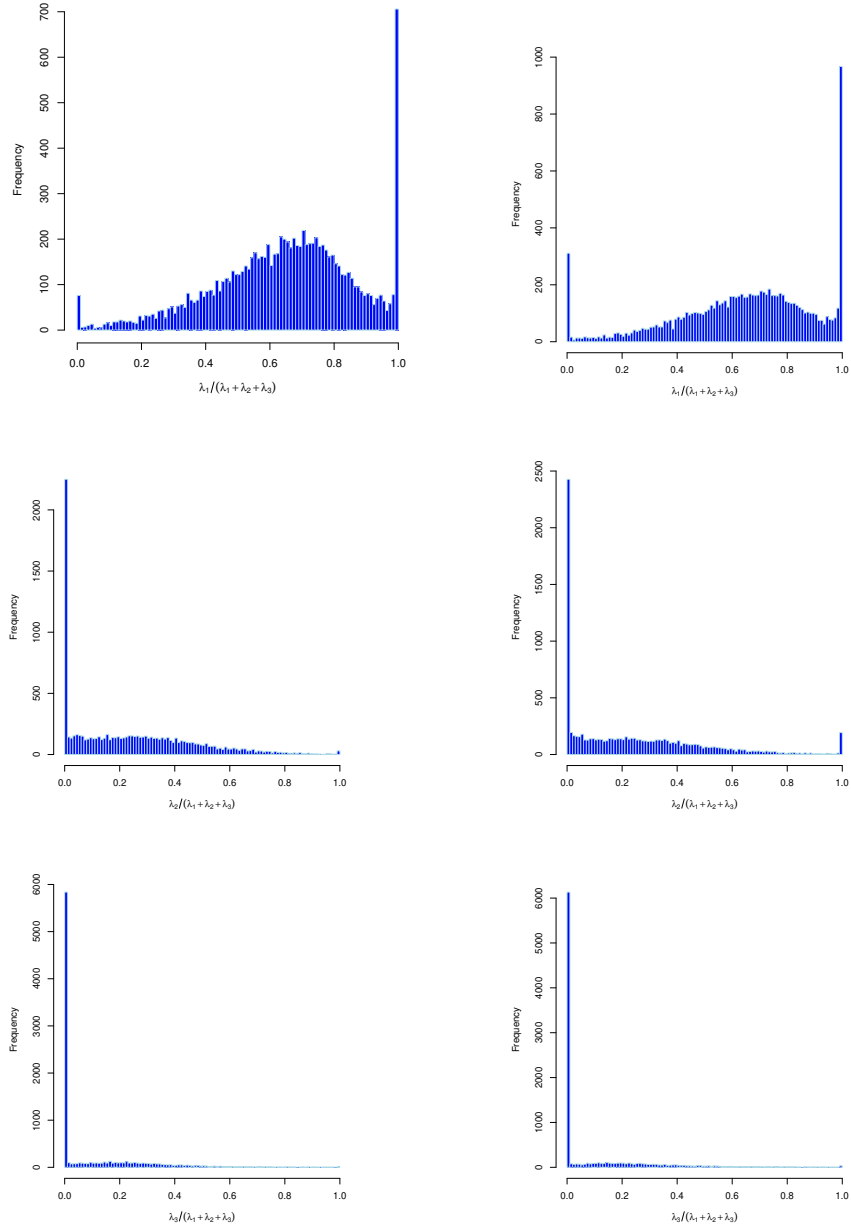


Figure 16: Three-target case with $p = 50$. Histograms of the elements of the optimal α as determined by leave-one-out cross-validation. Left and right column correspond to samples $n = 50$ and $n = 100$, respectively. The rows, from top to bottom, correspond to the first to last elements of α_{opt} , representing the weights of each target's contribution to the penalty.

SM IVd: Three-target case, $p = 100$

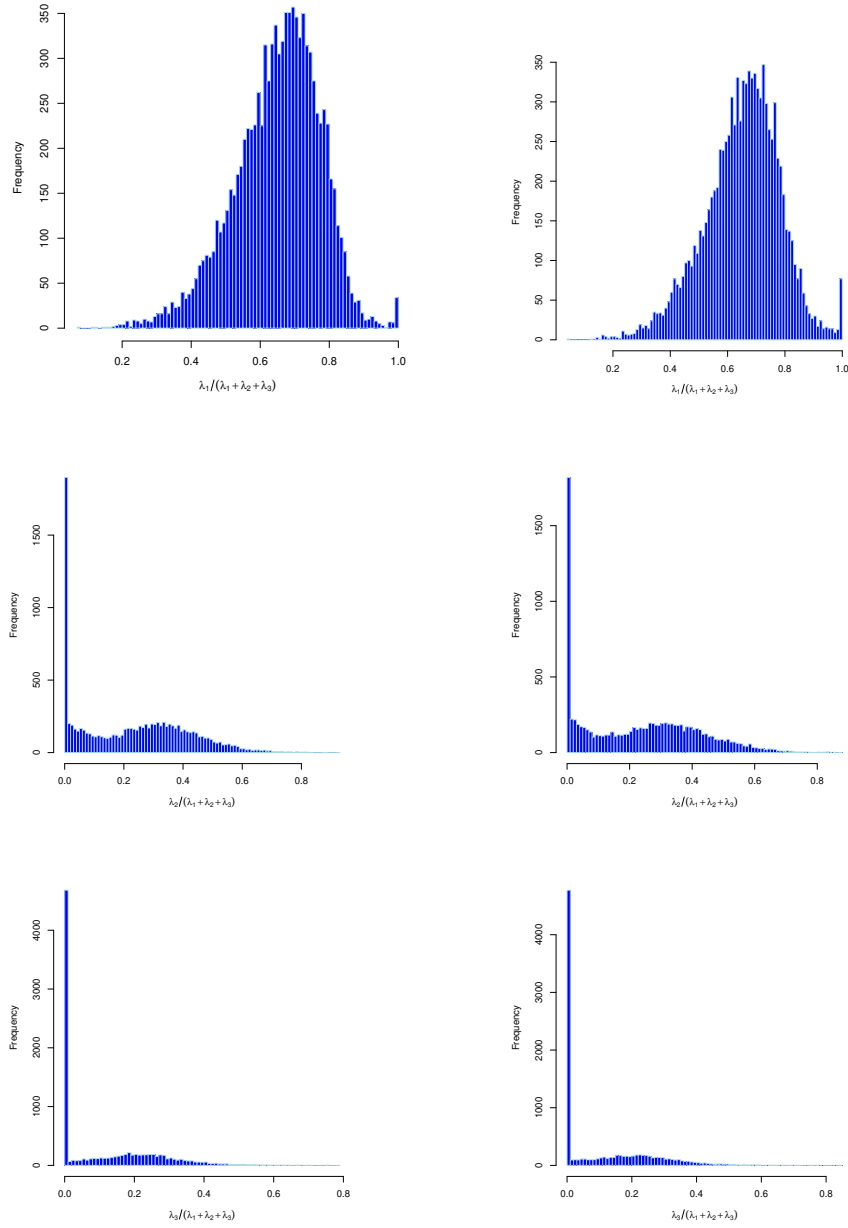


Figure 17: Three-target case with $p = 100$. Histograms of the elements of the optimal α as determined by leave-one-out cross-validation. Left and right column correspond to samples $n = 10$ and $n = 25$, respectively. The rows, from top to bottom, correspond to the first to last elements of α_{opt} , representing the weights of each target's contribution to the penalty.

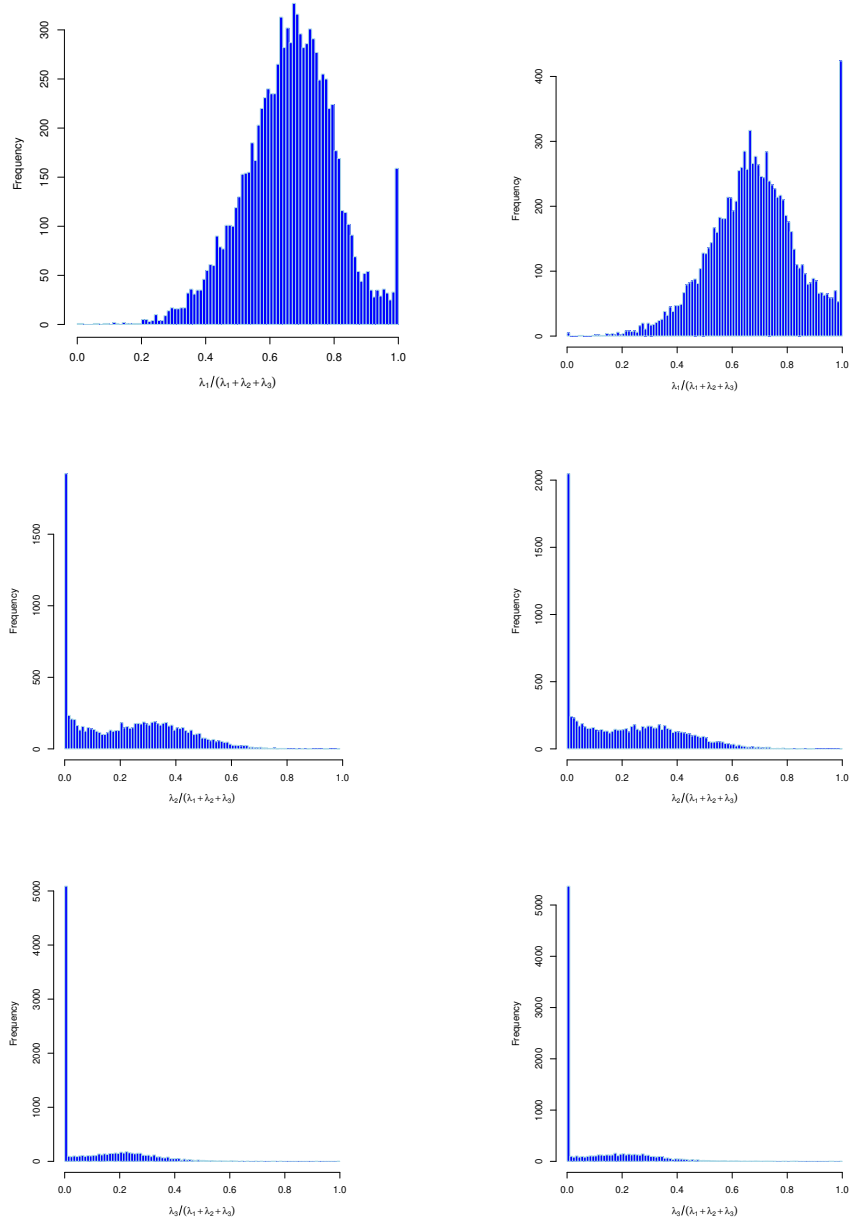


Figure 18: Three-target case with $p = 100$. Histograms of the elements of the optimal α as determined by leave-one-out cross-validation. Left and right column correspond to samples $n = 50$ and $n = 100$, respectively. The rows, from top to bottom, correspond to the first to last elements of α_{opt} , representing the weights of each target's contribution to the penalty.

Multi-target case, R- and cpp-code

The simulation studying the selection of the penalty parameters in the presence of a multi-target ridge penalty uses the following scripts:

- `twoTargetsLOOCV.r`
- `threeTargetsLOOCV.r`
- `mridgePfunctions.cpp`
- `plotHistograms.cpp`

The R-script work on a ‘copy+paste’ basis, but makes use of the C++-code provided through the script `mridgePfunctions.cpp`. The latter file is to be stored in the working directory, in order for it to be compiled by the R-script.

SM V: Application

The proposed updating of a Gaussian graphical model via iterative targeted penalized estimation is illustrated on data from five subsequently published breast cancer studies. The illustration aims to estimate the precision matrix and therefrom reconstruct the gene-gene interaction network, as operationalized by the conditional independence graph underlying the Gaussian graphical model, of ten pathways by updating. The first published data set is used for initiation, in which the effect of incorporating existing molecular biological knowledge on the network is studied.

Gene expression data of the five breast cancer studies are available through the Bioconductor repository (Gentleman et al., 2004). The studies were conducted in different hospitals but the breast cancer samples were all interrogated transcriptomically using the same Affymetrix 133a platform. The available data sets have been curated and preprocessed by Schröder et al. (2011). The samples of each study are limited to those that are estrogen positive (ER+), as breast cancer is a hormone related form of cancer and pathways may behave differently between estrogen groups (e.g., Creighton et al., 2015). The data sets, in order of appearance of the corresponding publication, (with their ER+ sample size in brackets), are called: **VDX** ($n = 209$), **UPP** ($n = 213$), **UNT** ($n = 86$), **TRANSBIG** ($n = 134$), and **MAINZ** ($n = 162$), where acronyms have been adopted from Schröder et al. (2011). The data sets are further subsetted (in their dimension p) to form data sets with data of ten pathways. The pathways, the genes that constitute them and the relationship among these genes, are taken from the KEGG repository (Ogata et al., 1999), available through the R-packages **KEGG.db** and **KEGGgraph**. The latter package provides directed graphs of the pathway, which are moralized to obtain undirected ones. Subsequently, the nomenclature of the genes was matched between pathways and the data sets. This amounted to the conversion of gene names to standardized EntrezID’s. The resulting pathway dimensions range from $p = 29$ to $p = 240$ and their number of edges from 5% to 29%. Finally, the data have been gaussianized (see Section 4 of the main document).

The topology of the ten pathways is reconstructed by updating the ridge estimates of their precision matrices from the transcriptomic data. Updating is initiated by a target precision matrix constructed from the first **VDX** data set by the graphical lasso (Friedman, Hastie, and Tibshirani, 2008), with either a zero or infinitely large penalty parameter for precision elements that correspond to edges present or absent (respectively) in the pathway topology of the repository. For reference a simple diagonal precision target, corresponding to an empty network, is taken along. The initial targets are used in the estimation of the precision matrices from the next **UPP** data set, with the penalty parameter chosen by leave-one-out cross-validation. The resulting estimate serves in turn as target precisions for the subsequent data set. This last step is repeated till the fifth **MAINZ** data set.

A first but indirect view of the updating results concentrates on the chosen penalty parameter values. They represent a weighing of the information provided by current and preceding data. Should the subsequent ridge precision estimates approach a precision matrix that is assumed common to all data sets, it is expected that the penalty parameter increases with each update. The top left panel of Figure 19 shows the leave-one-out penalty parameter values initiated with the diagonal target. The plots (hardly different between the initial targets) reveal a increasing trend of these penalty parameter values, consistently over the pathways. The varying sample sizes of the data sets seem not to affect this trend. By virtue of the weighted average interpretation (as pointed out in Section 2 of the main document), the target thus increasingly contains more relevant information for the estimation of the precision matrix from the next data set.

The ridge precision matrix estimates are used to evaluate whether updating via shrinkage is beneficial. Would each data set indeed provide information on the same model system, they may be considered as draws from the same distribution. Then, the asymptotic unbiasedness and consistency of the updated ridge precision estimate (Corollary 1 of the main document) implies the convergence of the subsequent estimate. While this is an asymptotic statement, the obtained updated estimates may be used to investigate whether there is indeed a tendency towards such a trend. Hereto the top right panel of Figure 19 shows the squared error loss of the difference between subsequent estimates, $\|\hat{\mathbf{\Omega}}_{r,k+1}(\lambda_{k+1,\text{opt}}) - \hat{\mathbf{\Omega}}_{r,k}(\lambda_{k,\text{opt}})\|_F^2$. Indeed the plot shows, consistently over the pathways, a tendency towards a diminishing difference with an increasing number of updates. In part, this is not surprising

after the increasing penalty parameters. Finally, while the chosen penalty parameters exhibit little differences between the two initial targets, the differences between subsequent precision matrix estimates are slightly higher for the updated precision initiated with the KEGG informed target (see SM IV).

The fit of the updated estimates of the precision matrices is evaluated by means of their quadratic loss in relation to the sample covariance matrix, $\|\hat{\mathbf{\Omega}}_{r,k}(\lambda_{k,\text{opt}})\mathbf{S}_k - \mathbf{I}_{pp}\|_F^2$. The left panel of the middle row of Figure 19 shows these losses. They appear to be stable, with a little ‘hick-up’ in the UNT data set. Closer inspection reveals that, from the first to the last data set, there is even a small decrease. Perhaps not spectacular, but in light of precision matrix estimates that are more alike, i.e. borrow more from the preceding data sets, it is reassuring that the fit does not deteriorate, and even improves. For comparison the latter analysis has been repeated for the ridge covariance matrix estimator from the pooled data as well as the pooled sample covariance matrix (as defined at the end of Section 2 of the main document). The result for the former (which performs better than the latter) is shown in the right panel of the middle row of Figure 19. Generally, the pooled ridge precision matrix estimator exhibits the same behavior as $\hat{\mathbf{\Omega}}_{r,k}(\lambda_{k,\text{opt}})$ in terms of fit. It does not show the aforementioned ‘hick-up’, which is smoothed out due to the pooling. However, its quadratic loss is generally worse, especially in the last two updates – except for the Citrate cycle pathway which is smallest in dimension. The above picture, i.e. a better loss for the updated ridge precision estimator, is preserved when studying the Frobenius loss (instead of the quadratic loss).

To assess the value of the qualitative (topological) information stored on the KEGG repository, the updated precision estimates are sparsified. Sparsification uses the empirical Bayes procedure proposed by Efron (2004) and is implemented for the screening of nonzero elements of a precision matrices by Strimmer (2008). Note that the dimension of the Citrate cycle pathway, $p = 29$, is too small to produce reliable inference of the precision matrices’ support with the employed procedure, and is therefore ignored in the remainder. The gene-gene interaction networks inferred from the updated precision estimates initiated with a non-informative target are much sparser than those inferred from their KEGG informed counterparts. But while the former sequence of networks gradually gains edges, the latter networks become sparser over the updates. The stability of the selected edges is assessed by the overlap between subsequently inferred networks (see SM IV). In line with the previous observation this overlap in- and decreases for the network sequences initiated with non- and KEGG informed, respectively. Moreover, the overlap between the networks inferred from the differently initiated updated precision matrices is determined per data set (see SM IV). Initially, until the UNT this overlap grows, suggesting the data prevails. But the overlap shows a decrease in the TRANSBIG data set, to recover slightly in the final MAINZ data set. More updates are needed to assess whether this will stabilize. The discussed plot also reveals that some pathways, e.g., apoptosis, exhibit hardly any overlap. These generally also have a low number of selected edges (irrespective of the chosen initialization). This is most likely due to the fact that they have been inactivated in ER+ breast cancer tissue. Finally, the initiation can be seen to have a lasting influence. The bottom panels of figure 19 show the inferred Cell cycle pathway with a non- and KEGG informed initiation (left and right panel, respectively). The edge widths in these network plots are proportional to the number of data sets in which the edge has been selected. Apart from edges present or absent in one but not the other, most notable is that the KEGG informed network has more edges present in all updates of the pathway’s network. Some of these edges have indeed been reported in the literature to be active interactions in ER+ breast cancer. Hence, the KEGG informed initiation may indeed be a good initial guess, thereby speeding up convergence to biologically sensible results.

The analysis above has been repeated for the reversed and a random order of the data sets. Similar trends in relation between e.g. the penalty parameter, the loss, the sequential difference, et cetera versus the order of the data sets are observed. These trends are most monotone for the random order, with the chronological and the reversed order showing a small ‘hick-up’ from monotonicity. This suggests that one (or a couple of the) data set(s) differs a little from the others. This difference, however, does not ruin the global trend.

Table 1: Data sets in order of appearance, the related publication, their ER+ sample size, and the R-package with its data.

| Dataset | Publication | # ER+ | R-package |
|----------|------------------------|-------|-----------------------------------|
| VDX | Wang et al. (2005) | 209 | <code>breastCancerVDX</code> |
| UPP | Miller et al. (2005) | 213 | <code>breastCancerUPP</code> |
| UNT | Sotiriou et al. (2006) | 86 | <code>breastCancerUNT</code> |
| TRANSBIG | Desmedt et al. (2007) | 134 | <code>breastCancerTRANSBIG</code> |
| MAINZ | Schmidt et al. (2008) | 162 | <code>breastCancerMAINZ</code> |

Table 2: Pathways, names and KEGG ID, number of genes and edge (absolute and in percentage).

| Pathway | KEGG ID | # genes | # edges | # edges (%) |
|---------------|----------|---------|---------|-------------|
| mapk | hsa04010 | 240 | 1770 | 6.17 |
| p53 | hsa04115 | 62 | 101 | 5.34 |
| erbb | hsa04012 | 82 | 299 | 9.00 |
| apoptosis | hsa04210 | 354 | 79 | 11.49 |
| wnt | hsa04310 | 126 | 1104 | 14.02 |
| TGFb | hsa04350 | 80 | 386 | 12.22 |
| VEGF | hsa04370 | 68 | 285 | 12.51 |
| Citrate cycle | hsa00020 | 29 | 121 | 29.06 |
| JakSTAT | hsa04630 | 138 | 1638 | 17.33 |
| Cell cycle | hsa04110 | 104 | 866 | 16.03 |

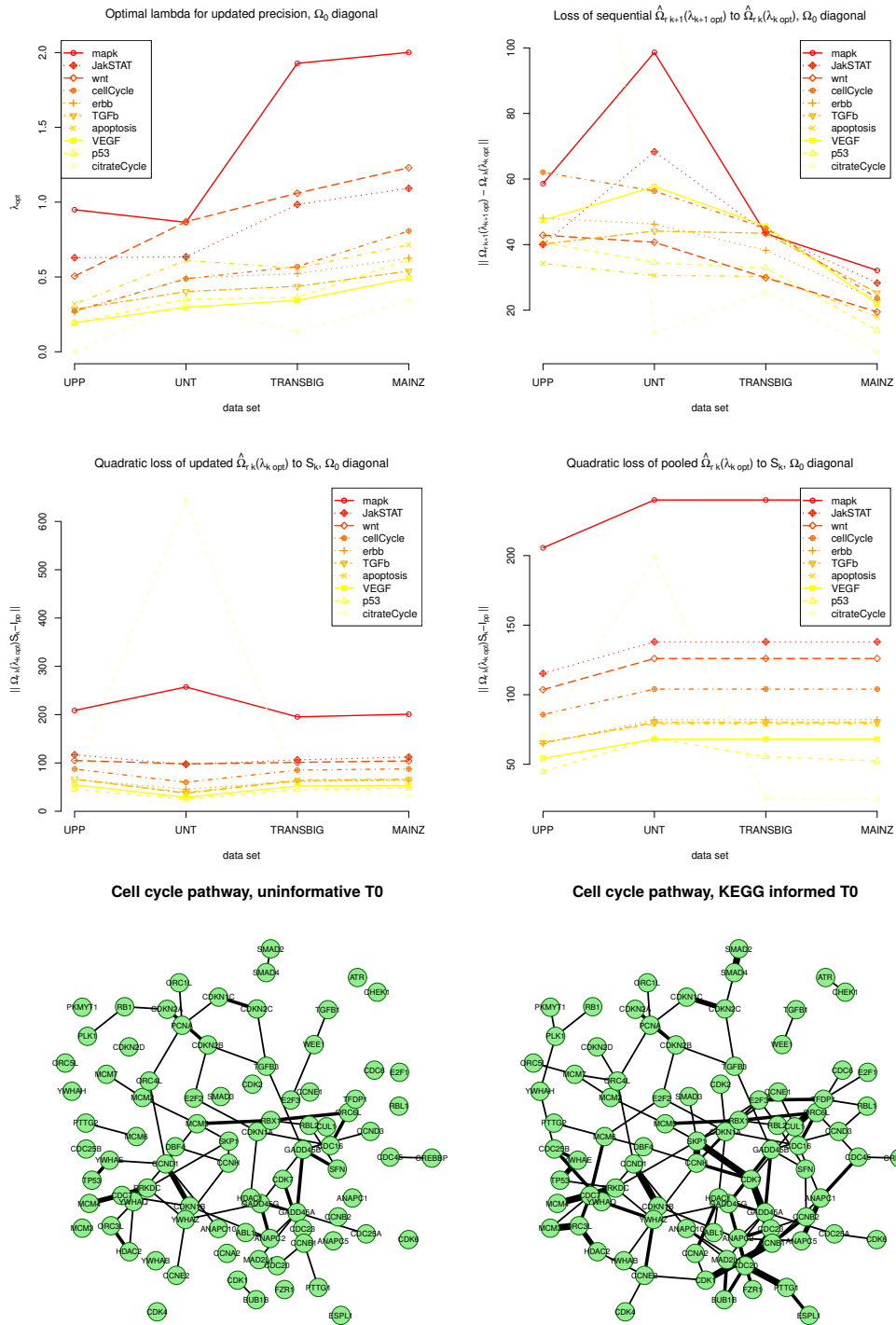


Figure 19: Top left panel: the optimal penalty parameter for each data set, connected by a line within each pathway. The coloring of lines and symbols ranges (throughout all panels) from red to light yellow, which corresponds to the pathways with the largest and smallest dimensions, respectively, and intermediate colors representing intermediate dimension sizes. Top right panel: the squared error between sequential ridge precision estimates. Middle panels: the quadratic loss of the updated ridge (left) and pooled ridge (right) precision estimate to the sample covariance matrix. Bottom panels: gene-gene interaction networks of the Cell cycle pathway inferred from the updated ridge precision estimates initiated with a diagonal target (left) and a KEGG inspired target (right). The edge width is proportional to the number of data sets in which the edge is selected.

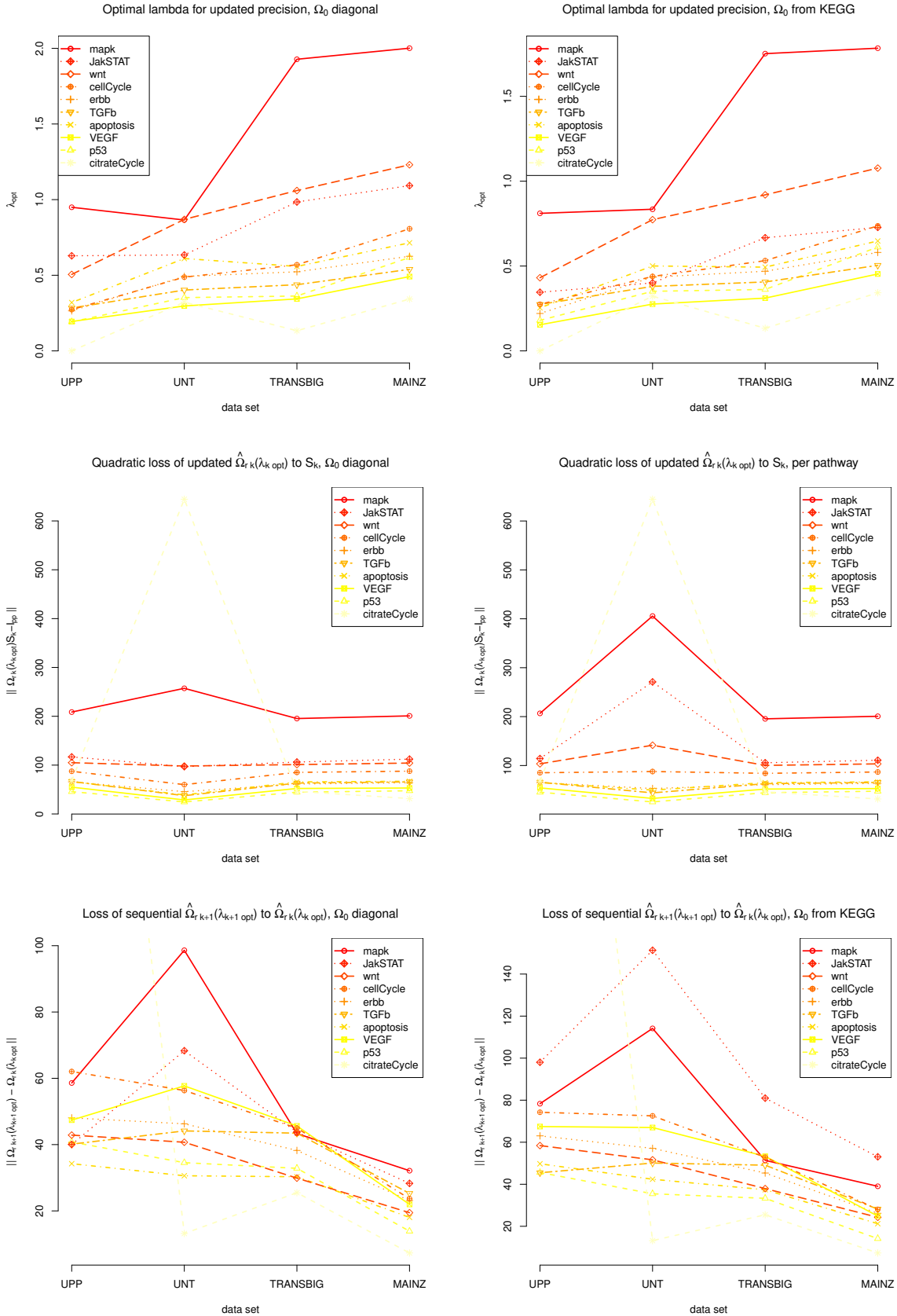


Figure 20: Analysis results with datasets in chronological order. Top panels: the optimal penalty parameter for each dataset for connected by a line within each pathway. The coloring of lines and symbols ranges (throughout all panels) from red to light yellow, which corresponds to the pathways with the largest and smallest dimensions, respectively, and intermediate colors representing intermediate dimension sizes. Middle panels: the quadratic loss of the updated ridge precision

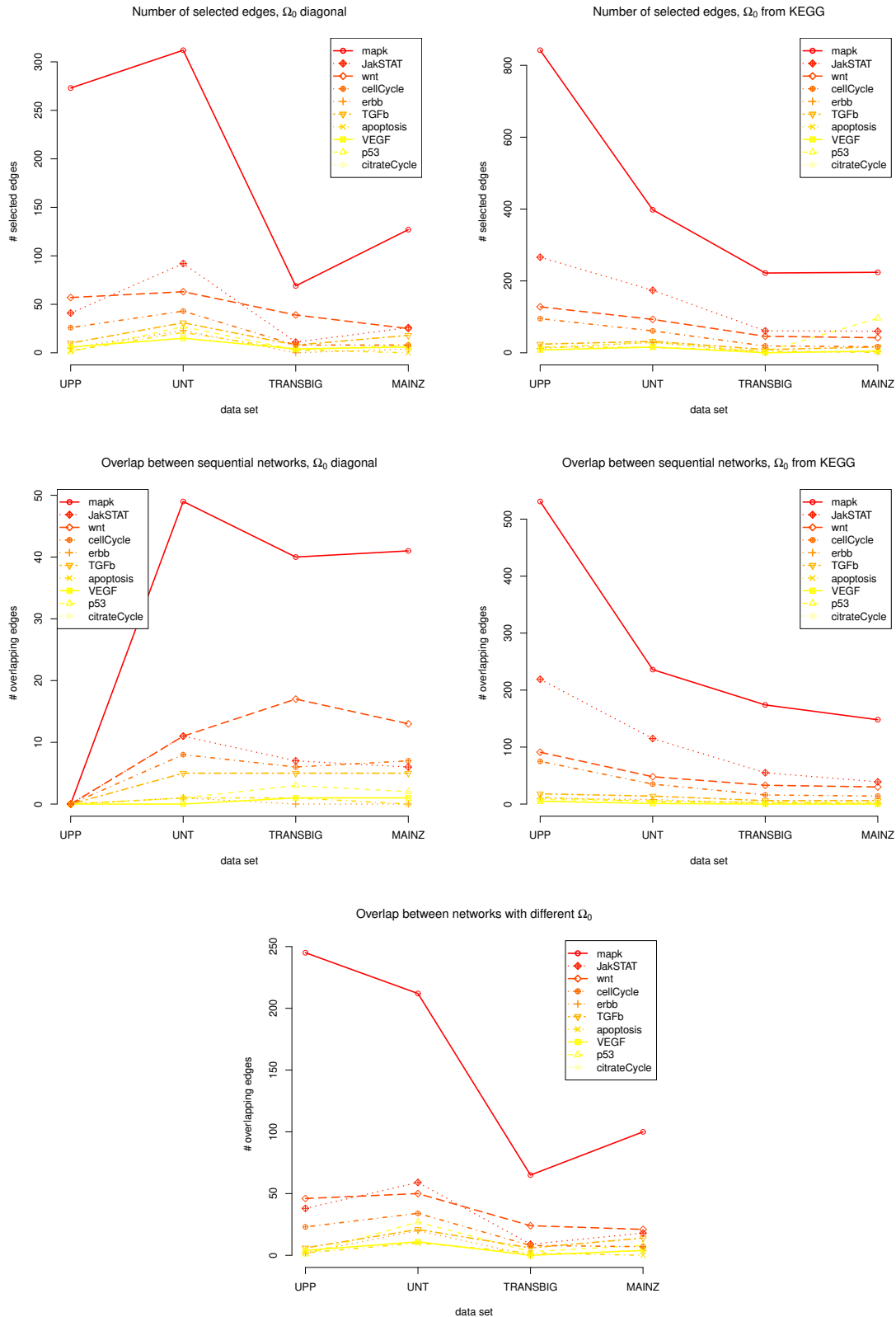


Figure 21: Analysis results with datasets in chronological order. Top panels: the number of edges in the support inferred from updated ridge precision estimates initiated with a diagonal target (left) and a KEGG inspired target (right). Middle panels: the number of overlapping edges in the support inferred from subsequent updated ridge precision estimates initiated with a diagonal target (left) and a KEGG inspired target (right). Bottom panel: the number of overlapping edges between the support from the update ridge precision initiated with a diagonal and a KEGG inspired target.

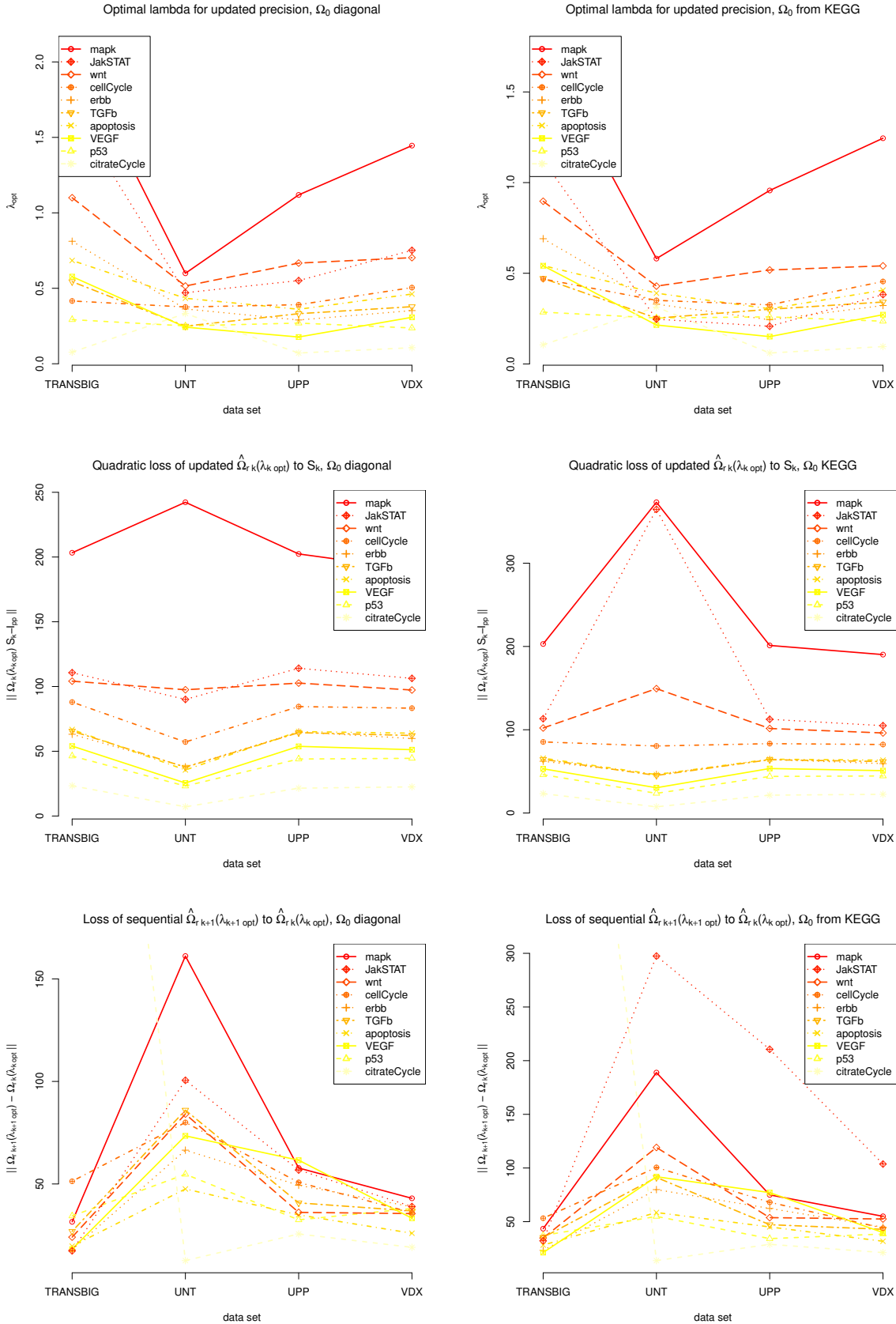


Figure 22: Analysis results with datasets in reversed order. Top panels: the optimal penalty parameter for each dataset for connected by a line within each pathway. The coloring of lines and symbols ranges (throughout all panels) from red to light yellow, which corresponds to the pathways with the largest and smallest dimensions, respectively, and intermediate colors representing intermediate dimension sizes. Middle panels: the quadratic loss of the updated ridge precision

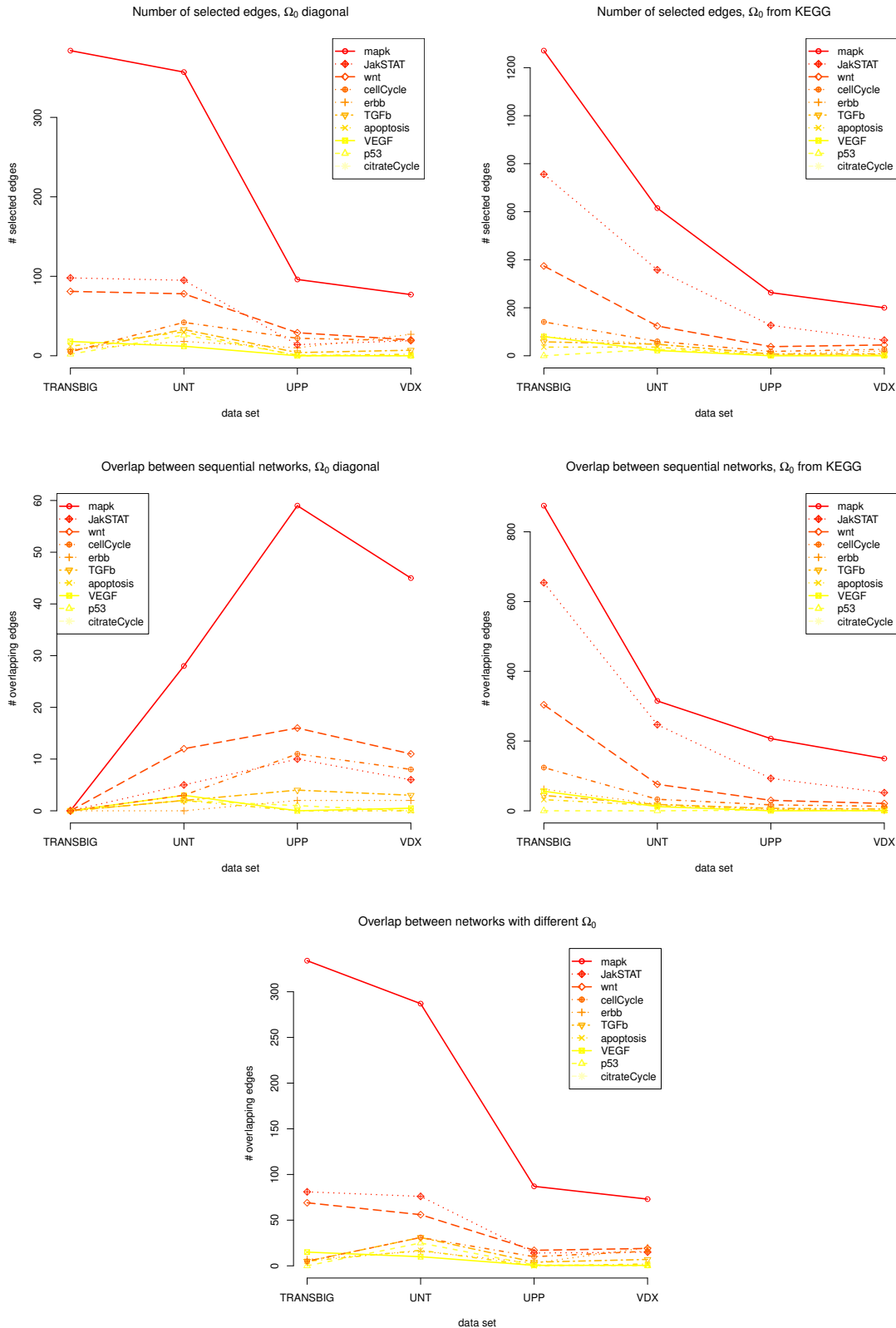


Figure 23: Analysis results with datasets in reversed order. Top panels: the number of edges in the support inferred from updated ridge precision estimates initiated with a diagonal target (left) and a KEGG inspired target (right). Middle panels: the number of overlapping edges in the support inferred from subsequent updated ridge precision estimates initiated with a diagonal target (left) and a KEGG inspired target (right). Bottom panel: the number of overlapping edges between the support from the update ridge precision initiated with a diagonal and a KEGG inspired target.

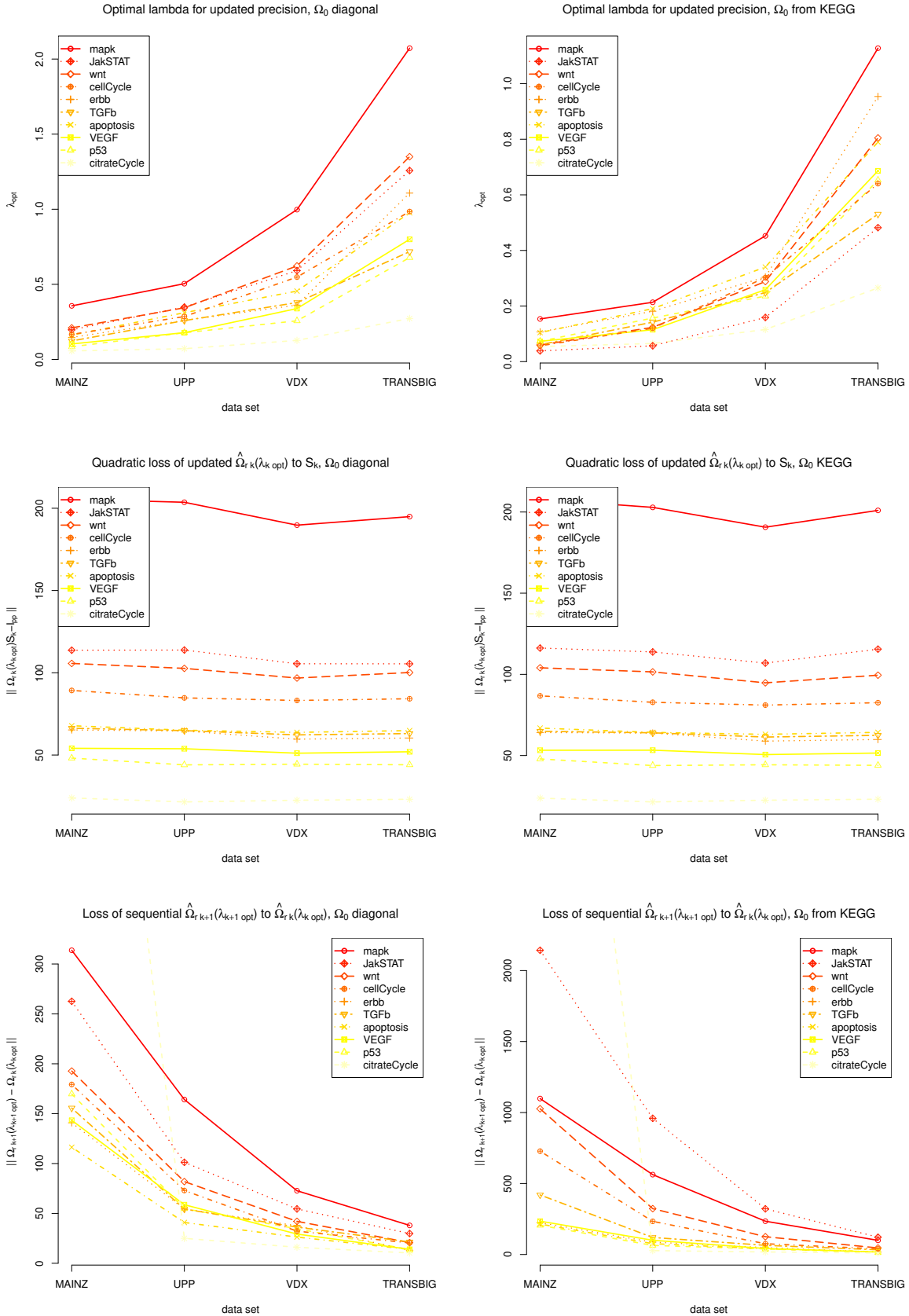


Figure 24: Analysis results with datasets in random order. Top panels: the optimal penalty parameter for each dataset for connected by a line within each pathway. The coloring of lines and symbols ranges (throughout all panels) from red to light yellow, which corresponds to the pathways with the largest and smallest dimensions, respectively, and intermediate colors representing intermediate dimension sizes. Middle panels: the quadratic loss of the updated ridge precision

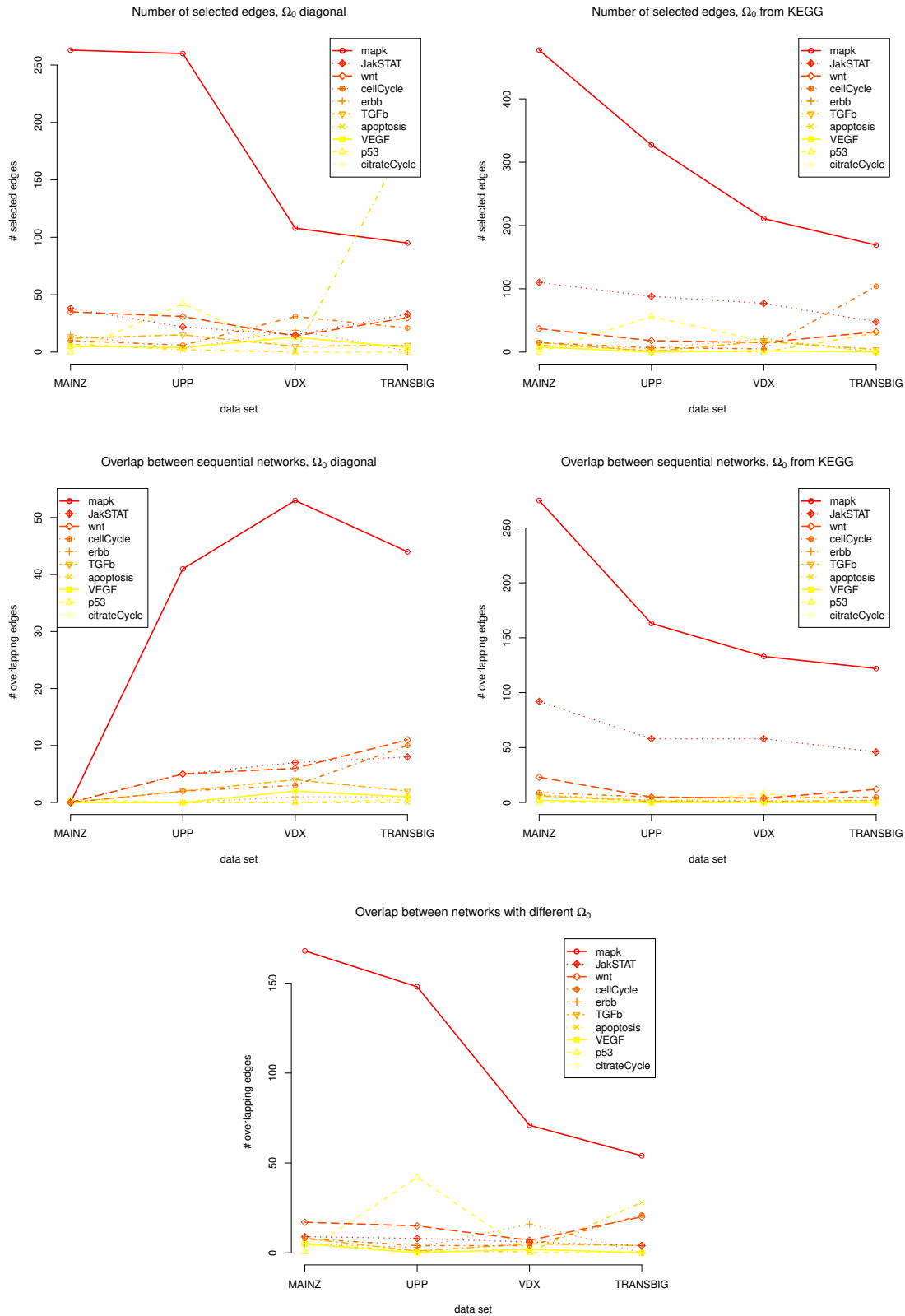


Figure 25: Analysis results with datasets in random order. Top panels: the number of edges in the support inferred from updated ridge precision estimates initiated with a diagonal target (left) and a KEGG inspired target (right). Middle panels: the number of overlapping edges in the support inferred from subsequent updated ridge precision estimates initiated with a diagonal target (left) and a KEGG inspired target (right). Bottom panel: the number of overlapping edges between the support from the update ridge precision initiated with a diagonal and a KEGG inspired target.

Application, R-code

The application uses the following scripts:

- `application1_preprocessingData.r`
- `application2 Updating.r`
- `application3_supportComparison.r`

The R-script work on a 'copy+paste' basis.

References

- Bilgrau, A.E., Peeters, C.F.W, Eriksen, P.S., Bogsted, M., and van Wieringen, W.N. (2015). Targeted fused ridge estimation of inverse covariance matrices from multiple high-dimensional data classes. *arXiv preprint*, arXiv:1509.07982.
- Creighton, C.J., Fu, X., Hennessy, B.T. and others (2010). Proteomic and transcriptomic profiling reveals a link between the PI3K pathway and lower estrogen-receptor (ER) levels and activity in ER+ breast cancer. *Breast Cancer Research*, 12(3), p.R40.
- Davis, C., and Kahan, W.M. (1970). The rotation of eigenvectors by a perturbation. III. *SIAM Journal on Numerical Analysis*, 7(1), 1–46.
- Desmedt, C., Piette, F., Loi, S., Wang, Y., Lallemand, F., Haibe-Kains, B., Viale, G., Delorenzi, M., Zhang, Y., d’Assignies, M.S. and Bergh, J. (2007). Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the TRANSBIG multicenter independent validation series. *Clinical Cancer Research*, 13(11), 3207–3214.
- Efron, B. (2004). Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *Journal of the American Statistical Association*, 99(465), 96–104.
- Friedman, J., Hastie, T. and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3), 432–441.
- Gentleman, R.C., Carey, V.J., Bates, D.M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J. and Hornik, K. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*, 5(10), R80.
- Horn, A., and Johnson, C.R. (1990). *Matrix Analysis*. Cambridge University Press.
- Jamison, B., Orey, S., and Pruitt, W. (1965). Convergence of weighted averages of independent random variables. *Zeitschrift fr Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 4(1), 40–44.
- Kolar, M., and Liu, H. (2012). Marginal regression for multitask learning. *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, Volume 22, 647–655.
- Miller, L.D., Smeds, J., George, J., Vega, V.B., Vergara, L., Ploner, A., Pawitan, Y., Hall, P., Klaar, S., Liu, E.T., and Bergh, J. (2005). An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. *Proceedings of the National Academy of Sciences of the United States of America*, 102(38), 13550–13555.
- 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(1):28–34.
- Rudin, W. (1964). *Principles of Mathematical Analysis* (3rd edition). New York: McGraw-hill.
- Schmidt, M., Böhm, D., von Törne, C., Steiner, E., Puhl, A., Pilch, H., Lehr, H.A., Hengstler, J.G., Kölbl, H. and Gehrman, M. (2008). The humoral immune system has a key prognostic impact in node-negative breast cancer. *Cancer Research*, 68(13), 5405–5413.
- Schröder, M., Haibe-Kains, B., Culhane, A., Sotiriou, C., Bontempi, G., and Quackenbush, J. (2011). `breastCancerMAINZ`; `breastCancerTRANSBIG`; `breastCancerUNT`; `breastCancerUPP`; `breastCancerVDX`. R packages, versions 1.16.0.
- Sotiriou, C., Wirapati, P., Loi, S., Harris, A., Fox, S., Smeds, J., Nordgren, H., Farmer, P., Praz, V., Haibe-Kains, B., Desmedt, C., Larsimont, D., Cardoso, F., Peterse, H., Nuyten, D., Marc Buyse, M., Van de Vijver, M.J., Bergh, J., Piccart, M., and Delorenzi, M. (2006). Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. *Journal of the National Cancer Institute*, 98(4), 262–272.

- Stachurski, J. (2009). *Economic Dynamics: Theory and Computation*. Cambridge, MA: MIT press.
- Stewart, G.W. and Sun, J.G. (1990). *Matrix Perturbation Theory*. Academic Press Boston.
- Stachurski, J. (2009). *Economic Dynamics: Theory and Computation*. Cambridge, MA: MIT press.
- Strimmer, K. (2008). A unified approach to false discovery rate estimation. *BMC Bioinformatics*, 9(1):303.
- van Wieringen, W. N. (2017). On the mean squared error of the ridge estimator of the covariance and precision matrix. *Statistics and Probability Letters*, 123, 88–92.
- 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.
- Vershynin, R. (2012). Introduction to the non-asymptotic analysis of random matrices. Chapter 5 of: *Compressed Sensing, Theory and Applications*, edited by Y. Eldar and G. Kutyniok. Cambridge University Press, 2012
- Wang, Y., Klijn, J.G., Zhang, Y., Sieuwerts, A.M., Look, M.P., Yang, F., Talantov, D., Timmermans, M., Meijer-van Gelder, M.E., Yu, J. and Jatke, T. (2005). Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *The Lancet*, 365(9460), 671–679.
- Weinstein, J.N., Collisson, E.A., Mills, G.B., Mills Shaw, K.R., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., Stuart, J.M., Cancer Genome Atlas Research Network and others. 2013. The cancer genome atlas pan-cancer analysis project. *Nature Genetics*, 45(10), 1113–1120.
- Yu, Y., Wang, T., and Samworth, R.J. (2015). A useful variant of the Davis-Kahan theorem for statisticians. *Biometrika*, 102(2), 315–323.