TEST (2019) 28:694–727
https://doi.org/10.1007/s11749-018-0592-4

ORIGINAL PAPER

NOVELIST estimator of large correlation and covariance matrices and their inverses

Na Huang[1] · **Piotr Fryzlewicz**[1]

Received: 14 June 2017 / Accepted: 7 June 2018 / Published online: 11 July 2018

Abstract
We propose a "NOVEL Integration of the Sample and Thresholded covariance" (NOVELIST) estimator to estimate the large covariance (correlation) and precision matrix. NOVELIST estimator performs shrinkage of the sample covariance (correlation) towards its thresholded version. The sample covariance (correlation) component is non-sparse and can be low rank in high dimensions. The thresholded sample covariance (correlation) component is sparse, and its addition ensures the stable invertibility of NOVELIST. The benefits of the NOVELIST estimator include simplicity, ease of implementation, computational efficiency and the fact that its application avoids eigenanalysis. We obtain an explicit convergence rate in the operator norm over a large class of covariance (correlation) matrices when the dimension p and the sample size n satisfy $\log p/n \to 0$, and its improved version when $p/n \to 0$. In empirical comparisons with several popular estimators, the NOVELIST estimator performs well in estimating covariance and precision matrices over a wide range of models and sparsity classes. Real-data applications are presented.

Keywords Covariance regularisation · High-dimensional covariance · Long memory · Non-sparse modelling · Singular sample covariance · High dimensionality

Mathematics Subject Classification 62G05 · 62H12

1 Introduction

Estimating the covariance matrix and its inverse, also known as the concentration or precision matrix, has always been an important part of multivariate analysis and arises prominently, for example, in financial risk management (Markowitz 1952; Longerstaey et al. 1996), linear discriminant analysis (Fisher 1936; Guo et al. 2007),

✉ Piotr Fryzlewicz
p.fryzlewicz@lse.ac.uk

[1] Department of Statistics, London School of Economics, London, UK

principal component analysis (Pearson 1901; Croux and Haesbroeck 2000) and network science (Jeong et al. 2001; Gardner et al. 2003). Naturally, this is also true of the correlation matrix, and the following discussion applies to it, too. The sample covariance matrix is a straightforward and often used estimator of the covariance matrix. However, when the dimension p of the data is larger than the sample size n, the sample covariance matrix is singular. Even if p is smaller than but of the same order of magnitude as n, the number of parameters to estimate is $p(p+1)/2$, which can significantly exceed n. In this case, the sample covariance matrix is not reliable, and alternative estimation methods are needed.

We would categorise the most commonly used alternative covariance estimators into two broad classes. Estimators in the first class rely on various structural assumptions on the underlying true covariance. One prominent example is ordered covariance matrices, often appearing in time-series analysis, spatial statistics and spatio-temporal modelling; these assume that there is a metric on the variable indices. Bickel and Levina (2008a) use banding to achieve consistent estimation in this context. Furrer and Bengtsson (2007) and Cai et al. (2010) regularise estimated ordered covariance matrices by tapering. Cai et al. (2010) derive the optimal estimation rates for the covariance matrix under the operator and Frobenius norms, a result which implies sub-optimality of the convergence rate of the banding estimator of Bickel and Levina (2008a) in the operator norm. The estimator of Cai et al. (2010) only achieves the optimal rate if the bandwidth parameter is chosen optimally; however, the optimal bandwidth depends crucially on the underlying unknown covariance matrix, and therefore, this estimator's optimality is only oracular. The banding technique is also applied to the estimated Cholesky factorisation of the covariance matrix (Bickel and Levina 2008a; Wu and Pourahmadi 2003).

Another important example of a structural assumption on the true covariance or precision matrices is sparsity; it is often made, e.g. in the statistical analysis of genetic regulatory networks (Gardner et al. 2003; Jeong et al. 2001). El Karoui (2008) and Bickel and Levina (2008b) regularise the estimated sparse covariance matrix by universal thresholding. Adaptive thresholding, in which the threshold is a random function of the data (Cai and Liu 2011; Fryzlewicz 2013), leads to more natural thresholding rules and hence, potentially, more precise estimation. The Lasso penalty is another popular way to regularise the covariance and precision matrices (Zou 2006; Rothman et al. 2008; Friedman et al. 2008). Focusing on model selection rather than parameter estimation, Meinshausen and Bühlmann (2006) propose the neighbourhood selection method. One other commonly occurring structural assumption in covariance estimation is the factor model, often used, e.g. in financial applications. Fan et al. (2008) impose sparsity on the covariance matrix via a factor model. Fan et al. (2013) propose the POET estimator, which assumes that the covariance matrix is the sum of a part derived from a factor model, and a sparse part.

Estimators in the second broad class do not assume a specific structure of the covariance or precision matrices, but shrink the sample eigenvalues of the sample covariance matrix towards an assumed shrinkage target (Ledoit and Wolf 2012). A considerable number of shrinkage estimators have been proposed along these lines. Ledoit and Wolf (2004) derive an optimal linear shrinkage formula, which imposes the same shrinkage intensity on all sample eigenvalues but leave the sample eigenvectors unchanged. Nonlinear shrinkage is considered in Ledoit and Péché (2011) and Ledoit and Wolf (2012,

2015). Lam (2016) introduces a Nonparametric Eigenvalue-Regularised Covariance Matrix Estimator (NERCOME) through subsampling of the data, which is asymptotically equivalent to the nonlinear shrinkage method of Ledoit and Wolf (2012). Shrinkage can also be applied on the sample covariance matrix directly. Ledoit and Wolf (2003) propose a weighted average estimator of the covariance matrix with a single-index factor target. Schäfer and Strimmer (2005) review six different shrinkage targets. Naturally related to the shrinkage approach is Bayesian estimation of the covariance and precision matrices. Evans (1965), Chen (1979) and Dickey et al. (1985) use possibly the most natural covariance matrix prior, the inverted Wishart distribution. Other notable references include Leonard and John (2012) and Alvarez et al. (2014).

The POET method of Fan et al. (2013) proposes to estimate the covariance matrix as the sum of a non-sparse, low-rank matrix coming from the factor model part, and a certain sparse matrix, added on to ensure invertibility of the resulting covariance estimator. In this paper, we are motivated by the general idea of building a covariance estimator as the sum of a non-sparse and a sparse part. By following this route, the resulting estimator can be hoped to perform well in estimating both non-sparse and sparse covariance matrices if the amount of sparsity is chosen well. At the same time, the addition of the sparse part can guarantee stable invertibility of the estimated covariance, a prerequisite for the successful estimation of the precision matrix. On the other hand, we wish to move away from the heavy modelling assumptions used by the POET estimator; indeed, our empirical results presented later suggest that POET can underperform if the factor model assumption does not hold.

Motivated by this observation, this paper proposes a simple, practically assumption-free estimator of the covariance and correlation matrices, termed NOVELIST (NOVEL Integration of the Sample and Thresholded covariance/correlation estimators). NOVELIST arises as the linear combination of two parts: the sample covariance (correlation) estimator, which is always non-sparse and has low rank if $p > n$, and its thresholded version, which is sparse. The inclusion of the sparse thresholded part means that NOVELIST can always be made stably invertible. NOVELIST can be viewed as a shrinkage estimator where the sample covariance (correlation) matrix is shrunk towards a flexible, nonparametric, sparse target. By selecting the appropriate amount of contribution of either of the two components, NOVELIST can adapt to a wide range of underlying covariance structures, including sparse but also non-sparse ones. In the paper, we show consistency of the NOVELIST estimator in the operator norm uniformly under a class of covariance matrices introduced by Bickel and Levina (2008b), as long as $\log p/n \to 0$, and offer an improved version of this result if $p/n \to 0$. The benefits of the NOVELIST estimator include simplicity, ease of implementation, computational efficiency and the fact that its application avoids eigenanalysis, which is unfamiliar to some practitioners. In our simulation studies, NOVELIST performs well in estimating both covariance and precision matrices for a wide range of underlying covariance structures, benefitting from the flexibility in the selection of its shrinkage intensity and thresholding level.

The rest of the paper is organised as follows. In Sect. 2, we introduce the NOVELIST estimator and its properties. Section 3 discusses the case where the two components of the NOVELIST estimator are combined in a non-convex way. Section 4 describes

the procedure for selecting its parameters. Section 5 shows empirical improvements of NOVELIST. Section 6 exhibits practical performance of NOVELIST in comparison with the state of the art. Section 7 presents real-data performance in portfolio optimisation problems and concludes the paper, and proofs appear in "Appendix" section. The R package "novelist" is available on CRAN.

2 Method, motivation and properties

2.1 Notation and method

We observe n i.i.d. p-dimensional observations $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$, distributed according to a distribution F, with $E(\boldsymbol{X}) = 0$, $\Sigma = \{\sigma_{ij}\} = E(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}})$, and $R = \{\rho_{ij}\} = D^{-1}\Sigma D^{-1}$, where $D = (\mathrm{diag}(\Sigma))^{1/2}$. In the case of heteroscedastic data, we apply NOVELIST to the sample correlation matrix and only then obtain the corresponding covariance estimator. The NOVELIST estimator of the correlation matrix is defined as

$$\hat{R}^N(\hat{R}, \lambda, \delta) = \underbrace{(1-\delta)\,\hat{R}}_{\text{non-sparse part}} + \underbrace{\delta\, T(\hat{R}, \lambda)}_{\text{sparse part}}, \tag{1}$$

and the corresponding covariance estimator is defined as $\hat{\Sigma}^N = \hat{D}\hat{R}^N\hat{D}$, where $\hat{\Sigma} = \{\hat{\sigma}_{ij}\}$ and $\hat{R} = \{\hat{\rho}_{ij}\}$ are the sample covariance and correlation matrices, respectively, $\hat{D} = (\mathrm{diag}(\hat{\Sigma}))^{1/2}$, δ is the weight or shrinkage intensity, which is usually within the range [0, 1] but can also lie outside it, λ is the thresholding value, which is a scalar parameter in [0, 1], and $T(\cdot, \cdot)$ is a function that applies any generalised thresholding operator (Rothman et al. 2009) to each off-diagonal entry of its first argument, with the threshold value equal to its second argument. The generalised thresholding operator refers to any function satisfying the following conditions for all $z \in \mathbb{R}$, (i) $|\, T(z, \lambda)\, | \leq |\, z\, |$; (ii) $T(z, \lambda) = 0$ for $|\, z\, | \leq \lambda$; (iii) $|\, T(z, \lambda) - z\, | \leq \lambda$. Typical examples of T include soft thresholding T_s with $T(z, \lambda) = (z - \mathrm{Sign}(z)\lambda)\mathbb{1}(|\, z\, | > \lambda)$, hard thresholding T_h with $T(z, \lambda) = z\mathbb{1}(|\, z\, | > \lambda)$ and SCAD (Fan and Li 2001). Note that $\hat{\Sigma}^N$ can also be written directly as a NOVELIST estimator with a $p \times p$ adaptive threshold matrix Λ, $\hat{\Sigma}^N = (1-\delta)\,\hat{\Sigma} + \delta\, T(\hat{\Sigma}, \Lambda)$, where $\Lambda = \{\lambda\hat{\sigma}_{ii}\hat{\sigma}_{jj}\}$.

NOVELIST is a shrinkage estimator, in which the shrinkage target is assumed to be sparse. The degree of shrinkage is controlled by the δ parameter and the amount of sparsity in the target by the λ parameter. Numerical results shown in Fig. 1 suggest that the eigenvalues of the NOVELIST estimator arise as a certain nonlinear transformation of the eigenvalues of the sample correlation (covariance) matrix, although the application of NOVELIST avoids explicit eigenanalysis.

2.2 Motivation: link to ridge regression

In this section, we show how the NOVELIST estimator can arise in a penalised solution to the linear regression problem, which is linked to ridge regression. For linear

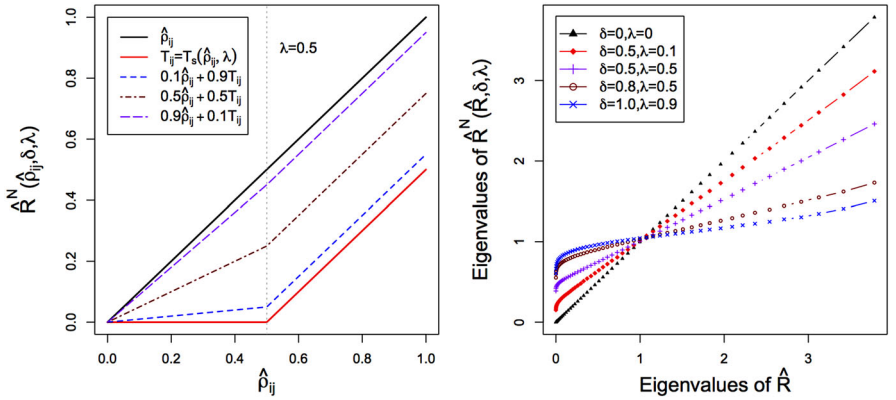


Fig. 1 Left: Illustration of NOVELIST operators for any off-diagonal entry of the correlation matrix $\hat{\rho}_{ij}$ with soft thresholding target T_s ($\lambda = 0.5$, $\delta = 0.1, 0.5$ and 0.9). Right: ranked eigenvalues of NOVELIST plotted versus ranked eigenvalues of the sample correlation matrix

regression $Y = \tilde{\boldsymbol{X}}\beta + \varepsilon$, the traditional OLS solution $(\tilde{\boldsymbol{X}}^{\mathrm{T}}\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{X}}^{\mathrm{T}}Y$ cannot be used if $p > n$ because of the non-invertibility of $\tilde{\boldsymbol{X}}^{\mathrm{T}}\tilde{\boldsymbol{X}}$. The OLS solution rewrites as $[(1-\delta)\tilde{\boldsymbol{X}}^{\mathrm{T}}\tilde{\boldsymbol{X}} + \delta\tilde{\boldsymbol{X}}^{\mathrm{T}}\tilde{\boldsymbol{X}}]^{-1}\tilde{\boldsymbol{X}}^{\mathrm{T}}Y$, where $\delta \in [0, 1]$. Using this as a starting point, we consider a regularised solution

$$\left[(1-\delta)\tilde{\boldsymbol{X}}^{\mathrm{T}}\tilde{\boldsymbol{X}} + \delta f\left(\tilde{\boldsymbol{X}}^{\mathrm{T}}\tilde{\boldsymbol{X}}\right)\right]^{-1}\tilde{\boldsymbol{X}}^{\mathrm{T}}Y \doteq A^{-1}\tilde{\boldsymbol{X}}^{\mathrm{T}}Y, \tag{2}$$

where $f(\tilde{\boldsymbol{X}}^{\mathrm{T}}\tilde{\boldsymbol{X}})$ is any elementwise modification of the matrix $\tilde{\boldsymbol{X}}^{\mathrm{T}}\tilde{\boldsymbol{X}}$ designed (a) to make A invertible and (b) to ensure adequate estimation of β. The expression in (2) is the minimiser of a generalised ridge regression criterion

$$(1-\delta)\ ||\ Y - \tilde{\boldsymbol{X}}\beta\ ||_2^2 + \delta\beta^{\mathrm{T}} f\left(\tilde{\boldsymbol{X}}^{\mathrm{T}}\tilde{\boldsymbol{X}}\right)\beta, \tag{3}$$

where δ acts as a tuning parameter. If $f(\tilde{\boldsymbol{X}}^{\mathrm{T}}\tilde{\boldsymbol{X}}) = I$, (3) is reduced to ridge regression and A is the shrinkage estimator with the identity matrix target. If $f(\tilde{\boldsymbol{X}}^{\mathrm{T}}\tilde{\boldsymbol{X}}) = T(\tilde{\boldsymbol{X}}^{\mathrm{T}}\tilde{\boldsymbol{X}}, \lambda\hat{\sigma}_{ii}\hat{\sigma}_{jj})$, A is the NOVELIST estimator of the covariance matrix.

From formula (3), NOVELIST penalises the regression coefficients in a pairwise manner which can be interpreted as follows: for a given threshold λ, we place a penalty on the products $\beta_i\beta_j$ of those coefficients of β for which the sample correlation between $\tilde{\boldsymbol{X}}_i$ and $\tilde{\boldsymbol{X}}_j$, the ith and jth column of $\tilde{\boldsymbol{X}}$ (respectively), exceeds λ. In other words, if the sample correlation is high, we penalise the product of the corresponding β's, hoping that the resulting estimated β_i and β_j are not simultaneously large.

2.3 Asymptotic properties of NOVELIST

2.3.1 Consistency of the NOVELIST estimators

In this section, we establish consistency of NOVELIST in the operator norm and derive the rates of convergence under different scenarios. Bickel and Levina (2008b) introduce a uniformity class of covariance matrices invariant under permutations as

$$\mathcal{U}(q, c_0(p), M, \epsilon_0) = \left\{ \Sigma : \sigma_{ii} \leq M, \sum_{j=1}^{p} \mid \sigma_{ij} \mid^q \leq c_0(p), \text{ for all } i \text{ and } \lambda_{\min}(\Sigma) \geq \epsilon_0 > 0 \right\}, \quad (4)$$

where $0 \leq q < 1$, c_0 is a function of p, the parameters M and ϵ_0 are constants, and $\lambda_{\min}()$ is the smallest eigenvalue operator. Analogously, we define a uniformity class of correlation matrices as

$$\mathcal{V}(q, s_0(p), \varepsilon_0) = \left\{ R : \sum_{j=1}^{p} \mid \rho_{ij} \mid^q \leq s_0(p), \text{ for all } i \text{ and } \lambda_{\min}(R) \geq \varepsilon_0 > 0 \right\}, \quad (5)$$

where $0 \leq q < 1$ and ε_0 is a constant. The parameters q and $s_0(p)$ (equiv. $c_0(p)$) together control the permitted degree of "sparsity" of the members of the given class. In the remainder of the paper, where it does not cause confusion, we mostly work with $s_0(p)$ rather than $c_0(p)$, noting that these two parameterisations are equivalent.

Next, we establish consistency of the NOVELIST estimator in the operator norm, $|| A ||_2^2 = \lambda_{\max}(AA^{\mathrm{T}})$, where $\lambda_{\max}()$ is the largest eigenvalue operator.

Proposition 1 *Let F satisfy $\int_0^\infty \exp(\gamma t) dG_j(t) < \infty$ for $0 < |\gamma| < \gamma_0$, where $\gamma_0 > 0$ and G_j is the cdf of X_{1j}^2. Let $R = \{\rho_{ij}\}$ and $\Sigma = \{\sigma_{ij}\}$ be the true correlation and covariance matrices with $1 \leq i, j \leq p$, and $\sigma_{ii} \leq M$, where $M > 0$. Then, uniformly on $\mathcal{V}(q, s_0(p), \varepsilon_0)$, for sufficiently large M', if $\lambda = M'\sqrt{\log p/n}$ and $\log p/n = o(1)$,*

$$|| \hat{R}^N - R || = \underbrace{O_p((1-\delta)p\sqrt{\log p/n})}_{(A)} + \underbrace{O_p(\delta s_0(p)(\log p/n)^{(1-q)/2})}_{(B)} = || (\hat{R}^N)^{-1} - R^{-1} || \quad (6)$$

$$|| \hat{\Sigma}^N - \Sigma || = O_p((1-\delta)p\sqrt{\log p/n}) + O_p(\delta s_0(p)(\log p/n)^{(1-q)/2}) = || (\hat{\Sigma}^N)^{-1} - \Sigma^{-1} || . \quad (7)$$

Proposition 2 *Let the length-p column vector $\boldsymbol{X}_i$ satisfy the sub-Gaussian condition $P(|v^T \boldsymbol{X}_i| > t) \leq \exp(-t^2\rho/2)$ for a certain $\rho > 0$, for all $t > 0$ and $\|v\|_2 = 1$. Let $R = \{\rho_{ij}\}$ and $\Sigma = \{\sigma_{ij}\}$ be the true correlation and covariance matrices with*

$1 \leq i, j \leq p$, and $\sigma_{ii} \leq M$, where $M > 0$. Then, uniformly on $\mathcal{V}(q, s_0(p), \varepsilon_0)$, for sufficiently large M', if $\lambda = M'\sqrt{\log p/n}$ and $p = o(n)$,

$$|| \hat{R}^N - R || = \underbrace{O_p((1-\delta)\sqrt{(p+\log n)/n})}_{(A)} + \underbrace{O_p(\delta s_0(p)(\log p/n)^{(1-q)/2})}_{(B)} = || (\hat{R}^N)^{-1} - R^{-1} || \quad (8)$$

$$|| \hat{\Sigma}^N - \Sigma || = O_p((1-\delta)\sqrt{(p+\log n)/n}) + O_p(\delta s_0(p)(\log p/n)^{(1-q)/2}) = || (\hat{\Sigma}^N)^{-1} - \Sigma^{-1} || . \quad (9)$$

The proofs are given in "Appendix" section. The NOVELIST estimators of the correlation and covariance matrices and their inverses yield the same convergence rate.

We now discuss the optimal asymptotic δ under the settings of Propositions 1 and 2. Proposition 1 can be thought of as a "large p" setting, while Proposition 2 applies to moderately large and small p.

2.3.2 Optimal δ and rate of convergence in Proposition 1

Proposition 1 corresponds to "large p" scenarios, in which p can be thought of as being $O(n)$ or larger (indeed, the case $p = o(n)$ is covered by Proposition 2). For such a large p, the pre-condition for the consistency of the NOVELIST estimator is that $\delta \to 1$, i.e. that the estimator asymptotically degenerates to the thresholding estimator. To see this, take $p = n^{1+\Delta}$ with $\Delta \geq 0$. If $\delta \not\to 1$, the error in part (A) of formula (6) would be of order $n^{1/2+\Delta}\sqrt{\log n^{1+\Delta}}$ and therefore would not converge to zero.

Focusing on $\hat{R}^N$ without loss of generality, the optimal rate of convergence is obtained by equating parts (A) and (B) in formula (6). The resulting optimal shrinkage intensity $\tilde{\delta}$ is

$$\tilde{\delta} = \frac{p(\log p/n)^{q/2}}{s_0(p) + p(\log p/n)^{q/2}} = \frac{(\log p/n)^{q/2}}{s_0(p)/p + (\log p/n)^{q/2}}. \quad (10)$$

In typical scenarios, bearing in mind that p is at least of order n or larger, and that $q < 1$, the term $s_0(p)/p$ will tend to zero much faster than the term $(\log p/n)^{q/2}$, which will result in $\tilde{\delta} \to 1$ and in the rate of convergence of NOVELIST being $O_p(s_0(p)(\log p/n)^{(1-q)/2})$. Examples or such scenarios are given directly below.

Scenario 1 $q = 0$, $s_0(p) = o((n/\log p)^{1/2})$.

When $q = 0$, the uniformity class of correlation matrices controls the maximum number of nonzero entries in each row. The typical example is the moving-average (MA) autocorrelation structure in time series.

Scenario 2 $q \neq 0$, $s_0(p) \leq C$ *as* $p \to \infty$.

A typical example of this scen is the auto-regressive (AR) autocorrelation structure.

We now show a scen in which NOVELIST is inconsistent, under the setting of Proposition 1. Consider the long-memory autocorrelation matrix, $\rho_{ij} \sim \mid i-j \mid^{-\alpha}, 0 < \alpha \leq 1$, for which $s_0(p) = \max_{1 \leq i \leq p} \sum_{j=1}^{p} \max(1, |i-j|)^{-\alpha q} = O(p^{1-\alpha q})$. Take $q \neq 0$. Note a sufficient condition for $\tilde{\delta}$ to tend to 1 is that $(\log\ p)^{(1/2)} n^{-1/2} p^{\alpha} \to \infty$. This more easily happens for larger α's, i.e. for "less long"-memory processes. However, considering the implied rate of convergence, we have $s_0(p)(\log\ p/n)^{(1-q)/2} = p^{1-\alpha q}(\log\ p/n)^{(1-q)/2}$, which is divergent even if $\alpha = 1$.

2.3.3 Optimal δ and rate of convergence in Proposition 2

Similarly, in the setting of Proposition 2, the resulting optimal shrinkage intensity $\tilde{\delta}$ is

$$\tilde{\delta} = \frac{((p + \log n)/n)^{1/2}}{((p + \log n)/n)^{1/2} + s_0(p)(\log p/n)^{(1-q)/2}}. \tag{11}$$

We now highlight a few special-case scenarios.

Scenario 3 *p fixed (and hence* $q = 0$*).*

Note that in the case of p being fixed or bounded in n, one can take $q = 0$ (to obtain as fast a rate for the thresholding part as possible) as the implied $s_0(p)$ will also be bounded in n. In this case, we have $\tilde{\delta} \to 1$ (and hence NOVELIST degenerates to the thresholding estimator with its corresponding speed of convergence), but the speed at which $\tilde{\delta}$ approaches 1 is extremely slow ($O(\log^{-1/2} n)$).

Scenario 4 $p \to \infty$ *with n, and* $q = 0$.

In this case, the quantity $\{(p + \log\ n)/\log\ p\}^{1/2}$ acts as a transition phase: if $s_0(p)$ is of a larger order, then we have $\tilde{\delta} \to 0$; if it is of a smaller order, then $\tilde{\delta} \to 1$; if it is of this order and if $\tilde{\delta}$ has a limit, then its limit lies in (0, 1). Therefore, NOVELIST will be closer to the sample covariance (correlation) if the truth is dense (i.e. if $s_0(p)$ is large), and closer to the thresholding estimator if $s_0(p)$ is small.

Scenario 5 $p \to \infty$ *with n, and* $q \neq 0$.

Here, the transition-phase quantity is $\frac{(p+\log\ n)^{1/2}}{(\log\ p)^{\frac{1-q}{2}} n^{q/2}}$ and conclusions analogous to those of the preceding Scenario can be formed.

In the context of Scenario 5, we now revisit the long-memory example from before. The most "difficult" case still included in the setting of Proposition 2 is when p is "almost" the size of n; therefore, we assume $p = n^{1-\Delta}$, with Δ being a small positive constant. Neglecting the logarithmic factors, the transition-phase quantity $\frac{(p+\log\ n)^{1/2}}{(\log\ p)^{\frac{1-q}{2}} n^{q/2}}$ reduces to $n^{\frac{1-\Delta-q}{2}}$. We have $s_0(p) = O(n^{(1-\Delta)(1-\alpha q)})$, and therefore $s_0(p)$ is of a larger order than $n^{\frac{1-\Delta-q}{2}}$ if $\alpha < \frac{1-\Delta+q}{2q(1-\Delta)}$; in this case, $\tilde{\delta} \to 0$, and the

NOVELIST estimator degenerates to the sample covariance (correlation) estimator, which is consistent in this setting at the rate of $n^{-\Delta/2}$ (neglecting the log-factors). The other case, $\alpha \geq \frac{1-\Delta+q}{2q(1-\Delta)}$, is impossible as we must have $\alpha \leq 1$. Therefore, the NOVELIST estimator is consistent for the long-memory model under the setting of Proposition 2, i.e. when $p = o(n)$ (and degenerates to the sample covariance estimator). This is in contrast to the setting of Proposition 1, where, as argued before, the consistency of NOVELIST in the long-memory model cannot be shown.

3 δ outside [0, 1]

Some authors (Ledoit and Wolf 2003; Schäfer and Strimmer 2005; Savic and Karlsson 2009), more or less explicitly, discuss the issue of the shrinkage intensity (for other shrinkage estimators) falling within versus outside the interval [0, 1]. Ledoit and Wolf (2003) "expect" it to lie between zero and one, Schäfer and Strimmer (2005) truncate it at zero or one, and Savic and Karlsson (2009) view negative shrinkage as a "useful signal for possible model misspecification". We are interested in the performance of the NOVELIST estimator with $\delta \notin [0, 1]$ and have reasons to believe that $\delta \notin [0, 1]$ may be a good choice in certain scenarios.

We use the diagrams below to briefly illustrate this point. When the target T is appropriate, the "oracle" NOVELIST estimator (by which we mean one where δ is computed with the knowledge of the true R by minimising the spectral norm distance to R) will typically be in the convex hull of $\hat{R}$ and T, i.e. $\delta \in [0, 1]$ as shown in the left graph. However, the target may also be misspecified. For example, if the true correlation matrix is highly non-sparse, the sparse target may be inappropriate, to the extent that R will be further away from T than from $\hat{R}$, as shown in the middle graph. In that case, the optimal δ should be negative to prevent NOVELIST being close to the target. By contrast, when the sample correlation is far from the (sparse) truth, perhaps because of high dimensionality, the optimal delta may be larger than one (Diagram 1).

4 Empirical choices of (λ, δ) and algorithm

The choices of the shrinkage intensity (for shrinkage estimators) and the thresholding level (for thresholding estimators) are intensively studied in the literature. Bickel and

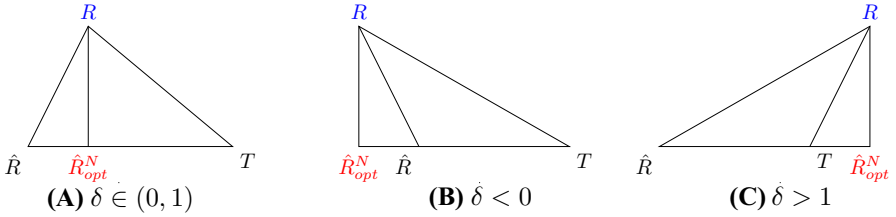


Diagram 1 Geometric illustration of shrinkage estimators. R is the truth, T is the target, $\hat{R}$ is the sample correlation, $\hat{R}^N_{opt}$ is the "oracle" NOVELIST estimator defined as the linear combination of T and $\hat{R}$ with minimum spectral norm distance to R. LEFT: $\delta \in (0, 1)$ if target T is appropriate; MIDDLE: $\delta < 0$ if target T is misspecified; RIGHT: $\delta > 1$ if $\hat{R}$ is far from R

Levina (2008b) propose a cross-validation method for choosing the threshold value for their thresholding estimator. However, NOVELIST requires simultaneous selection of the two parameters λ and δ, which makes straight cross-validation computationally intensive. Ledoit and Wolf (2003) and Schäfer and Strimmer (2005) give an analytic solution to the problem of choosing the optimal shrinkage level, under the Frobenius norm, for any shrinkage estimator. Since NOVELIST can be viewed as a shrinkage estimator, we borrow strength from this result and proceed by selecting the optimal shrinkage intensity $\delta^*(\lambda)$ in the sense of Ledoit and Wolf (2003) for each λ, and then perform cross-validation to select the best pair $(\lambda', \delta^*(\lambda'))$. This process significantly accelerates computation.

Cai and Liu (2011) and Fryzlewicz (2013) use adaptive thresholding for covariance matrices, in order to make thresholding insensitive to changes in the variance of the individual variables. This, effectively, corresponds to thresholding sample correlations rather than covariances. In the same vein, we apply NOVELIST to sample correlation matrices. We use soft thresholding as it often exhibits better and more stable empirical performance than hard thresholding, which is partly due to its being a continuous operation. Let $\hat{\Sigma}$ and $\hat{R}$ be the sample covariance and correlation matrices computed on the whole dataset, and let $T = \{T_{ij}\}$ be the soft thresholding estimator of the correlation matrix. The algorithm proceeds as follows.

For estimating the covariance matrix,

LW (Ledoit–Wolf) step Using all available data, for each $\lambda \in (0, 1)$ chosen from a uniform grid of size m, find the optimal empirical δ as

$$\begin{aligned}\delta^*(\lambda) &= \frac{\sum_{1\le i\ne j\le n} \mathrm{Var}(\hat{R}_{ij}) - \mathrm{Cov}(\hat{R}_{ij}, T_{ij})}{\sum_{1\le i\ne j\le n}(\hat{R}_{ij} - T_{ij})^2} \\ &= \frac{\sum_{1\le i\ne j\le n} \mathrm{Var}(\hat{R}_{ij})\mathcal{I}(\hat{R}_{ij} < \lambda)}{\sum_{1\le i\ne j\le n}(\hat{R}_{ij} - T_{ij})^2},\end{aligned} \tag{12}$$

to obtain the pair $(\lambda, \delta^*(\lambda))$.

The first equality comes from Ledoit and Wolf (2003), and the second follows because of the fact that our shrinkage target T is the soft thresholding estimator with threshold λ (applied to the off-diagonal entries only).

CV (Cross-validation) step For each $z = 1, \ldots, Z$, split the data randomly into two equal-size parts A (training data) and B (test data), letting $\hat{\Sigma}_A^{(z)}$ and $\hat{\Sigma}_B^{(z)}$ be the sample covariance matrices of these two datasets, and $\hat{R}_A^{(z)}$ and $\hat{R}_B^{(z)}$ – the sample correlation matrices.

1. For each λ, obtain the NOVELIST estimator of the correlation matrix $\hat{R}_A^{N^{(z)}}(\lambda) = \hat{R}^N(\hat{R}_A^{(z)}, \lambda, \delta^*(\lambda))$, and of the covariance matrix $\hat{\Sigma}_A^{N^{(z)}}(\lambda) = \hat{D}_A \hat{R}_A^{N^{(z)}}(\lambda)\hat{D}_A$, where $\hat{D}_A = (\mathrm{diag}\,(\hat{\Sigma}_A^{(z)}))^{1/2}$.
2. Compute the spectral norm error $\mathrm{Err}(\lambda)^{(z)} = || \hat{\Sigma}_A^{N^{(z)}}(\lambda) - \hat{\Sigma}_B^{(z)} ||_2^2$.
3. Repeat steps 1 and 2 for each z and obtain the averaged error $\mathrm{Err}(\lambda) = \frac{1}{Z}\sum_{z=1}^{Z} \mathrm{Err}(\lambda)^{(z)}$. Find $\lambda' = \min_\lambda \mathrm{Err}(\lambda)$, then obtain the optimal pair $(\lambda', \delta') = (\lambda', \delta^*(\lambda'))$.

4. Compute the cross-validated NOVELIST estimators of the correlation and covariance matrices as

$$\hat{R}^N_{cv} = \hat{R}^N(\hat{R}, \lambda', \delta'), \tag{13}$$
$$\hat{\Sigma}^N_{cv} = \hat{D}\hat{R}^N_{cv}\hat{D}, \tag{14}$$

where $\hat{D} = (\mathrm{diag}(\hat{\Sigma}))^{1/2}$.

For estimating the inverses of the correlation and the covariance matrices, the difference lies in step 2, where the error measure is adjusted as follows. If $n > 2p$ (i.e. in the case when $\hat{\Sigma}^{(z)}_B$ is invertible), we use the measure $\mathrm{Err}(\lambda)^{(z)} = || \ (\hat{\Sigma}^{N^{(z)}}_A(\lambda))^{-1} - (\hat{\Sigma}^{(z)}_B)^{-1} \ ||^2_2$; otherwise, use $\mathrm{Err}(\lambda)^{(z)} = || \ (\hat{\Sigma}^{N^{(z)}}_A(\lambda))^{-1}\hat{\Sigma}^{(z)}_B - \mathcal{I} \ ||^2_2$, where $\mathcal{I}$ is the identity matrix. In step 4, we compute the cross-validated NOVELIST estimators of the inverted correlation and covariance matrices as

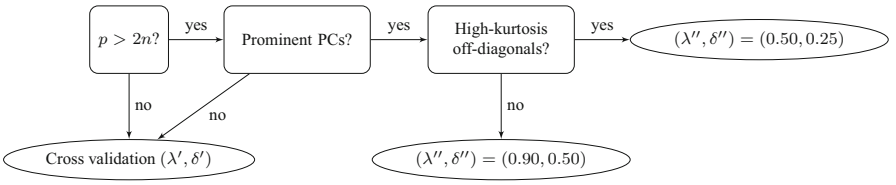
$$(\hat{R}^N_{cv})^{-1} = (\hat{R}^N(\hat{R}, \lambda', \delta'))^{-1}, \tag{15}$$
$$(\hat{\Sigma}^N_{cv})^{-1} = (\hat{D}\hat{R}^N_{cv}\hat{D})^{-1}. \tag{16}$$

We note that a closely related procedure for choosing δ has also been described in Lam and Feng (2017).

5 Empirical improvements of NOVELIST

5.1 Fixed parameters

As shown in the simulation study of Sect. 6.2, the performance of cross-validation is generally adequate, except in estimating large precision matrices with highly non-sparse covariance structures, such as in factor models and long-memory autocovariance structures. To remedy this problem, we suggest that fixed, rather than cross-validated parameters be used, if the eigenanalysis of the sample correlation matrix indicates that there are prominent principal components, when $p > 2n$ or close. We suggest the following rules of thumb: first, we look for the evidence of "elbows" in the scree plot of eigenvalues, by examining if $\sum_{k=1}^{p} \mathbb{I}\{\gamma_{(k)} + \gamma_{(k+2)} - 2\gamma_{(k+1)} > 0.1p\} > 0$, where $\gamma(k)$ is the kth principal component. If so, then we look for the evidence of long-memory decay, by examining if the off-diagonals of the sample correlation matrix follow a high-kurtosis distribution. If the sample kurtosis ≤ 3.5, this suggests that the factor structure may be present, and we use the fixed parameters $(\lambda'', \delta'') = (0.90, 0.50)$; if the sample kurtosis > 3.5, this may point to long memory, and we use the fixed parameters $(\lambda'', \delta'') = (0.50, 0.25)$. The above decision procedure, including all the specific parameter values, has been obtained through extensive numerical experiments not shown in this paper. It is sketched in the following flowchart (Flowchart 1).



Flowchart 1 Decision procedure for using cross-validated or fixed parameters in estimating precision matrices

5.2 Principal-component-adjusted NOVELIST

NOVELIST can further benefit from any prior knowledge about the underlying covariance matrix, such as the factor model structure. If the underlying correlation matrix follows a factor model, we can decompose the sample correlation matrix as

$$\hat{R} = \sum_{k=1}^{K} \hat{\gamma}_{(k)} \hat{\xi}_{(k)} \hat{\xi}_{(k)}^{\mathrm{T}} + \hat{R}_{\mathrm{rem}}, \tag{17}$$

where $\hat{\gamma}_{(k)}$ and $\hat{\xi}_{(k)}$ are the kth eigenvalue and eigenvector of sample correlation matrix, K is the number up to which the principal components are considered to be "large" and $\hat{R}_{\mathrm{rem}}$ is the sample correlation matrix after removing the first K principal components. Instead of applying NOVELIST on $\hat{R}$ directly, we keep the first K components unchanged and only apply NOVELIST to $\hat{R}_{\mathrm{rem}}$. Principal-component-adjusted NOVELIST estimators are obtained by

$$\hat{R}_{\mathrm{rem}}^{N} = \sum_{k=1}^{K} \hat{\gamma}_{(k)} \hat{\xi}_{(k)} \hat{\xi}_{(k)}^{\mathrm{T}} + \hat{R}^{N}(\hat{R}_{\mathrm{rem}}, \lambda, \delta), \tag{18}$$

$$\hat{\Sigma}_{\mathrm{rem}}^{N} = \hat{D} \hat{R}_{\mathrm{rem}}^{N} \hat{D}. \tag{19}$$

In the remainder of the paper, we always use the not-necessarily-optimal value $K = 1$. We suggest that PC-adjusted NOVELIST should only be used with prior knowledge or if empirical testing indicates that there are prominent principal components.

6 Simulation study

In this section, we investigate the performance of the NOVELIST estimator of covariance and precision matrices based on optimal and data-driven choices of (λ, δ) for seven different models and in comparison with five popular competitors. According to the algorithm in Sect. 4, the NOVELIST estimator of the correlation is obtained first; the corresponding estimator of the covariance follows by formula (13) and the inverse of the covariance estimator is obtained by formula (16). In all simulations, the sample size $n = 100$, and the dimension $p \in \{10, 100, 200, 500\}$. We perform $N = 50$ repetitions.

6.1 Simulation models

We use the following models for Σ.

(A) *Identity* $\sigma_{ij} = 1\mathbb{I}\{i = j\}$, for $1 \le i, j \le p$.

(B) *MA(1) autocovariance structure*

$$\sigma_{ij} = \begin{cases} 1, & \text{if } i = j; \\ \rho, & \text{if } \mid i - j \mid = 1; \\ 0, & \text{otherwise} \end{cases} \tag{20}$$

for $1 \le i, j \le p$. We set $\rho = 0.5$.

(C) *AR(1) autocovariance structure*

$$\sigma_{ij} = \rho^{|i-j|}, \quad \text{for } 1 \le i, j \le p, \tag{21}$$

with $\rho = 0.9$.

(D) *Non-sparse covariance structure* We generate a positive definite matrix as

$$\Sigma = Q \Lambda Q^{\mathrm{T}}, \tag{22}$$

where Q has iid standard normal entries and Λ is a diagonal matrix with its diagonal entries drawn independently from the χ^2_5 distribution. The resulting Σ is non-sparse and lacks an obvious pattern.

(E) *Factor model covariance structure* Let Σ be the covariance matrix of $\mathbf{X} = \{X_1, X_2, \cdots, X_p\}^{\mathrm{T}}$, which follows a three-factor model

$$\mathbf{X}_{p\times n} = \mathbf{B}_{p\times 3}\mathbf{Y}_{3\times n} + \mathbf{E}_{p\times n}, \tag{23}$$

where

$\mathbf{Y} = \{Y_1, Y_2, Y_3\}^{\mathrm{T}}$ is a three-dimensional factor, generated independently from the standard normal distribution, i.e. $\mathbf{Y} \sim \mathcal{N}(\mathbf{0}, \mathcal{I}_{\mathbf{3}})$,

$\mathbf{B} = \{\beta_{ij}\}$ is the coefficient matrix, $\beta_{ij} \overset{i.i.d.}{\sim} U(0, 1)$, $1 \le i \le p$, $1 \le j \le 3$,

$\mathbf{E} = \{\epsilon_1, \epsilon_2, \cdots, \epsilon_p\}^{\mathrm{T}}$ is p-dimensional random noise, generated independently from the standard normal distribution, $\sim \mathcal{N}(\mathbf{0}, \mathbf{1})$.

Based on this model, we have $\sigma_{ij} = \begin{cases} \sum_{k=1}^{3} \beta_{ik}^2 + 1 & \text{if } i = j; \\ \sum_{k=1}^{3} \beta_{ik}\beta_{jk} & \text{if } i \ne j. \end{cases}$

(F) *Long-memory autocovariance structure* We use the autocovariance matrix of the fractional Gaussian noise (FGN) process, with

$$\sigma_{ij} = \frac{1}{2}[|| \, i - j \mid +1 \mid^{2H} -2 \mid i - j \mid^{2H} + || \, i - j \mid -1 \mid^{2H}] \qquad 1 \le i, j \le p. \tag{24}$$

The model is taken from Bickel and Levina (2008a), Sect. 6.1, and is non-sparse. We take $H = 0.9$ in order to investigate the case with strong long memory.

(G) *Seasonal covariance structure*

$$\sigma_{ij} = \rho^{|i-j|}\mathbb{1}\{|\, i-j \,| = l\mathbb{Z}_{\geq 0}\}, \quad \text{for } 1 \leq i, j \leq p, \tag{25}$$

where $\mathbb{Z}_{\geq 0}$ is the set of non-negative integers. We take $l = 3$ and $\rho = 0.9$.

The models can be broadly divided into three groups. (A)–(C) and (G) are sparse, (D) is non-sparse, and (E) and (F) are highly non-sparse. In models (B), (C) (F) and (G), the covariance matrix equals the correlation matrix. In order to depart from the case of equal variances, we also work with modified versions of these models, denoted by (B*), (C*) (F*) and (G*), in which the correlation matrix $\{\rho_{ij}\}$ is generated as in (B), (C) (F) and (G), respectively, and which have unequal variances independently generated as $\sigma_{ii} \sim \chi^2_5$. As a result, in the "starred" models, we have $\sigma_{ij} = \rho_{ij}\sqrt{\sigma_{ii}\sigma_{jj}}$, $i, j \in (1, p)$.

The performance of the competing estimators is presented in two parts. In the first part, we compare the estimators with optimal parameters identified with the knowledge of the true covariance matrix. These include (a) the soft thresholding estimator T_s, which applies the soft thresholding operator to the off-diagonal entries of $\hat{R}$ only, as described in Sect. 2.1, (b) the banding estimator B (Section 2.1 in Bickel and Levina (2008a)), (c) the optimal NOVELIST estimator $\hat{\Sigma}^N_{opt}$ and (d) the optimal PC-adjusted NOVELIST estimator $\hat{\Sigma}^N_{\text{opt.rem}}$. In the second part, we compare the data-driven estimators including (e) the linear shrinkage estimator S [Target D in Table 2 from Schäfer and Strimmer (2005)], which estimates the correlation matrix by "shrinkage of the sample correlation towards the identity matrix" and estimates the variances by "shrinkage of the sample variances towards their median", (f) the POET estimator P (Fan et al. 2013), (g) the cross-validated NOVELIST estimator $\hat{\Sigma}^N_{cv}$, (h) the PC-adjusted NOVELIST $\hat{\Sigma}^N_{\text{rem}}$ and (i) the nonlinear shrinkage estimator NS (Ledoit and Wolf 2015). The sample covariance matrix $\hat{\Sigma}$ is also listed for reference. We use the R package *corpcor* to compute S and the R package *POET* to compute P. In the latter, we use $k = 7$ as suggested by the authors and use soft thresholding in NOVELIST and POET as it tends to offer better empirical performance. We use $Z = 50$ for $\hat{\Sigma}^N_{cv}$ and extend the interval for δ to $[-0.5, 1.5]$. $\hat{\Sigma}^N_{cv}$ with fixed parameters are only considered for estimating precision matrix under model (E), (F) and (F*) when $p = 100, 200, 500$. We use $K = 1$ for $\hat{\Sigma}^N_{\text{opt.rem}}$ and $\hat{\Sigma}^N_{\text{rem}}$. NS is performed by using the commercial package SNOPT for Matlab (Ledoit and Wolf 2015).

6.2 Simulation results

Performance of $\hat{\Sigma}^N$ *as a function of* (λ, δ) Examining the results presented in Figs. 2 and 3 and Table 1, it is apparent that the performance of NOVELIST depends on the combinations of λ and δ used. Generally speaking, the average operator norm errors increase as sparsity decreases and dimension p increases. The positions of empirically optimal λ^* and δ^* are summarised as follows.

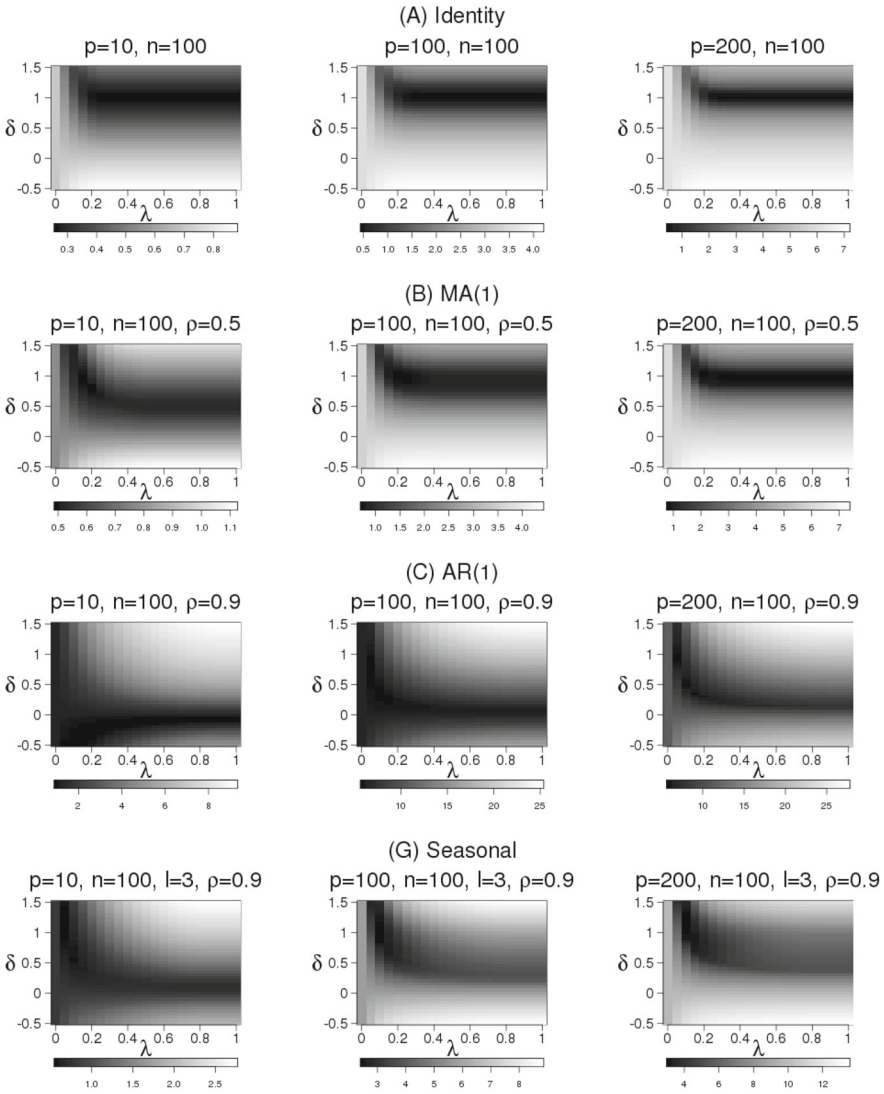


Fig 2 Image plots of operator norm errors of NOVELIST estimators of Σ with different λ and δ under models (**a**)–(**c**) and (**g**), $n = 100$, $p = 10$ (left), 100 (middle), 200 (right), simulation times = 50. The darker the area, the smaller the error

1. The higher the degree of sparsity, the closer δ^* is to 1. The δ^* parameter tends to be close to 1 or slightly larger than 1 for the sparse group, around 0.5 for the non-sparse group and about 0 or negative for the highly non-sparse group.
2. δ^* moves closer to 1 as p increases. This is especially true for the sparse group.
3. Unsurprisingly, the choice of λ is less important when δ is closer to 0.
4. Occasionally, $\delta^* \notin [0, 1]$. In particular, for the AR(1) and seasonal models, $\delta^* \in (1, 1.5]$, while in the highly non-sparse group, δ^* can take negative values, which

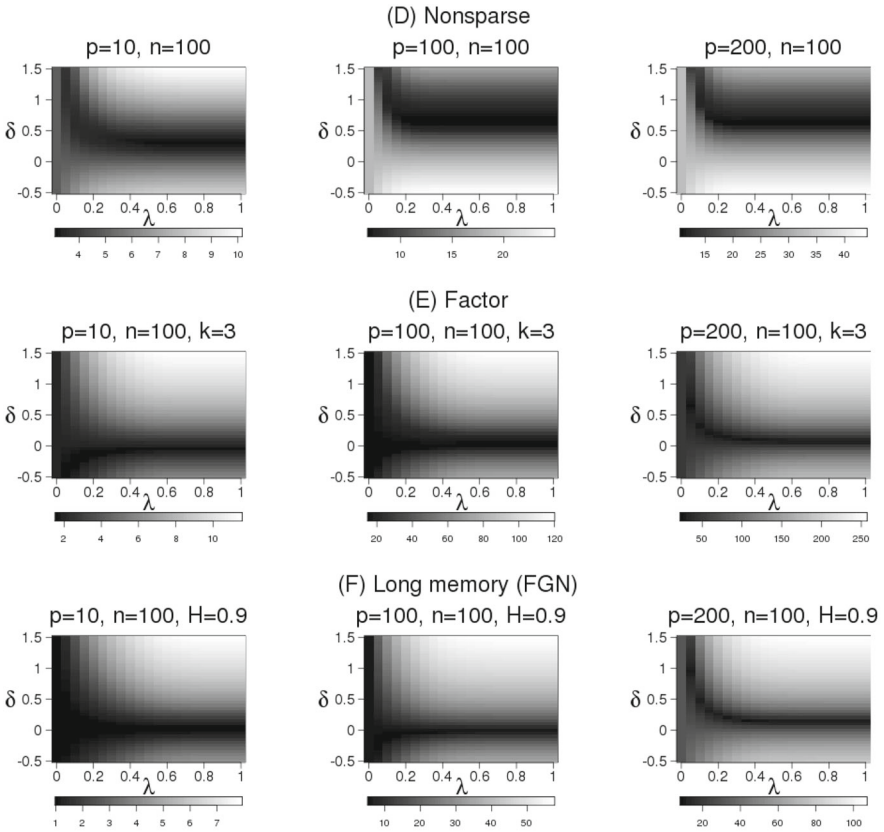


Fig 3 Image plots of operator norm errors of NOVELIST estimators of Σ with different λ and δ under models (**d**)–(**f**), $n = 100$, $p = 10$ (left), 100 (middle), 200 (right), simulation times $= 50$. The darker the area, the smaller the error

is a reflection of the fact that $\hat{\Sigma}^N_{\text{opt}}$ attempts to reduce the effect of the strongly misspecified sparse target.

Performance of cross-validated choices of (λ, δ) Table 1 shows that the cross-validated choices of the parameter (λ', δ') for $\hat{\Sigma}^N_{cv}$ are close to the optimal (λ^*, δ^*) for most models when $p = 10$, but there are bigger discrepancies between (λ', δ') and (λ^*, δ^*) as p increases, especially for the highly non-sparse group. Again, Fig. 4, which only includes representative models from each sparsity category, shows that the choices of (λ', δ') are consistent with (λ^*, δ^*) in most of the cases. For models (A) and (C), cross-validation works very well: the vast majority of (λ', δ') lead to the error lying in the 1st decile of the possible error range, whereas for models (D) and (G) with $p = 10$, in the 1st or 2nd decile.

However, as given in Tables 4 and 5, the performance of cross-validation in estimating Σ^{-1} with highly non-sparse covariance structures, such as in factor models and long-memory autocovariance structures, is less good (a remedy to this was described in Sect. 5.1).

Table 1 Choices of (λ^*, δ^*) and $(\lambda^{'}, \delta^{'})$ for $\hat{\Sigma}^N$ (50 replications)

	$\hat{\Sigma}^N_{\text{opt}}$		$\hat{\Sigma}^N_{cv}$		$\hat{\Sigma}^N_{\text{opt}}$		$\hat{\Sigma}^N_{cv}$	
	λ^*	δ^*	$\lambda^{'}$	$\delta^{'}$	λ^*	δ^*	$\lambda^{'}$	$\delta^{'}$
	$p = 10,\ n = 100$				$p = 100,\ n = 100$			
(A) Identity	(0.50,1.00)	1.00	0.60	1.00	(0.50,1.00)	1.00	0.60	1.00
(B) MA(1)	0.15	1.00	0.25	0.80	0.20	1.00	0.20	0.95
(B*) MA(1)*	0.15	0.95	0.30	0.65	0.15	1.00	0.30	0.90
(C) AR(1)	0.50	0.00	0.40	0.15	0.15	0.50	0.10	0.70
(C*) AR(1)*	0.50	0.05	0.40	0.00	0.30	0.60	0.30	0.85
(D) Non-sparse	0.40	0.50	0.55	0.40	0.45	0.60	0.35	0.80
(E) Factor	0.40	0.00	0.65	0.10	0.20	−0.15	0.50	0.05
(F) FGN	0.50	−0.05	0.50	0.00	0.30	−0.10	0.55	0.05
(F*) FGN*	0.50	−0.05	0.50	0.00	0.40	−0.05	0.65	0.05
(G) Seasonal	0.15	0.75	0.15	0.70	0.10	1.30	0.05	1.50
(G*) Seasonal*	0.25	0.75	0.20	0.65	0.10	1.30	0.05	1.50
	$p = 200,\ n = 100$				$p = 500,\ n = 100$			
(A) Identity	0.55	1.00	0.60	1.00	0.55	1.00	0.60	1.00
(B) MA(1)	0.25	1.00	0.20	1.00	0.30	1.00	0.25	1.00
(B*) MA(1)*	0.25	1.00	0.25	0.95	0.25	1.00	0.20	1.00
(C) AR(1)	0.05	1.00	0.05	1.00	0.10	1.10	0.05	0.80
(C*) AR(1)*	0.05	1.10	0.05	1.30	0.10	0.95	0.10	1.10
(D) Non-sparse	0.30	0.65	0.55	0.40	0.40	0.75	0.40	0.90
(E) Factor	0.10	−0.10	0.60	0.05	0.20	−0.10	0.50	0.05
(F) FGN	0.30	0.05	0.65	0.10	0.35	0.10	0.40	0.10
(F*) FGN*	0.25	0.05	0.50	0.05	0.15	−0.10	0.35	0.10
(G) Seasonal	0.10	1.10	0.05	1.50	0.10	1.30	0.10	1.20
(G*) Seasonal*	0.10	1.10	0.05	1.50	0.10	1.30	0.10	1.20

Comparison with competing estimators For the estimators with the optimal parameters (Tables 2, 3), NOVELIST performs the best for $p = 10$ for both Σ and Σ^{-1} and beats the competitors across the non-sparse and highly non-sparse model classes when $p = 100$, 200 and 500. The banding estimator beats NOVELIST in covariance matrix estimation in the homoscedastic sparse models by a small margin in the higher-dimensional cases. For the identity matrix, banding, thresholding and the optimal NOVELIST attain the same results. Optimal PC-adjusted NOVELIST achieves better relative results for estimating Σ^{-1} than for Σ.

In the competitions based on the data-driven estimators (Tables 4, 5), when $p = 10$, the cross-validation NOVELIST is the best for most of the models with heteroscedastic variances and only slightly worse than linear or nonlinear shrinkage estimator for the other models. When $p = 100$, 200 or 500, the cross-validation NOVELIST is the best for most of the models in the sparse and the non-sparse groups (more so for heteroscedastic models) for both Σ and Σ^{-1}, but is beaten by POET for the factor model

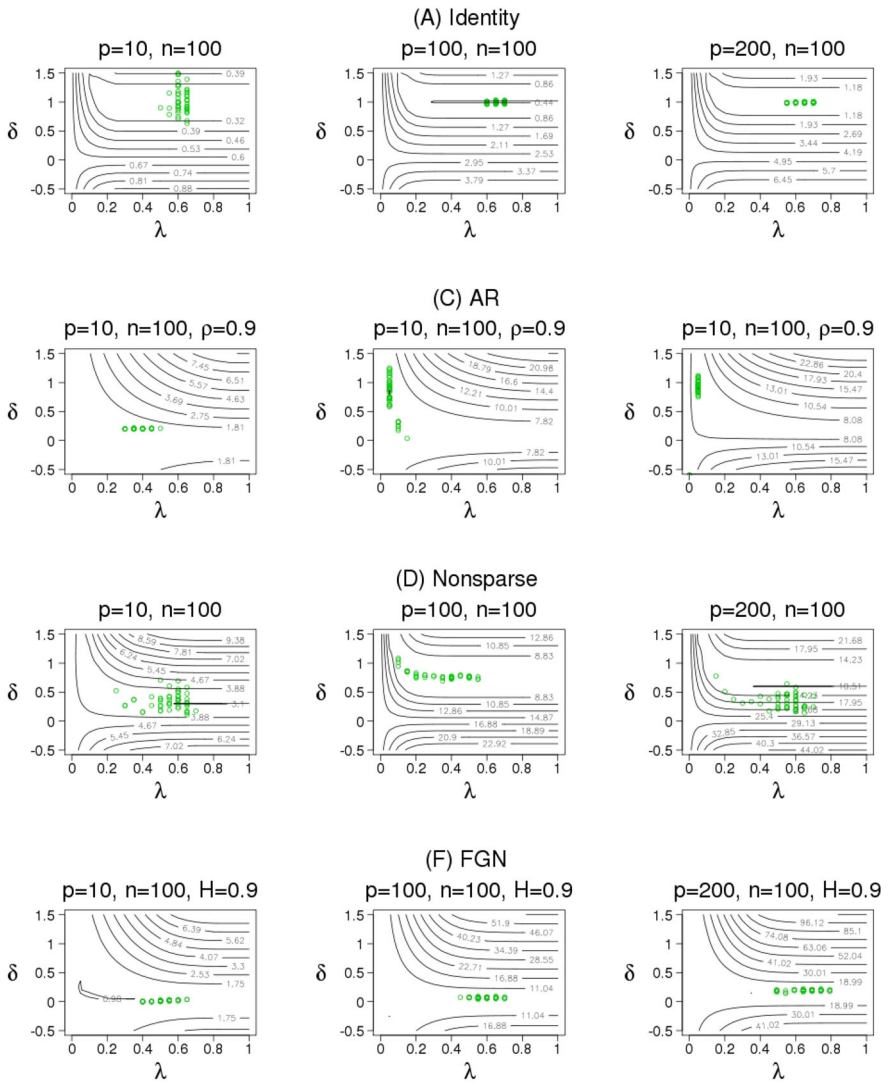


Fig 4 50 replicated cross-validation choices of (δ', λ') (green circles) against the background of contour lines of operator norm distances to Σ under models (**a**), (**c**), (**d**) and (**f**) [equivalent to Figs. 2 and 3], $n = 100$, $p = 10$ (Left), 100 (Middle), 200 (Right). The area inside the first contour line contains all combinations of (λ, δ) for which $\|\hat{\Sigma}^N(\lambda, \delta) - \Sigma\|$ is in the 1st decile of $[\min_{(\lambda,\delta)} \|\hat{\Sigma}^N(\lambda, \delta) - \Sigma\|, \max_{(\lambda,\delta)} \|\hat{\Sigma}^N(\lambda, \delta) - \Sigma\|]$

and the FGN model by a small margin, and is slightly worse than nonlinear shrinkage for homoscedastic sparse models. However, POET underperforms for the sparse and non-sparse models for Σ, and nonlinear shrinkage does worse than NOVELIST for heteroscedastic sparse models. The cases where the cross-validation NOVELIST performs the worst are rare. NOVELIST with fixed parameters as in Flowchart 1 for highly non-sparse cases improves the results for Σ^{-1}. PC-adjusted NOVELIST can

Table 2 Average operator norm error to Σ for competing estimators with optimal parameters (50 replications)

	$\hat{\Sigma}$	T_s	B	$\hat{\Sigma}^N_{\text{opt}}$	$\hat{\Sigma}^N_{\text{opt.rem}}$	$\hat{\Sigma}$	T_s	B	$\hat{\Sigma}^N_{\text{opt}}$	$\hat{\Sigma}^N_{\text{opt.rem}}$
	$p = 10,\ n = 100$					$p = 100,\ n = 100$				
(A) Identity	**0.578**	0.246	0.246	0.246	–	**2.946**	0.436	0.436	0.436	–
(B) MA(1)	**0.623**	0.447	0.361	0.435	–	**3.055**	0.670	0.554	0.668	–
(B*) MA(1)*	**1.400**	1.008	0.871	0.988	–	**6.458**	1.890	1.370	1.800	–
(C) AR(1)	**1.148**	0.762	1.072	0.475	–	**6.112**	4.977	3.999	4.703	–
(C*) AR(1)*	**2.010**	1.707	2.004	1.020	–	**16.338**	8.353	8.786	7.992	–
(D) Non-sparse	**3.483**	2.954	3.127	2.812	–	**25.844**	11.302	11.539	10.717	–
(E) Factor	**1.811**	1.462	1.742	1.120	1.221	**14.350**	13.675	13.993	9.881	9.921
(F) FGN	**1.110**	0.751	0.970	0.527	0.711	**7.824**	6.777	7.478	5.135	7.033
(F*) FGN*	**2.239**	1.617	2.108	1.129	1.683	**15.666**	13.383	15.147	10.878	13.782
(G) Seasonal	**0.850**	0.564	0.797	0.527	–	**4.290**	2.493	2.205	2.460	–
(G*) Seasonal*	**1.664**	1.228	1.594	1.158	–	**6.694**	3.028	2.362	2.959	–

Table 2 continued

	$\hat{\Sigma}$	T_s	B	$\hat{\Sigma}^N_{\text{opt}}$	$\hat{\Sigma}^N_{\text{opt.rem}}$	$\hat{\Sigma}$	T_s	B	$\hat{\Sigma}^N_{\text{opt}}$	$\hat{\Sigma}^N_{\text{opt.rem}}$
	$p = 200,\ n = 100$					$p = 500,\ n = 100$				
(A) Identity	**4.661**	0.440	0.440	0.440	–	**9.321**	0.467	0.467	0.467	–
(B) MA(1)	**4.886**	0.717	0.626	0.716	–	**9.828**	0.761	0.729	0.761	–
(B*) MA(1)*	**10.727**	1.884	1.545	1.881	–	**21.233**	2.041	1.775	2.041	–
(C) AR(1)	**10.291**	6.922	4.898	6.768	–	**17.877**	9.311	5.584	9.261	–
(C*) AR(1)*	**20.277**	14.691	14.943	14.426	–	**39.241**	18.780	11.738	18.728	–
(D) Non-sparse	**26.729**	10.990	11.240	10.322	–	**50.915**	13.917	13.284	12.913	–
(E) Factor	**31.183**	28.053	29.819	20.463	20.432	**82.451**	65.234	73.807	48.104	48.928
(F) FGN	**14.732**	12.729	13.877	9.906	15.881	**35.041**	30.201	31.272	23.939	30.782
(F*) FGN*	**32.370**	26.692	29.862	20.357	28.983	**68.154**	66.833	66.320	49.853	55.998
(G) Seasonal	**6.913**	2.961	2.418	2.930	–	**13.157**	3.582	2.499	3.460	–
(G*) Seasonal*	**14.709**	6.427	5.171	6.350	–	**27.627**	7.873	5.660	7.538	–

The results of $\hat{\Sigma}^N_{\text{opt.rem}}$ are only presented for the highly non-sparse group, i.e. models (E), (F) and (F*). The best results and those up to 5% worse than the best are boxed. The worst results are in bold

further improve the results for estimating Σ^{-1} but not for Σ. We would argue that NOVELIST is the overall best performer, followed by nonlinear shrinkage, linear shrinkage and POET.

7 Portfolio selection

In this section, we apply the NOVELIST algorithm and the competing methods to share portfolios composed of the constituents of the FTSE 100 index. Similar competitions were previously conducted to compare the performance of different covariance matrix estimators (Ledoit and Wolf 2003; Lam 2016). We compare the performance for risk minimisation purposes. The data were provided by Bloomberg.

Daily returns Our first dataset consists of $p = 85$ stocks of FTSE 100 (we removed all those constituents that contained missing values) and 2606 daily returns $\{r_t\}$ for the period 1 January 2005 to 31 December 2015. We use data from the first $n = 120$ days to estimate the initial covariance matrices of the returns based on six different competing covariance matrix estimators and create six portfolios with weights given by the well-known weight formula

$$\hat{w}_t = \frac{\left\{\hat{\Sigma}_t^{(120)}\right\}^{-1}\mathbf{1}_p}{\mathbf{1}_p^{\mathrm{T}}\left\{\hat{\Sigma}_t^{(120)}\right\}^{-1}\mathbf{1}_p}, \tag{26}$$

where $\hat{\Sigma}_t^{(120)}$ is an estimator of the $p \times p$ covariance matrix of the past 120-trading-day returns on trading day t (i.e. computed over days $t - 119$ to t) and $\mathbf{1}_p$ is the column vector of p ones. We hold these portfolios for the next 22 trading days and compute their out-of-sample standard deviations as (Ledoit and Wolf 2003)

$$\mathrm{STD} = \left\{\hat{w}_t \frac{1}{22}\sum_{i=1}^{22} r_{t+i} r_{t+i}^T \hat{w}_t^{\mathrm{T}}\right\}^{1/2}, \tag{27}$$

which is a measure of risk. On the 23rd day, we liquidate the portfolios and start the process all over again based on the past 120 trading days. The dataset is composed of 113 instances of such 22-trading-day blocks, and the average STD of each portfolio is computed.

5-min returns The second dataset consists of $p = 100$ constituents of FTSE 100 and 13,770 5-min returns $\{y_t\}$ for the period 2 March 2015 to 4 September 2015 (135 trading days). The procedure is similar to the one above, and only the differences are explained here. We use the first 2 days ($n = 204$) to estimate the initial covariance matrices of the returns and create portfolios with weights given by

$$\hat{y}_t = \frac{\left\{\hat{\Sigma}_t^{(204)}\right\}^{-1}\mathbf{1}_p}{\mathbf{1}_p^{\mathrm{T}}\left\{\hat{\Sigma}_t^{(204)}\right\}^{-1}\mathbf{1}_p}, \tag{28}$$

Table 3 Average operator norm error to Σ^{-1} for competing estimators with optimal parameters (50 replications)

	$\hat{\Sigma}$	T_s	B	$\hat{\Sigma}^N_{\text{opt}}$	$\hat{\Sigma}^N_{\text{opt.rem}}$	$\hat{\Sigma}$	T_s	B	$\hat{\Sigma}^N_{\text{opt}}$	$\hat{\Sigma}^N_{\text{opt.rem}}$
	$p = 10,\ n = 100$					$p = 100,\ n = 100$				
(A) Identity	**0.917**	0.281	0.281	0.281	–	–	0.469	0.469	0.469	–
(B) MA(1)	**1.177**	0.681	0.656	0.605	–	–	1.244	**1.300**	1.166	–
(B*) MA(1)*	0.626	0.489	**0.732**	0.442	–	–	**0.846**	0.779	0.745	–
(C) AR(1)	**9.078**	7.751	**9.078**	5.502	–	–	14.313	**18.064**	10.792	–
(C*) AR(1)*	**4.491**	2.736	**4.491**	2.339	–	–	**8.915**	7.298	6.001	–
(D) Non-sparse	**0.378**	0.256	0.297	0.210	–	–	2.670	**2.775**	1.793	–
(E) Factor	**0.846**	0.403	0.610	0.370	0.400	–	0.712	**0.715**	0.653	0.518
(F) FGN	**2.995**	1.727	2.980	1.560	1.535	–	3.585	**4.650**	3.112	2.734
(F*) FGN*	**1.571**	1.193	1.212	1.001	1.018	–	2.029	**2.038**	1.948	1.761
(G) Seasonal	**2.688**	1.538	2.685	1.302	–	–	3.806	**5.444**	3.260	–
(G*) Seasonal*	1.340	1.091	**1.726**	0.827	–	–	2.526	**4.345**	1.971	–

Table 3 continued

	$\hat{\Sigma}$	T_S	B	$\hat{\Sigma}^N_{\text{opt}}$	$\hat{\Sigma}^N_{\text{opt.rem}}$	$\hat{\Sigma}$	T_S	B	$\hat{\Sigma}^N_{\text{opt}}$	$\hat{\Sigma}^N_{\text{opt.rem}}$
	$p = 200,\ n = 100$					$p = 500,\ n = 100$				
(A) Identity	–	0.527	0.527	0.527	–	–	0.599	0.599	0.599	–
(B) MA(1)	–	1.358	**1.530**	1.258	–	–	1.405	**1.562**	1.377	–
(B*) MA(1)*	–	**1.100**	0.795	0.850	–	–	1.040	**1.145**	0.962	–
(C) AR(1)	–	15.023	**18.122**	11.469	–	–	15.622	**18.136**	11.064	–
(C*) AR(1)*	–	14.509	**20.358**	7.362	–	–	18.392	**23.740**	7.155	–
(D) Non-sparse	–	**2.460**	2.016	1.459	–	–	**5.986**	5.896	4.289	–
(E) Factor	–	**0.711**	**0.711**	0.677	0.537	–	**0.744**	**0.744**	0.730	0.557
(F) FGN	–	3.972	**4.658**	3.317	3.024	–	4.267	**4.737**	3.527	3.306
(F*) FGN*	–	2.974	**4.096**	2.083	1.849	–	4.426	**5.674**	2.250	2.083
(G) Seasonal	–	4.029	**5.469**	3.538	–	–	4.188	**5.477**	3.673	–
(G*) Seasonal*	–	3.328	**4.885**	2.259	–	–	3.726	**5.479**	2.358	–

The results of $\hat{\Sigma}^N_{\text{opt.rem}}$ are only presented for the highly non-sparse group, i.e. models (E), (F) and (F*). The worst results for model (A) with $p = 100$, 200 and 500 are not labelled, as T, B and $\hat{\Sigma}^N_{\text{opt}}$ obtain exactly the same results. The best results and those up to 5% worse than the best are boxed. The worst results are in bold

Table 4 Average operator norm error to Σ for competing estimators with data-driven parameters (50 replications)

	$\hat{\Sigma}$	S	P	$\hat{\Sigma}^N_{cv}$	$\hat{\Sigma}^N_{\text{rem}}$	NS	$\hat{\Sigma}$	S	P	$\hat{\Sigma}^N_{cv}$	$\hat{\Sigma}^N_{\text{rem}}$	NS
	$p = 10,\ n = 100$						$p = 100,\ n = 100$					
(A) Identity	0.578	0.084	**0.823**	0.263	–	0.116	2.946	0.088	**3.657**	0.446	–	0.087
(B) MA(1)	0.623	0.444	**0.732**	0.493	–	0.481	3.055	0.670	**3.730**	0.704	–	0.694
(B*) MA(1)*	1.400	1.165	**1.546**	1.159	–	1.191	6.458	1.985	**8.015**	1.877	–	2.449
(C) AR(1)	1.148	1.013	1.135	**1.153**	–	1.017	6.112	5.423	**6.257**	5.390	–	5.892
(C*) AR(1)*	2.010	2.190	**2.291**	2.114	–	2.190	16.338	8.878	**19.468**	8.446	–	12.095
(D) Non-sparse	3.483	3.120	**3.860**	3.046	–	2.934	25.844	12.453	**29.355**	11.739	–	11.730
(E) Factor	1.811	1.793	**1.866**	1.741	1.763	1.537	14.350	**17.681**	14.304	16.497	16.438	15.285
(F) FGN	**1.110**	0.849	1.020	1.021	1.024	0.980	**7.824**	6.628	7.798	7.799	7.732	7.554
(F*) FGN*	**2.239**	2.218	2.221	2.222	2.227	1.960	15.666	14.795	15.611	15.225	15.254	**16.561**
(G) Seasonal	0.850	0.666	**0.852**	0.687	–	0.659	4.290	3.200	**4.826**	2.534	–	3.098
(G*) Seasonal*	**1.664**	1.647	1.652	1.452	–	1.480	6.694	4.268	**7.171**	3.016	–	6.979

Table 4 continued

	$\hat{\Sigma}$	S	P	$\hat{\Sigma}^N_{cv}$	$\hat{\Sigma}^N_{\text{rem}}$	NS	$\hat{\Sigma}$	S	P	$\hat{\Sigma}^N_{cv}$	$\hat{\Sigma}^N_{\text{rem}}$	NS
	$p = 200,\ n = 100$						$p = 500,\ n = 100$					
(A) Identity	4.661	0.058	**5.414**	0.443	–	0.067	9.321	0.064	**10.076**	0.468	–	0.047
(B) MA(1)	4.886	0.658	**5.615**	0.744	–	0.694	9.828	0.645	**10.566**	0.819	–	0.683
(B*) MA(1)*	10.727	2.094	**12.458**	1.956	–	2.729	21.233	2.060	**23.034**	2.116	–	3.004
(C) AR(1)	10.291	8.123	**11.446**	8.217	–	7.759	17.877	12.785	**18.496**	12.484	–	12.036
(C*) AR(1)*	20.277	18.172	**23.721**	16.251	–	18.751	39.241	26.571	**40.903**	18.903	–	24.581
(D) Non-sparse	26.729	11.920	**30.108**	11.220	–	10.993	50.915	13.758	**54.462**	13.636	–	12.996
(E) Factor	31.183	**34.237**	31.064	33.224	33.194	31.020	82.451	**83.101**	81.489	81.697	81.382	80.852
(F) FGN	**14.732**	12.961	14.376	14.640	14.593	14.125	35.041	26.672	34.344	31.296	30.992	**36.299**
(F*) FGN*	**32.370**	31.165	30.263	31.470	31.042	32.188	68.154	**84.958**	69.133	75.546	75.377	74.432
(G) Seasonal	**6.913**	4.126	7.403	2.972	–	4.016	13.157	4.994	**13.722**	3.471	–	4.949
(G*) Seasonal*	14.709	9.225	**15.855**	6.494	–	9.064	27.627	11.030	**28.949**	7.561	–	11.132

For $\hat{\Sigma}^N_{\text{rem}}$, (λ'', δ'') is fixed to be (0.10, 0.30) in (E), and (0.30, 0.50) in (F) and (F*). The best results and those up to 5% worse than the best are boxed. The worst results are in bold

Table 5 Average operator norm error to Σ^{-1} for competing estimators with data-driven parameters (50 replications)

	$\hat{\Sigma}$	S	P	$\hat{\Sigma}^N_{cv}$	$\hat{\Sigma}^N_{\text{rem}}$	NS	$\hat{\Sigma}$	S	P	$\hat{\Sigma}^N_{cv}$	$\hat{\Sigma}^N_{\text{rem}}$	NS
	$p = 10,\ n = 100$						$p = 100,\ n = 100$					
(A) Identity	0.917	0.090	**4.472**	0.469	–	0.146	–	0.045	**0.882**	0.472	–	0.109
(B) MA(1)	1.123	0.799	**6.474**	0.824	–	0.780	–	1.273	1.403	**1.439**	–	1.405
(B*) MA(1)*	0.626	0.526	**4.892**	0.448	–	0.440	–	1.358	0.993	0.935	–	**1.748**
(C) AR(1)	9.078	7.309	**40.142**	8.574	–	5.396	–	13.410	**15.704**	12.605	–	12.272
(C*) AR(1)*	4.941	5.390	**27.593**	4.841	–	3.264	–	12.508	**13.649**	10.167	–	13.446
(D) Non-sparse	0.378	0.500	**1.705**	0.328	–	0.340	–	2.937	2.916	2.910	–	**2.979**
(E) Factor	0.846	1.142	**1.806**	0.864	–	0.296	–	**2.603**	0.893	1.608	–	0.343
					(0.854)					(0.695)	(0.526)	
(F) FGN	2.995	1.864	**16.530**	2.097	–	1.701	–	**4.565**	3.060	4.212	–	3.122
					(2.081)					(3.159)	(2.773)	
(F*) FGN*	1.571	1.174	**10.284**	2.017	–	1.101	–	**4.474**	2.965	3.431	–	4.432
					(2.001)					(2.075)	(1.843)	
(G) Seasonal	2.688	1.897	**13.175**	2.103	2.115	1.687	–	4.229	**4.721**	3.839	–	3.947
(G*) Seasonal*	1.340	1.284	**8.436**	1.143	–	1.219	–	3.510	3.799	2.743	–	**4.538**

Table 5 continued

	$\hat{\Sigma}$	S	P	$\hat{\Sigma}^N_{cv}$	$\hat{\Sigma}^N_{\text{rem}}$	NS	$\hat{\Sigma}$	S	P	$\hat{\Sigma}^N_{cv}$	$\hat{\Sigma}^N_{\text{rem}}$	NS
	$p = 200,\ n = 100$						$p = 500,\ n = 100$					
(A) Identity	–	0.046	**0.930**	0.529	–	0.136	–	0.078	**0.923**	0.601	–	0.139
(B) MA(1)	–	1.449	1.371	1.401	–	**1.463**	–	1.473	1.445	**1.540**	–	1.487
(B*) MA(1)*	–	1.293	1.256	1.169	–	**1.906**	–	1.914	1.140	1.221	–	**2.463**
(C) AR(1)	–	15.066	**17.128**	14.125	–	13.907	–	16.526	**17.700**	16.025	–	15.924
(C*) AR(1)*	–	17.480	18.286	13.201	–	**19.037**	–	22.833	23.053	19.169	–	**23.740**
(D) Non-sparse	–	2.602	2.842	2.563	–	**3.206**	–	5.998	**6.171**	5.994	–	5.660
(E) Factor	–	**3.701**	0.892	1.450	–	0.348	–	**5.672**	0.962	4.106	–	0.347
				(0.710)	(0.546)					(0.937)	(0.558)	
(F) FGN	–	**9.397**	3.552	5.670	–	3.434	–	**8.621**	3.933	6.652	–	3.752
				(3.582)	(3.045)					(4.364)	(3.326)	
(F*) FGN*	–	**6.649**	2.765	4.024	–	5.519	–	6.241	3.083	5.442	–	**6.519**
				(2.589)	(2.199)					(3.002)	(2.887)	
(G) Seasonal	–	4.676	**5.019**	4.176	–	4.526	–	5.045	**5.256**	4.548	–	5.001
(G*) Seasonal*	–	4.540	4.643	3.514	–	**6.068**	–	5.632	5.254	4.489	–	**6.988**

For models (E), (F) and (F*), results by both cross-validation and fixed parameters (in brackets) are presented for NOVELIST when $n < 2p$. For $\hat{\Sigma}^N_{cv}$, fixed parameters (λ'', δ'') are (0.90, 0.50) for model (E), and (0.50, 0.25) for models (F) and (F*). For $\hat{\Sigma}^N_{\text{rem}}$, (λ'', δ'') is fixed to be (0.50, 0.90) for (E), and (0.25, 0.65) for (F) and (F*). The best results and those up to 5% worse than the best are boxed. The worst results are in bold

Table 6 Standard deviation of minimum variance portfolios in percentage (daily and 5-min returns)

	STD (daily returns)	STD (5-min returns)
Sample	1.256	10.675
Linear shrinkage	0.851	7.809
Nonlinear shrinkage	0.733	7.670
POET	0.760	7.253
NOVELIST	0.709	6.987
PC-adjusted NOVELIST	0.715	8.577

where $\hat{\Sigma}_t^{(204)}$ is an estimator of the $p \times p$ covariance matrix of the 5-min returns over the past 204 data points (2 days) at trading time t. We hold them for the next day and the out-of-sample standard deviations are calculated by

$$\text{STD} = \left\{ \hat{w}_t \frac{1}{102} \sum_{i=1}^{102} r_{t+i} r_{t+i}^{\mathrm{T}} \hat{w}_t^{\mathrm{T}} \right\}^{1/2}. \tag{29}$$

We rebalance the portfolios every day and compute the sum of out-of-sample STD's over the 133 trading days.

Following the advice from Sect. 5.1, we apply fixed parameters for both NOVELIST and PC-adjusted NOVELIST. Table 6 shows the results. NOVELIST has the lowest risk for both daily and 5-min portfolios, followed by PC-adjusted NOVELIST and nonlinear shrinkage in the low-frequency case and by POET and nonlinear shrinkage in the high-frequency case. In summary, NOVELIST offers the best option in terms of risk minimisation.

8 Discussion

As many other covariance (correlation) matrix estimators which incorporate thresholding, the NOVELIST estimator is not guaranteed to be positive definite in finite samples. To remedy this, our advice is similar to other authors' (e.g. Cai et al. 2010; Fan et al. 2013; Bickel and Levina 2008b): we propose to diagonalise the NOVELIST estimator and replace any eigenvalues that fall under a certain small positive threshold by the value of that threshold. How to choose the threshold is, of course, an important matter, and we do not believe there is a generally accepted solution in the literature, partly because the value of the "best" such threshold will necessarily be problem-dependent. Denoting the such corrected estimator by $\hat{\Sigma}^N(\zeta)$ (in the covariance case) and $\hat{R}^N(\zeta)$ (in the correlation case), where ζ is the eigenvalue threshold, one possibility would be to choose the lowest possible ζ for which the matrix $\hat{\Sigma}^N(\hat{\Sigma}^N(\zeta))^{-1}$ (and analogously for the correlation case) resembles the identity matrix, in a certain user-specified sense.

We also note that either part of the NOVELIST estimator can be replaced by a banding-type estimator, for example, as defined by Cai et al. (2010). In this way, we

would depart from the particular construction of the NOVELIST estimator towards the more general idea of using convex combinations of two (or more) covariance estimators, which is conceptually and practically appealing but lies outside the scope of the current work.

To summarise, the flexible control of the degree of shrinkage and thresholding offered by NOVELIST means that it is able to offer competitive performance across most models, and in situations in which it is not the best, it tends not to be much worse than the best performer. We recommend NOVELIST as a simple, good all-round covariance, correlation and precision matrix estimator ready for practical use across a variety of models and data dimensionalities.

Acknowledgements Piotr Fryzlewicz's work has been supported by the Engineering and Physical Sciences Research Council Grant No. EP/L014246/1.

9 Appendix

9.1 Additional lemmas and proofs

Firstly, we briefly introduce two lemmas that will be used in the proof of Proposition 1.

Lemma 1 *If F satisfies $\int_0^\infty exp(\gamma t) dG_j(t) < \infty$, for $0 < |\gamma| < \gamma_0$, for some $\gamma_0 > 0$, where G_j is the cdf of X_{1j}^2, $R = \{\rho_{ij}\}$ and $\Sigma = \{\sigma_{ij}\}$ are the true correlation and covariance matrices, $1 \le i, j \le p$, and $\sigma_{ii} \le M$, where M is a constant, then, for sufficiently large M', if $\lambda = M'\sqrt{\log p/n}$ and $\log p/n = o(1)$, we have $\max_{1\le i,j\le p} |\hat{\rho}_{ij} - \rho_{ij}| = O_p(\sqrt{\log p/n})$, for $1 \le i, j \le p$.*

Proof of Lemma 1 By the sub-multiplicative norm property $\|AB\| \le \|A\| \, \|B\|$ (Golub and Loan 1989), we write

$$
\begin{aligned}
&\max_{1\le i,j\le p} |\hat{\rho}_{ij} - \rho_{ij}| \\
&= \max_{1\le i,j\le p} \left| \hat{\sigma}_{ij}/(\hat{\sigma}_{ii}\hat{\sigma}_{jj})^{1/2} - \sigma_{ij}/(\sigma_{ii}\sigma_{jj})^{1/2} \right| \\
&\le \max_{1\le i\le p} \left| \hat{\sigma}_{ii}^{-1/2} - \sigma_{ii}^{-1/2} \right| \max_{1\le i,j\le p} \left| \hat{\sigma}_{ij} - \sigma_{ij} \right| \max_{1\le j\le p} \left| \hat{\sigma}_{jj}^{-1/2} - \sigma_{jj}^{-1/2} \right| \\
&\quad + \max_{1\le i\le p} \left| \hat{\sigma}_{ii}^{-1/2} - \sigma_{ii}^{-1/2} \right| \max_{1\le i,j\le p} \left(\left| \hat{\sigma}_{ij} \right| \left| \sigma_{jj}^{-1/2} \right| + \left| \hat{\sigma_{ii}}^{-1/2} \right| \left| \sigma_{ij} \right| \right) \\
&\quad + \max_{1\le i,j\le p} \left| \hat{\sigma}_{ij} - \sigma_{ij} \right| \max_{1\le i\le p} \left| \hat{\sigma_{ii}}^{-1/2} \right| \max_{1\le i\le p} \left| \sigma_{ii}^{-1/2} \right| \\
&= O_p(\sqrt{\log p/n}). \qquad (30)
\end{aligned}
$$

The last equality holds as we have $\max_{1\le i,j\le p}|\hat{\sigma}_{ij}-\sigma_{ij}| = O_p(\sqrt{\log p/n}) = \max_{1\le i,j\le p}|\hat{\sigma}_{ij}^{-1}-\sigma_{ij}^{-1}|$ (Bickel and Levina 2008b) and $\max_{1\le i,j\le p}|\hat{\sigma}_{ij}| = O_p(\sqrt{\log p/n}) = \max_{1\le i,j\le p}|\hat{\sigma}_{ij}^{-1}|$, and $\sigma_{ii}\le M$, $1\le i,j\le p$. □

Lemma 2 *If F satisfies $\int_0^\infty exp(\gamma t)dG_j(t)<\infty$, for $0<|\gamma|<\gamma_0$, for some $\gamma_0>0$, where G_j is the cdf of X_{1j}^2, $R=\{\rho_{ij}\}$ is the true correlation matrix, $1\le i,j\le p$, then, uniformly on $\mathcal{V}(q,s_0(p),\varepsilon_0)$, for sufficiently large M', if $\lambda=M'\sqrt{\log p/n}$ and* $\log p/n=o(1)$,

$$\|T(\hat{R},\lambda)-R\| = O_p(s_0(p)(\log p/n)^{(1-q)/2}), \tag{31}$$

where T is any kind of generalised thresholding estimator.

Lemma 2 is a correlation version of Theorem 1 in Rothman et al. (2009) and follows in a straightforward way by replacing $\hat{\Sigma}$, Σ, $\mathcal{U}(q,c_0(p),M,\epsilon_0)$ and $c_0(p)$ by $\hat{R}$, R, $\mathcal{V}(q,s_0(p),\varepsilon_0)$ and $s_0(p)$ in the proof of the theorem.

Proof of Proposition 1 We first show the result for $\hat{R}^N$. By the triangle inequality,

$$\begin{aligned}\|\hat{R}^N-R\| &= \|(1-\delta)\hat{R}+\delta T(\hat{R},\lambda)-R\|\\ &\le (1-\delta)\|\hat{R}-R\|+\delta\|T(\hat{R},\lambda)-R\|\\ &= I+II.\end{aligned} \tag{32}$$

Using Lemma 2, we have

$$II = O_p\left\{\delta s_0(p)(\log p/n)^{(1-q)/2}\right\}. \tag{33}$$

For symmetric matrices M, Corollary 2.3.2 in Golub and Loan (1989) states that

$$\|M\|\le(\|M\|_{(1,1)}\|M\|_{(\infty,\infty)})^{1/2} = \|M\|_{(1,1)} = \max_{1\le i\le p}\sum_{j=1}^{p}|m_{ij}|. \tag{34}$$

Then by Lemma 1,

$$\|\hat{R}-R\|\le\max_{1\le i\le p}\sum_{j=1}^{p}|\hat{R}_{ij}-R_{ij}|\le p\max_{1\le i,j\le p}|\hat{\rho}_{ij}-\rho_{ij}| = O_p(p\sqrt{\log p/n}). \tag{35}$$

Thus, we have

$$I=(1-\delta)\|\hat{R}-R\|\le O_p((1-\delta)p\sqrt{\log p/n}). \tag{36}$$

Combining formulae (33) and (36) yields the first equality. The second equality follows because

$$\|(\hat{R}^N)^{-1} - R^{-1}\| \asymp \|\hat{R}^N - R\| \tag{37}$$

uniformly on $\mathcal{V}(q, s_0(p), \varepsilon_0)$.

For the $\hat{\Sigma}^N$ estimator, recalling that $T = T(\hat{R}, \lambda)$ and $D = (\mathrm{diag}(\Sigma))^{1/2}$, we have

$$\begin{aligned}\|\hat{\Sigma}^N - \Sigma\| &= \|\hat{D}\hat{R}^N\hat{D} - DRD\| \\ &= \|\hat{D}((1-\delta)\hat{R} + \delta\, T)\hat{D} - DRD\| \\ &\le (1-\delta)\|\hat{\Sigma} - \Sigma\| + \delta\|\hat{D}T\hat{D} - DRD\| \\ &= III + IV. \end{aligned} \tag{38}$$

Similarly as in 36, we obtain $III = O_p((1-\delta)p\sqrt{\log p/n})$. For IV, we write

$$\begin{aligned}&\|\hat{D}T\hat{D} - DRD\| \\ &\quad\le \|\hat{D} - D\|\,\|T - R\|\,\|\hat{D} - D\| + \|\hat{D} - D\|(\|T\|\,\|D\| + \|\hat{D}\|\,\|R\|) \\ &\qquad + \|T - R\|\,\|\hat{D}\|\,\|D\| \\ &\quad= O_p((1 + s_0(p)(\log p/n)^{-q/2})\sqrt{\log p/n}). \end{aligned} \tag{39}$$

The last equality holds as we have $\|T - R\| = O_p(s_0(p)(\log p/n)^{(1-q)/2})$, $\|\hat{D} - D\| = O_p(\sqrt{\log p/n})$, $\|\hat{D}\| = O_p(1) = \|T\|$, and $\|D\| = O(1)$ as $\sigma_{ii} < M$. Because $(\log p/n)^{q/2}(s_0(p))^{-1}$ is bounded from above by the assumption that $\log p/n = o(1)$ and $\|(\hat{\Sigma}^N)^{-1} - \Sigma^{-1}\| \asymp \|\hat{\Sigma}^N - \Sigma\|$ uniformly on $\mathcal{V}(q, s_0(p), \varepsilon_0)$, the result follows. □

Proof of Proposition 2 We only need to show the rate for the sample covariance (correlation) part as the arguments for the thresholding part are identical to those in Proposition 1. We first collect the relevant arguments from the proof of Lemma 3 in Cai et al. (2010). Let $\|\cdot\|$ denote the spectral norm of a matrix. From the proof of Lemma 3 in Cai et al. (2010), there exist vectors $v_1, v_2, \ldots, v_{5^m} \in S^{m-1}$, where S^{m-1} is the unit sphere in the Euclidean distance in $\mathbb{R}^m$, such that

$$\|A\| \le 4 \sup_{j \le 5^m} \left| v_j^{\mathrm{T}} A\, v_j \right|$$

for all $m \times m$ symmetric matrices A.

Consider now the sample covariance matrix $\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i \boldsymbol{X}_i^{\mathrm{T}}$ (recall that $E(\boldsymbol{X}) = 0$), satisfying a sub-Gaussian condition in the sense that the length-p column vector $\boldsymbol{X}_i$ satisfies

$$P\left(\left|v^{\mathrm{T}}\boldsymbol{X}_i\right| > t\right) \le \exp\left(-t^2\rho/2\right)$$

for a certain $\rho > 0$, for all $t > 0$ and $\|v\|_2 = 1$.

Then, by the same arguments as in the proof of Lemma 3 in Cai et al. (2010), there exists $\rho_1 > 0$ such that

$$P\left\{\left|\frac{1}{n}\sum_{i=1}^{n} v^{\mathrm{T}}\left(\boldsymbol{X}_i\boldsymbol{X}_i^{\mathrm{T}} - \Sigma\right)v\right| > x\right\} \leq \exp(-nx^2\rho_1/2),$$

where Σ is the population covariance matrix, for all $0 < x < \rho_1$ and $\|v\| = 1$.

We then bound

$$\begin{aligned} P(\|\hat{\Sigma} - \Sigma\| > x) &\leq P\left(4 \sup_{j \leq 5^p} |v_j^{\mathrm{T}}(\hat{\Sigma} - \Sigma)v_j| > x\right) \\ &\leq 5^p \sup_{v_j} P\left(|v_j^{\mathrm{T}}\left(\hat{\Sigma} - \Sigma\right)v_j| > y\right) \\ &\leq 5^p \exp\left(-ny^2\rho_1/2\right) \\ &= \exp\left(p\log 5 - ny^2\rho_1/2\right), \end{aligned}$$

with $y = x/4$.

As ρ_1 is unknown, the only "safe" y's to consider are such that $y \to 0$ as $n \to \infty$, uniformly over all permitted p. We now want

$$\exp(p\log 5 - ny^2\rho_1/2) \leq \frac{C}{n} = \exp(\log C - \log n),$$

which leads to

$$y \geq \sqrt{\frac{2(p\log 5 + \log n - \log C)}{n\rho_1}}.$$

This can only converge to zero if $p = o(n)$. Under this assumption, we therefore indeed have

$$\|\hat{\Sigma} - \Sigma\| = O_P\left(\sqrt{\frac{p + \log n}{n}}\right),$$

which completes the proof. □

References

Alvarez I, Niemi J, Simpson M (2014) Bayesian inference for a covariance matrix. Preprint

Bickel P, Levina E (2008a) Regularized estimation of large covariance matrices. Ann Stat 36:199–227

Bickel P, Levina E (2008b) Covariance regularization by thresholding. Ann Stat 36:2577–2604

Cai T, Liu W (2011) Adaptive thresholding for sparse covariance matrix estimation. J Am Stat Assoc 106:672–684
Cai TT, Zhang C, Zhou HH (2010) Optimal rates of convergence for covariance matrix estimation. Ann Stat 38:2118–2144
Chen C (1979) Bayesian inference for a normal dispersion matrix and its application to stochastic multiple regression analysis. J R Stat Soc Ser B 41:235–248
Croux C, Haesbroeck G (2000) Principal component analysis based on robust estimators of the covariance or correlation matrix: influence functions and efficiencies. Biometrika 87:603–618
Dickey JM, Lindley DV, Press SJ (1985) Bayesian estimation of the dispersion matrix of a multivariate normal distribution. Commun Stat Theory Methods 14:1019–1034
El Karoui N (2008) Operator norm consistent estimation of large-dimensional sparse covariance matrices. Ann Stat 36:2717–2756
Evans IG (1965) Bayesian estimation of parameters of a multivariate normal distribution. J R Stat Soc Ser B 27:279–283
Fan J, Li R (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. J Am Stat Assoc 96:1348–1360
Fan J, Fan Y, Lv J (2008) High dimensional covariance matrix estimation using a factor model. J Econ 147:186–197
Fan J, Liao Y, Mincheva M (2013) Large covariance estimation by thresholding principal orthogonal complements. J R Stat Soc Ser B 75:603–680
Fisher RA (1936) The use of multiple measurements in taxonomic problems. Ann Eugen 7:179–188
Friedman J, Hastie T, Tibshirani R (2008) Sparse inverse covariance estimation with the graphical lasso. Biostatistics 9:432–441
Fryzlewicz P (2013) High-dimensional volatility matrix estimation via wavelets and thresholding. Biometrika 100:921–938
Furrer R, Bengtsson T (2007) Estimation of high-dimensional prior and posteriori covariance matrices in Kalman filter variants. J Multivar Anal 98:227–255
Gardner TS, di Bernardo D, Lorenz D, Collins JJ (2003) Inferring genetic networks and identifying compound mode of action via expression profiling. Science 301:102–105
Golub GH, Van Loan CF (1989) Matrix computations, 2nd edn. Johns Hopkins University Press, Baltimore
Guo YQ, Hastie T, Tibshirani R (2007) Regularized linear discriminant analysis and its application in microarrays. Biostatistics 8:86–100
Jeong H, Mason SP, Barabási A-L, Oltvai ZN (2001) Lethality and centrality in protein networks. Nature 411:41–42
Lam C (2016) Nonparametric eigenvalue-regularized precision or covariance matrix estimation. Ann Stat 44:928–953
Lam C, Feng P (2017) Integrating regularized covariance matrix estimators. Preprint
Ledoit O, Péché S (2011) Eigenvectors of some large sample covariance matrix ensembles. Probab Theory Relat Fields 151:233–264
Ledoit O, Wolf M (2003) Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. J Empir Finance 10:603–621
Ledoit O, Wolf M (2004) A well-conditioned estimator for large-dimensional covariance matrices. J Multivar Anal 88:365–411
Ledoit O, Wolf M (2012) Nonlinear shrinkage and estimation of large-dimensional covariance matrices. Ann Stat 4:1024–1060
Ledoit O, Wolf M (2015) Spectrum estimation: a unified framework for covariance matrix estimation and PCA in large dimensions. J Multivar Anal 139:360–384
Leonard T, John SJH (2012) Bayesian inference for a covariance matrix. Ann Stat 20:1669–1696
Longerstaey J, Zangari A, Howard S (1996) Risk metrics^{TM}-technical document. Technical document. J.P. Morgan, New York
Markowitz H (1952) Portfolio selection. J Finance 7:77–91
Meinshausen N, Bühlmann P (2006) High-dimensional graphs and variable selection with the lasso. Ann Stat 34:1436–1462
Pearson K (1901) On lines and planes of closest fit to systems of points in space. Philos Mag 2:559–572
Rothman AJ, Bickel P, Levina E, Zhu J (2008) Sparse permutation invariant covariance estimation. Electron J Stat 2:494–515

Rothman AJ, Levina E, Zhu J (2009) Generalized thresholding of large covariance matrices. J Am Stat Assoc 104:177–186

Savic RM, Karlsson MO (2009) Importance of shrinkage in empirical bayes estimates for diagnostics: problems and solutions. Am Assoc Pharm Sci 11:558–569

Schäfer J, Strimmer K (2005) A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomic. Stat Appl Genet Mol Biol 4:1544–6115

Wu WB, Pourahmadi M (2003) Nonparametric estimation in the gaussian graphical model. Biometrika 90:831–844

Zou H (2006) The adaptive lasso and its oracle properties. J Am Stat Assoc 101:1418–1429