INTERNATIONAL MONETARY FUND

High-Dimensional Covariance Matrix Estimation: Shrinkage Toward a Diagonal Target

Sakai Ando and Mingmei Xiao

WP/23/257

2023
DEC

WORKING PAPER

 WP/23/257

IMF Working Paper
Research Department

High-Dimensional Covariance Matrix Estimation: Shrinkage Toward a Diagonal Target

Prepared by Sakai Ando and Mingmei Xiao*

Authorized for distribution by Prachi Mishra
December 2023

ABSTRACT: This paper proposes a novel shrinkage estimator for high-dimensional covariance matrices by extending the Oracle Approximating Shrinkage (OAS) of Chen et al. (2009) to target the diagonal elements of the sample covariance matrix. We derive the closed-form solution of the shrinkage parameter and show by simulation that, when the diagonal elements of the true covariance matrix exhibit substantial variation, our method reduces the Mean Squared Error, compared with the OAS that targets an average variance. The improvement is larger when the true covariance matrix is sparser. Our method also reduces the Mean Squared Error for the inverse of the covariance matrix.

JEL Classification Numbers:	C13, C55
Keywords:	High-Dimension; Covariance Matrix; Shrinkage; Diagonal Target
Author's E-Mail Address:	sando@imf.org ; mx235@cam.ac.uk

Contents

1 Introduction

Estimating a covariance matrix $\Sigma : p \times p$ and its inverse when the dimension of the matrix p is larger than the sample size n is central to many empirical applications, including financial portfolio selection and macroeconomic forecasting ((DeMiguel et al. (2009), Ban et al. (2018), Ando and Kim (2022)), and econometric methods, such as Generalized Method of Moments (Hansen (1982)) and Principal Component Analysis (Pearson (1901)). Although Ledoit and Wolf (2004) developed a shrinkage estimator based on an average variance target, and Chen et al. (2009) improved its finite sample performance under the normality assumption, the method leaves room for improvement when the diagonal elements of the true covariance matrix exhibit substantial variation. For example, in the setting of macroeconomic forecasting, GDP and output of, say, the fishing industry can differ by a hundredfold, so the shrinkage estimator that targets the average variance can overestimate the variance of the fishing industry's output and underestimate that of GDP.

To accommodate the case where the variance of random variables exhibit substantial variation, this paper proposes a shrinkage estimator that targets the diagonal elements of the sample covariance matrix. Our method extends the Oracle Approximating Shrinkage estimator (OAS) of Chen et al. (2009) that targets the average variance. Following Eldar and Chernoi (2008) and Chen et al. (2009), we derive the optimal shrinkage parameter given the true covariance matrix (Oracle estimator) and approximate this infeasible Oracle estimator with an iterative algorithm.

We use a simulation to show that our method generates a lower Mean Squared Error (MSE) than OAS when the diagonal elements of the true covariance matrix exhibit substantial variation. In the specification of decaying off-diagonal elements, the degree of improvement is higher when the true covariance matrix is sparser. Our method also generates a smaller MSE for the inverse of the covariance matrix, which is often an ultimate goal of estimating a covariance matrix in practice.

As Chen et al. (2009), our method is based on the optimality under the normal distri-

bution. Compared to Schäfer and Strimmer (2005) which also target diagonal elements of the covariance matrix but without imposing a distributional assumption, our method performs better when the distribution is normal. In addition, our method inherits the desirable property of OAS that the shrinkage parameter stays between 0 and 1. Thus, the estimated covariance matrix is positive-definite, even without manually restricting the shrinkage parameter as done in Schäfer and Strimmer (2005). The normality assumption also allows us to derive the optimal shrinkage parameter in a closed form, which involves less computation than the non-linear shrinkage method of Ledoit and Wolf (2012).

Our method, however, does not outperform existing methods in all circumstances, and thus, should be considered a complement to them. For example, when the variation in the diagonal elements of the true covariance matrix is small, the OAS tends to generate a lower MSE. This observation also suggests an alternative method of estimating the covariance matrix by applying OAS to the correlation matrix and scaling it back by multiplying sample variances. To examine the robustness, we conduct a simulation and show that the difference in MSE between OAS and our proposed method is small and that directly shrinking the sample covariance matrix performs better than applying OAS to the correlation matrix and scaling it back.

This paper is organized as follows. Section 2 describes the theoretical framework, section 3 uses simulation to assess the performance and evaluate robustness, and section 4 concludes.

2 Theoretical Framework

Suppose that the data $\{x_i\}_{i=1}^n$ are *i.i.d.* and has p dimensions. In a high-dimensional environment $p > n \geq 2$, the sample covariance matrix

$$S := \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x}) (x_i - \bar{x})^T, \quad \bar{x} := \frac{1}{n} \sum_{i=1}^{n} x_i, \tag{1}$$

is degenerate and is a poor estimate of the true covariance matrix Σ. Throughout the paper, we assume that the diagonal elements of sample covariance matrices are positive $S_{mm} > 0$ for all $m = 1, ..., p$ and the true covariance matrix is positive definite $\Sigma > 0$.

One way to address the issue is to use a linear shrinkage estimator of the covariance matrix

$$\hat{S}(\rho) := (1 - \rho) S + \rho T, \tag{2}$$

where T is called a target matrix. We use the diagonal elements of the sample covariance matrix S as the target $T = diag(S)$, while OAS targets the average variance $T = \frac{tr(S)}{p} I$. In either case, as long as the target matrix T is positive definite and the shrinkage parameter resides in $\rho \in (0, 1]$, the estimated covariance matrix $\hat{S}(\rho)$ is positive definite even when the sample covariance matrix S is degenerate

$$a' \hat{S}(\rho) a = (1 - \rho) \underbrace{a' S a}_{\geq 0} + \rho \underbrace{a' T a}_{> 0} > 0, \quad \forall a \neq 0, \quad \rho \in (0, 1]. \tag{3}$$

When the true covariance matrix Σ is known, the shrinkage parameter ρ can be pinned down by minimizing the MSE from the true covariance matrix

$$\rho_{OD}(\Sigma, T) := \arg\min_{\rho \in \mathbb{R}} E\left[\left\| \hat{S}(\rho) - \Sigma \right\|^2 \right], \quad \|A\|^2 := tr(A^T A) = \sum_{i,j} A_{i,j}^2, \tag{4}$$

where the resulting shrinkage parameter ρ_{OD} is called an Oracle estimator with a diagonal target. The problem (4) is quadratic in ρ, and thus, has the following closed-form solution.

Theorem 1 *Suppose S is the unbiased sample covariance matrix (1) and T is a symmetric target matrix. The optimal shrinkage parameter that solves (4) is*

$$\rho_{OD}(\Sigma, T) = \frac{E\left[tr(\Sigma - S)(T - S) \right]}{E\left[\|T - S\| \right]}. \tag{5}$$

If, in addition, x_i follows a joint normal distribution $N(\mu, \Sigma)$, and the target matrix is the diagonal elements of the covariance matrix $T = diag(S)$, (5) can be written as

$$\rho_{OD}(\Sigma) := \rho_{OD}(\Sigma, diag(S)) = \frac{1}{1 + (n-1)\phi(\Sigma)} \in (0, 1], \tag{6}$$

where $\phi(\Sigma)$ is

$$\phi(\Sigma) := \frac{tr(\Sigma^2) - tr(diag(\Sigma)^2)}{tr(\Sigma^2) + tr(\Sigma)^2 - 2tr(diag(\Sigma)^2)} \in [0, 1). \tag{7}$$

Proof. See Appedix A. ■

The oracle shrinkage parameter of (6) is optimal but infeasible since it relies on the true covariance matrix Σ. A natural sample analogue is

$$\rho_{OD} := \rho_{OD}(S) = \frac{1}{1 + (n-1)\phi}, \tag{8}$$

where $\phi := \phi(S)$ replaces the true covariance matrix Σ with the sample analogue S in (7).

It turns out that this OD estimator ρ_{OD} may not perform better than an alternative approach, which we call Oracle approximating shrinkage with diagonal target ($OASD$) and uses the limit of the following iteration indexed by j

$$\Sigma_j = (1 - \rho_j)S + \rho_j diag(S), \tag{9}$$

$$\rho_{j+1} = \frac{tr(\Sigma_j S) - 2tr(diag(\Sigma_j)^2) + tr(\Sigma_j)^2}{ntr(\Sigma_j S) - (n+1)tr(diag(\Sigma_j)^2) + tr(\Sigma_j)^2}. \tag{10}$$

The updating equation (10) replaces the true covariance matrix Σ in (6) by the sample covariance matrix S except for the squared terms Σ^2, in which case only one of them is replaced by the sample covariance matrix as $\Sigma_j S$. In this way, ρ_j^2 does not show up and the system of equations remains tractable.

The following main theorem shows that the iteration converges to a unique limit irrespective of the initial value $\rho_0 \in (0, 1)$.

Theorem 2 *For any initial value $\rho_0 \in (0,1)$, the sequence $\{\rho_j\}_j$ specified by (9) and (10) monotonically converges to*

$$\rho_{OASD} := \min\left\{\frac{1}{n\phi}, 1\right\} \in (0,1]. \tag{11}$$

Proof. See Appendix B ■

We note three observations. First, the shrinkage parameter satisfies $\rho_{OASD} \in (0,1]$, and thus, the covariance estimator

$$S_{OASD} := (1-\rho_{OASD})S + \rho_{OASD} diag(S) \tag{12}$$

is positive definite. Second, the shrinkage parameter ρ_{OASD} in (11) contains min operator, but this is a result of the convergence and is not manually imposed, as can be seen in the proof. Third, the formula does not contain the dimension of the sample covariance matrix p, unlike the OAS in (21).

2.1 Special Case: Known Mean

This section provides the formula for the special case where the mean is known to be zero $\mu = 0$. This specification has been used in the literature (Ledoit and Wolf (2004), Chen et al. (2009)), and thus, allows us to compare the performance of different methods, although the general setup with unknown mean is more useful in practice.

It turns out that the resulting formula replaces n in (6), (8), and (11) by $n+1$.

Theorem 3 *Suppose $x_i \sim N(0,\Sigma)$ is i.i.d., and the sample covariance matrix (1) is replaced by*

$$S := \frac{1}{n}\sum_{i=1}^{n} x_i x_i^T. \tag{13}$$

Then, the Oracle (6), its sample analogue (8), and OASD estimator (11) are replaced by

$$\rho_{OD}(\Sigma) := \frac{1}{1 + n\phi(\Sigma)} \in (0, 1], \tag{14}$$

$$\rho_{OD} := \rho_{OD}(S) = \frac{1}{1 + n\phi} \in (0, 1], \tag{15}$$

$$\rho_{OASD} := \min\left\{\frac{1}{(n+1)\phi}, 1\right\} \in (0, 1]. \tag{16}$$

Proof. See Appendix C ■

We note three observations. First, the formula for ϕ remains the same as (7), but with Σ replaced by (13) instead of (1). Second, as in Theorem 2, the shrinkage parameter satisfies $\rho_{OASD} \in (0, 1]$, so the covariance estimator

$$S_{OASD} := (1 - \rho_{OASD})S + \rho_{OASD} diag(S) \tag{17}$$

is positive definite. Third, the shrinkage parameter ρ_{OASD} contains min operator, but this is a result of the convergence and is not manually imposed, as can be seen in the proof.

3 Simulation

This section uses simulations to assess the performance of the $OASD$ estimator S_{OASD} in a high-dimensional environment with large variation in the diagonal elements of the true covariance matrix Σ. The $OASD$ performs better than other methods in most cases with different degrees of variation, sparsity of the true correlation matrix, as and sample sizes. The $OASD$ also exhibits a better performance when it is inverted and when it is compared with alternative methods that shrinks correlation matrices.

3.1 Setting

To conduct simulations in a high-dimensional environment, fix the dimension of the matrices by $p = 100$ and let the sample size n vary from 6 to 30. The true covariance matrix Σ is created from a correlation matrix Γ with a decaying off-diagonal elements $\Gamma_{ij} = \gamma^{|i-j|}$, where γ controls the sparsity and varies from 0 to .9.[1] Up to here, the high-dimensional simulation environment resembles the one in Chen et al. (2009).

To generate the variation across the diagonal elements of the true covariance matrix Σ, we assume half of variables have different unit,

$$\Sigma = \Lambda\Gamma\Lambda, \quad \Lambda = \Lambda^T = \begin{bmatrix} 1 & & & 0 \\ & \ddots & & \\ & & 1 \; sd & \\ & & & \ddots \\ 0 & & & sd \end{bmatrix}, \tag{18}$$

where the parameter for the standard deviation, sd, varies from 1 to 20. Large variations in scales are often of interest in applications, including macroeconomic forecasting. For example, GDP can be a summation of small industries' value added. The government's tax revenue can be a sum of small municipalities. In these cases, the units of variables can differ by hundreds of times.

We generate $\{x_i\}_{i=1}^n$ from a normal distribution $N(0, \Sigma)$ and repeat the sampling $B = 5000$ times. The number of sampling is so large that the sampling errors can be ignored. The performance criterion is the percentage relative improvement in average loss ($PRIAL$), defined as

$$PRIAL(\hat{S}) := \left(1 - \frac{\sum_{b=1}^{B} \left\| \hat{S}^{(b)} - \Sigma \right\|^2}{\sum_{b=1}^{B} \left\| S^{(b)} - \Sigma \right\|^2} \right) \times 100, \tag{19}$$

where $S^{(b)}$ and $\hat{S}^{(b)}$ denote the sample and estimated covariance matrices at the b^{th} sampling. $PRIAL$ can be considered a measure of improvement from the sample covariance matrix S to $\hat{S}$, taking 100 when the estimated covariance $\hat{S}$ coincides with the true covariance Σ, 0

[1] We set $\Gamma_{ij} = 1$ when $\gamma = 0$ and $i = j$.

when the estimated covariance $\hat{S}$ is as poor as the sample covariance matrix S, a negative value when $\hat{S}$ is worse than the sample covariance matrix S.

To assess the performance of $OASD$, we assume a known mean and compare the $OASD$ of Theorem 3 with three methods in the literature, which also assume the known mean in their derivations except for SS. For SS, we derive the formula with know mean. First, we denote by LW the estimator proposed by Ledoit and Wolf (2004)

$$S_{LW} := (1-\rho_{LW})S + \rho_{LW}\frac{tr(S)}{p}I, \quad \rho_{LW} := \min\left\{\frac{\sum_{i=1}^{n}\left\|x_i x_i^T - S\right\|^2}{n^2\left[tr(S^2) - \frac{tr(S)^2}{p}\right]}, 1\right\}. \tag{20}$$

Second, we denote by OAS the estimator proposed by Chen et al. (2009)[2]

$$S_{OAS} := (1-\rho_{OAS})S + \rho_{OAS}\frac{tr(S)}{p}I, \quad \rho_{OAS} := \min\left\{\frac{\left(1-\frac{2}{p}\right)tr(S^2) + tr(S)^2}{\left(n+1-\frac{2}{p}\right)\left[tr(S^2) - \frac{tr(S)^2}{p}\right]}, 1\right\}. \tag{21}$$

Third, we denote by SS the estimator proposed by Schäfer and Strimmer (2005)

$$S_{SS} := (1-\rho_{SS})S + \rho_{SS}diag(S), \quad \rho_{SS} := \min\left\{\frac{\sum_{k\neq l}\widehat{Var}(r_{kl})}{\sum_{k\neq l} r_{kl}^2}, 1\right\}, \tag{22}$$

where r_{kl} is the (k,l) element of the sample correlation matrix and $\widehat{Var}(r_{kl})$ is the sample variance estimator of r_{kl}. Note that the min operator appears as a natural consequence of the proof for OAS but is manually imposed for LW and SS. Finally, we also compare $OASD$ with the sample analogues of the Oracle estimator $S_{OD} = \hat{S}(\rho_{OD})$.

In summary, we compare five estimators, $\{S_{LW}, S_{OAS}, S_{OASD}, S_{OD}, S_{SS}\}$, by varying the three parameters $\{n, sd, \gamma\}$ that control the sample size, the variation in the variances, and the sparsity of the true correlation matrix Γ. For exposition, we move each parameter one by one, fixing others at their medians.

[2]This formula is a modified version of equation (23) of Chen et al. (2009), which has a typo in the numerator.

3.2 Main Results

The following subsections demonstrate that, compared with other methods, the $OASD$ exhibits a higher $PRIAL$ and that the shrinkage parameter ρ_{OASD} tracks the infeasible Oracle estimator $\rho_{OD}(\Sigma)$ closer in all three dimensions $\{n, sd, \gamma\}$.

3.2.1 Variation in Scales

Figure 1 shows the $PRIAL$ on the left and average shrinkage parameters on the right for each method over the variation in scales sd.

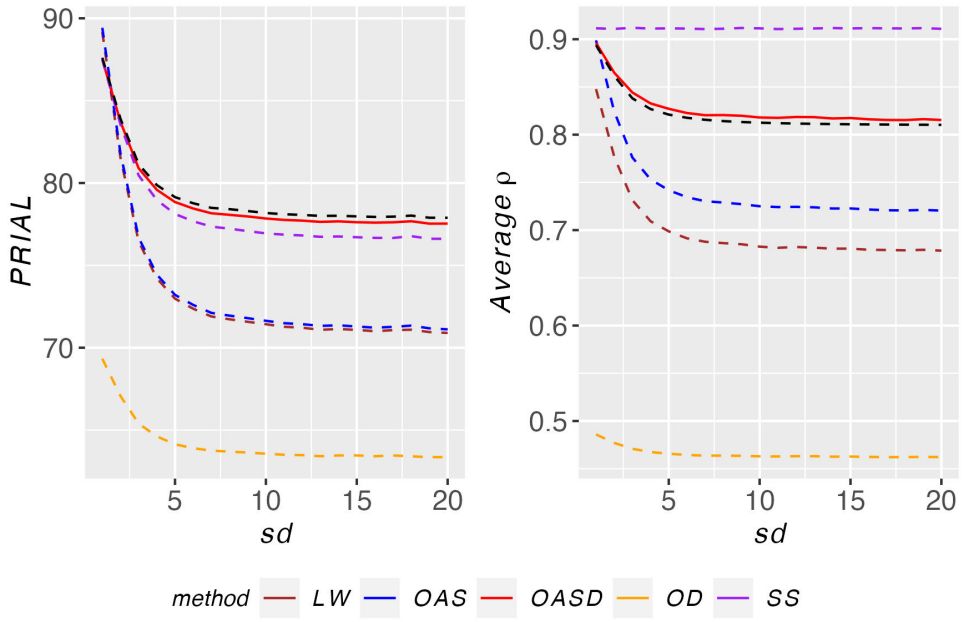
For most regions of the variation parameter sd except for the areas with small values, the $OASD$ exhibits a higher $PRIAL$ than the average variance methods, LW and OAS. This is not surprising since, the larger the variation in scales is, the closer the diagonal target is to the true covariance matrix compared with the average variance target. When the variation parameter sd is small, the methods with the average variance target, LW and OAS, perform better, suggesting that the proposed method $OASD$ is a complement to existing methods rather than a dominant choice, although the difference is relatively small.

The $OASD$ also shows a higher $PRIAL$ than SS by around 1 percent. The improvement can be attributed to the better approximation to the oracle weight $\rho_{OD}(\Sigma)$, as in the right chart of Figure 1. The shrinkage parameter ρ_{SS} remains constant since its formula only contains the elements of correlation matrix, which is constant over sd.

Interestingly, the sample analogue OD is a much poorer estimator than others. The difference between (14) and (15) is $1 - \phi$ in the denominator. Numerically, however, the resulting shrinkage parameter differs by about half with $n = 18$ and $\phi \approx .065$.

All methods exhibit a lower $PRIAL$ as the variation in scales becomes larger. This is because the variation in the off-diagonal elements of the true covariance matrix Σ is larger, and thus, the approximation by the target matrices with null off-diagonals becomes poorer. Accordingly, the shrinkage parameter decreases. This is also the case when the sparsity decreases as the next section shows.

Figure 1: Comparison of methods with different variable scales *sd*



Note: The black lines denote the unfeasible true values $(\hat{S}_{OD}(\Sigma), \rho_{OD}(\Sigma))$. The above results are generated under $(n, \gamma) = (18, .5)$.

3.2.2 Sparsity of Correlation Matrix

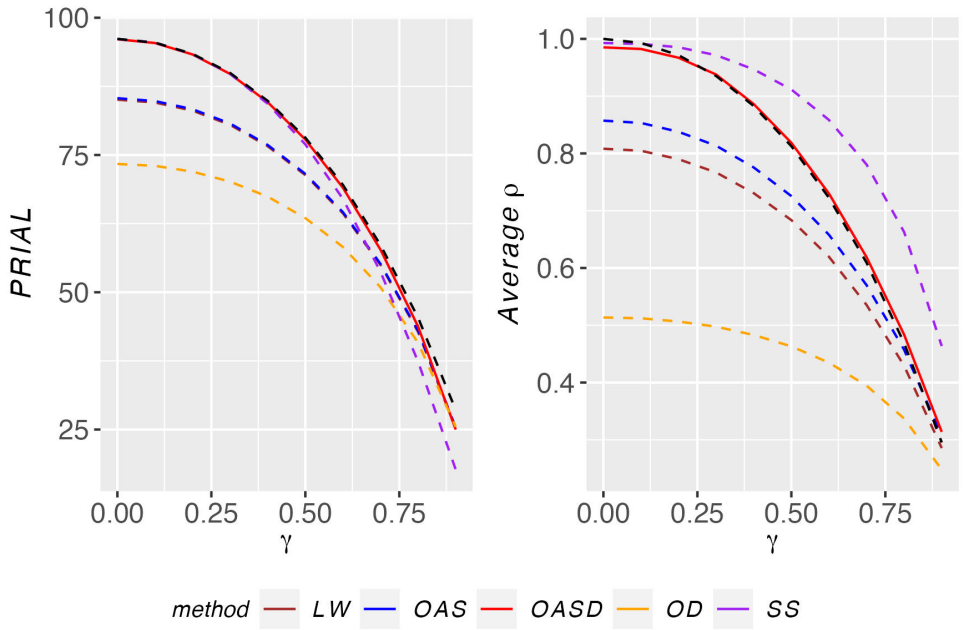
Figure 2 shows the $PRIAL$ on the left and average shrinkage parameters on the right for each method over the sparsity of the correlation matrix γ.

The $OASD$ exhibits a higher $PRIAL$ than other methods. The improvement compared to the methods with the average variance target, LW and OAS, can be up to 10 percent when the true covariance matrix Σ is sparser. One way to understand this comparative statics is to consider the limit case $\gamma \to 0$, where the true covariance matrix Σ is diagonal. The $OASD$ can shrink the off-diagonals without distorting diagonal elements, but LW and OAS face the trade-off of shrinking off-diagonals and distorting diagonal elements. When the true covariance matrix Σ becomes denser, the difference is smaller since most improvement comes from off-diagonals, so the difference in the target matrices matters less.

The $OASD$ also performs better than SS by up to 10 percent. The difference in $PRIAL$ is similar when the true covariance matrix Σ is sparse, but the difference becomes larger as

the sparsity decreases. This can be attributed to the better approximation of the shrinkage parameter ρ_{OASD} to the oracle weight $\rho_{OD}(\Sigma)$ compared to ρ_{SS}, as can be seen in the right chart of Figure 2. The better approximation of $OASD$ also explains the higher $PRIAL$ over OD.

Figure 2: Comparison of methods with different levels of correlation sparsity γ



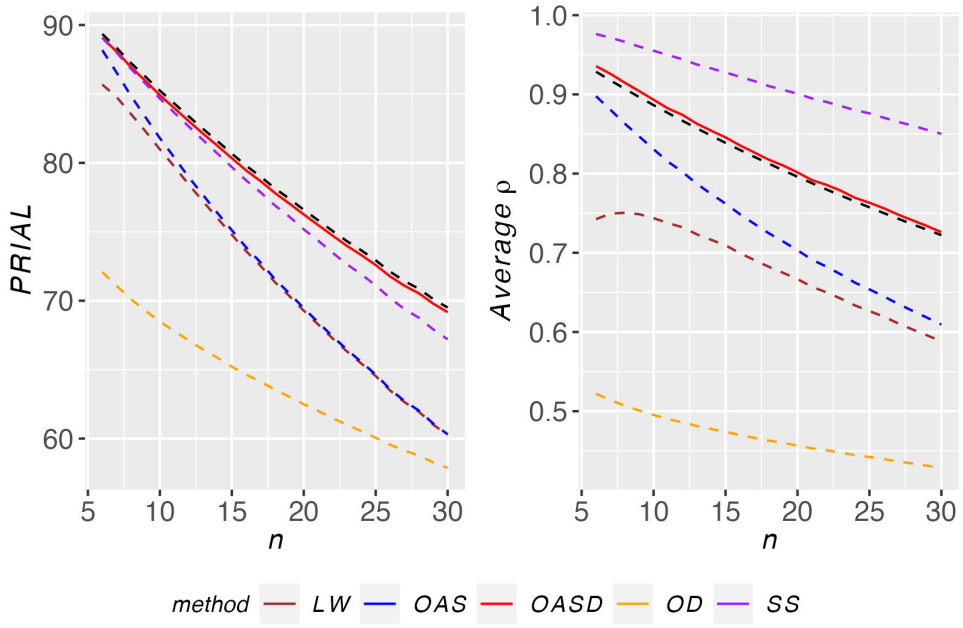
Note: The black lines denote the unfeasible true values $(\hat{S}_{OD}(\Sigma), \rho_{OD}(\Sigma))$. The above results are generated under $(n, sd) = (18, 10)$.

3.2.3 Sample Size

Figure 3 shows the $PRIAL$ on the left and average shrinkage parameters on the right for each method over the sample size n.

The $OASD$ performs best over all sample size n. On average, the $PRIAL$ of $OASD$ is 10 percent higher than LW and OAS and 2 percent higher than SS. The difference increases as the sample size n increases in the region of the chart, but the difference shrinks eventually as $n \to \infty$ and sample covariance matrix S converges to the true covariance matrix Σ. The shrinkage parameter ρ_{OASD} tracks the oracle weight $\rho_{OD}(\Sigma)$ closer for all sample sizes.

Figure 3: Comparison of methods with different sample sizes n



Note: The black lines denote the unfeasible true values $(\hat{S}_{OD}(\Sigma), \rho_{OD}(\Sigma))$. The above results are generated under $(sd, \gamma) = (10, .5)$.

3.3 Performance of Inverse Matrix

This section shows that the inverse of the S_{OASD} can approximate the inverse of the true matrix Σ^{-1} better than other methods. The result is of independent interest since all shrinkage methods try to minimize the MSE of the covariance matrix (4), not its inverse, although in practice, the inverse can often be the ultimate goal of estimating a covariance matrix.

To measure the performance, the criterion of $PRIAL$ is modified to

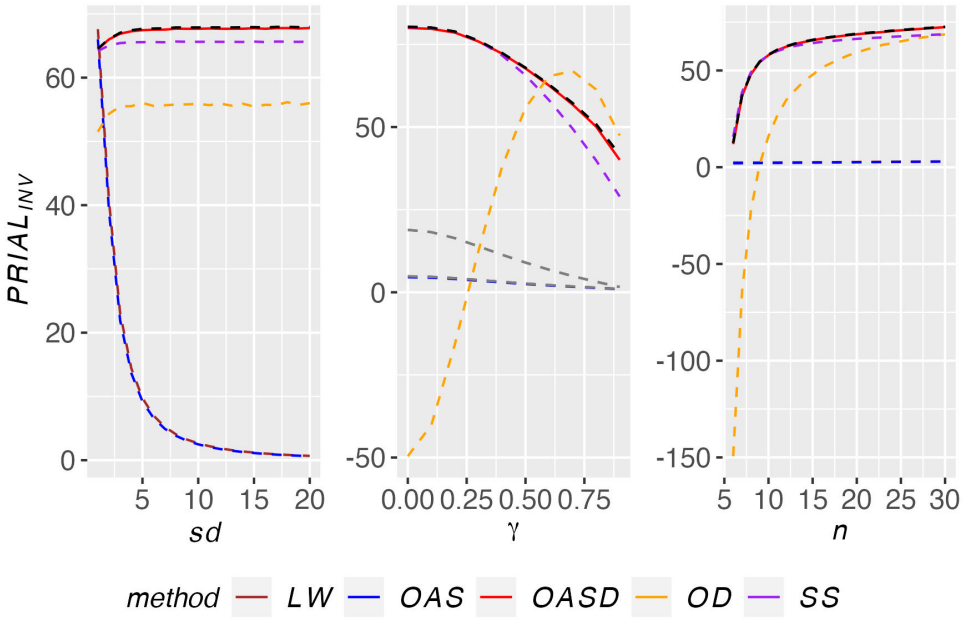
$$PRIAL_{INV}(\hat{S}) := \left(1 - \frac{\sum_{b=1}^{B} \left\| \{\hat{S}^{(b)}\}^{-1} - \Sigma^{-1} \right\|^2}{\sum_{b=1}^{B} \left\| \{S^{(b)}\}^{+} - \Sigma^{-1} \right\|^2}\right) \times 100, \tag{23}$$

where A^+ denotes the generalized inverse of A. This criterion shares the spirit with (19) but uses generalized inverse for the denominator since the sample covariance matrices are not invertible in a high-dimensional environment $p > n$.

Figure 4 shows that the inverse of $OASD$ tends to perform better than other methods.

Intuitively, suppose the true covariance matrix Σ is a 2×2 diagonal matrix with 1 and 10 on the diagonal. The inverse Σ^{-1} has 1 and .1 on the diagonal. If the sample covariance matrix S is close to the true covariance matrix Σ, the inverse of the diagonal target $diag(S)^{-1}$ is also close to the inverse of the true matrix Σ^{-1}. The inverse of the average variance target $[\frac{tr(S)}{2} I]^{-1}$, however, has $1/5.5 \approx 0.2$ on the diagonal, which is close to .1 but not to 1.

Figure 4: Comparison of methods for inverse matrices



Note: The black lines denote the $PRIAL_{INV}$ for the unfeasible true value $\hat{S}_{OD}(\Sigma)$. When one parameter is varied, other parameters are fixed at median $(n, sd, \gamma) = (18, 10, .5)$.

One interesting observation is that the $PRIAL_{INV}$ of the OD method is not monotonic over γ and can be higher than both $OASD$ and the Oracle method (black line), although in some regions, the $PRIAL_{INV}$ can be lower than the generalized inverse of the sample covariance matrix. Both OD and SS share the same diagonal target, but their $PRIAL_{INV}$ differ substantially because the shrinkage parameter of OD is much smaller than SS, especially when the sparsity parameter γ is small, as can seen in Figure 2. The shrinkage parameter, however, happens to give a lower $PRIAL_{INV}$ when γ is large.

3.4 Alternative Method Based on Shrinking Correlation Matrix

When the variation in variable scales is large, one can obtain a sample correlation matrix, apply existing shrinkage methods, and scale back the estimated correlation matrix by multiplying sample standard deviations. This section shows that such alternative methods can reduce the MSE, but the $OASD$, which directly shrinks the covariance matrix S without detouring through the correlation matrix, still outperforms others.

Specifically, let R be a sample correlation matrix. Each existing method e, denoted by $e = LWcorr$, $OAScorr$, and $SScorr$, yields a shrinkage estimator R_e by shrinking the correlation matrix R toward their targets, $\frac{tr(R)}{p}I$ and $diag(R)$ respectively, which both equal to an identity matrix I. The covariance matrix can then be estimated by

$$S_e = diag(S)^{\frac{1}{2}} R_e diag(S)^{\frac{1}{2}}, \quad e \in \{LWcorr, OAScorr, SScorr\}. \tag{24}$$

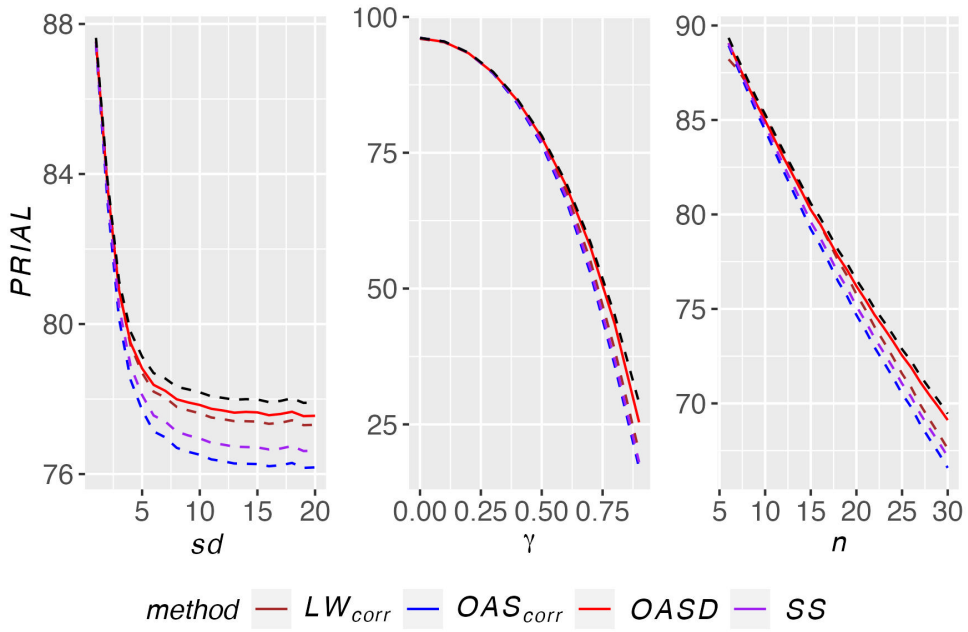
where $diag(S)^{\frac{1}{2}}$ is a diagonal matrix with the sample standard deviations on the diagonals. We note two observations. First, the diagonal elements of S_e equal to the sample variances for all estimation methods $e \in \{LWcorr, OAScorr, SScorr\}$, just like S_{OASD}, mitigating the disadvantage of the methods with the average variance target, LW and OAS. Second, the shrinkage parameter for $SScorr$ remains the same, $\rho_{SScorr} = \rho_{SS}$ since (22) only uses sample correlations, and thus, resulting covariance matrices are also identical $S_{SS} = S_{SS_{corr}}$.

Figure 5 shows that $OASD$ still outperforms other methods. Compared to the figures in section 3.2, the average variance methods, $LWcorr$ and $OAScorr$, exhibit lower MSE, but the performance still lags behind $OASD$. Intuitively, S_e is in the feasible set of the problem that defines the optimal shrinkage toward the diagonal target (4) for all estimation methods $e \in \{LWcorr, OAScorr, SScorr\}$ since

$$S_e = diag(S)^{\frac{1}{2}} \{(1-\rho_e)R + \rho_e I\} diag(S)^{\frac{1}{2}} = (1-\rho_e)S + \rho_e diag(S). \tag{25}$$

The shrinkage parameter ρ_e, however, may not be optimized to minimize the MSE of the covariance matrix, but ρ_{OASD} approximates $\rho_{OD}(\Sigma)$ that does minimize MSE, so ρ_{OASD} can generate a lower MSE.

Figure 5: Comparison of methods through correlation matrix shrinkage



Note: The black lines denote the $PRIAL$ for the unfeasible true value $\hat{S}_{OD}(\Sigma)$. When one parameter is varied, other parameters are fixed at median $(n, sd, \gamma) = (18, 10, .5)$.

4 Conclusion

This paper has proposed a novel covariance matrix estimator $OASD$ that achieves a smaller MSE than existing methods when the variation in variable scales is large. It is useful, for example, when different variables have different units.

We conclude by noting two caveats. First, despite the better performance in simulations, it is important to note that our results are based on a normality assumption. Normalization procedures, such as the Box-Cox transformation, may need to be used if the distribution of data deviate substantially from normality. Second, our methods leave the sample variances

intact and only shrink the off-diagonal entries. Extensions can allow different levels of shrinkage for diagonal and off-diagonal entries, which we leave for future research.

References

Ando, M. S. and Kim, M. T. (2022). *Systematizing Macroframework Forecasting: High-Dimensional Conditional Forecasting with Accounting Identities*. Number 2022-2110. International Monetary Fund.

Ban, G.-Y., El Karoui, N., and Lim, A. E. (2018). Machine learning and portfolio optimization. *Management Science*, 64(3):1136–1154.

Chen, Y., Wiesel, A., and Hero, A. O. (2009). Shrinkage estimation of high dimensional covariance matrices. In *2009 IEEE international conference on acoustics, speech and signal processing*, pages 2937–2940. IEEE.

DeMiguel, V., Garlappi, L., Nogales, F. J., and Uppal, R. (2009). A generalized approach to portfolio optimization: Improving performance by constraining portfolio norms. *Management science*, 55(5):798–812.

Eldar, Y. C. and Chernoi, J. S. (2008). A pre-test like estimator dominating the least-squares method. *Journal of Statistical Planning and Inference*, 138(10):3069–3085.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the econometric society*, pages 1029–1054.

Ledoit, O. and Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of multivariate analysis*, 88(2):365–411.

Ledoit, O. and Wolf, M. (2012). Nonlinear shrinkage estimation of large-dimensional covariance matrices.

Pearson, K. (1901). Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, 2(11):559–572.

Schäfer, J. and Strimmer, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical applications in genetics and molecular biology*, 4(1).

Appendix

A Proof of Theorem 1

The first result can be obtained by a direct calculation. Since T is symmetric,

$$E\left[\left\|\hat{S}(\rho)-\Sigma\right\|^2\right] = E\left[\|(1-\rho)S+\rho T-\Sigma\|^2\right]$$
$$= E\left[\|T-S\|^2\right]\rho^2 + 2E\left[tr\left(\{S-\Sigma\}\{T-S\}\right)\right]\rho + E\left[\|S-\Sigma\|^2\right].$$

The first order condition with respect to ρ leads to

$$\rho = \frac{E\left[tr(\{\Sigma-S\}\{T-S\})\right]}{E\left[\|T-S\|^2\right]}.$$

The second result uses the following lemma.

Lemma 1 *When $x_i \sim N(\mu,\Sigma)$ is i.i.d., the following equations hold.*

$$E\left[tr(\Sigma diag(S))\right] = tr\left(diag(\Sigma)^2\right),$$

$$E\left[tr(S^2)\right] = \frac{n}{n-1}tr(\Sigma^2) + \frac{1}{n-1}tr(\Sigma)^2,$$
$$E\left[tr\left(Sdiag(S)\right)\right] = E\left[tr\left(diag(S)^2\right)\right] = \frac{n+1}{n-1}tr\left(diag(\Sigma)^2\right).$$

Proof. The first equation is a direct calculation.

$$E\left[tr(\Sigma diag(S))\right] = E\left[\sum_{m=1}^{p}\Sigma_{mm}S_{mm}\right] = \sum_{m=1}^{p}(\Sigma_{mm})^2 = tr(diag(\Sigma)^2).$$

For the second equation, let $w_i = x_i - \bar{x}$. Since $x_i \sim N(\mu,\Sigma)$, the demeaned variable also follows a joint normal distribution

$$w_i = \frac{n-1}{n}x_i - \frac{1}{n}\sum_{k\neq i}x_k \sim N(0,U), \quad U = \frac{n-1}{n}\Sigma.$$

Note that U is symmetric, so it can be diagonalized as $U = VDV^T$, where V is an orthogonal matrix and D is a diagonal matrix. Since $n \geq 2$ and $\Sigma > 0$, $U^{\frac{1}{2}} := VD^{\frac{1}{2}}V^T$ is invertible and can be used to transform w_i into a standard normal distribution

$$z_i := V^T U^{-\frac{1}{2}} w_i \sim N(0, I).$$

We decompose the left hand side into two components.

$$\begin{aligned}
E\left[tr(S^2)\right] &= E\left[tr\left(\left\{\frac{1}{n-1}\sum_{i=1}^{n} w_i w_i^T\right\}^2\right)\right] \\
&= \frac{1}{(n-1)^2} E\left[tr\left(\sum_{i=1}^{n}(w_i w_i^T)^2 + \sum_{i=1, j\neq i}^{n} w_i w_i^T w_j w_j^T\right)\right] \\
&= \frac{1}{(n-1)^2} E\left[\sum_{i=1}^{n}(w_i^T w_i)^2 + \sum_{i=1, j\neq i}^{n}(w_i^T w_j)^2\right].
\end{aligned}$$

Let's zoom in on the first component

$$E\left[(w_i^T w_i)^2\right] = Var\left[w_i^T w_i\right] + E\left[w_i^T w_i\right]^2.$$

We can write the inner product as

$$w_i^T w_i = \left(V^T U^{-\frac{1}{2}} w_i\right)^T D \left(V^T U^{-\frac{1}{2}} w_i\right) = \sum_{m=1}^{p} \lambda_m z_{im}^2,$$

where λ_m is the m^{th} diagonal element of D and eigenvalue of U. Since $E[z_{im}^2] = 1$,

$$E\left[w_i^T w_i\right]^2 = \left(\sum_{m=1}^{p} \lambda_m E[z_{im}^2]\right)^2 = \left(\sum_{m=1}^{p} \lambda_m\right)^2 = tr(U)^2.$$

For the variance, note that the normality of z_{im} implies $Var[z_{im}^2] = E[z_{im}^4] - (E[z_{im}^2])^2 = 2$, and the joint normality $z_i \sim N(0, I)$ implies the independence of z_{ik} and z_{il}, which then

implies the independence of z_{ik}^2 and z_{il}^2 when $k \neq l$.

$$Var\left[w_i^T w_i\right] = \sum_{m=1}^{p} \lambda_m^2 Var\left[z_{im}^2\right] = 2\sum_{m=1}^{p} \lambda_m^2 = 2tr(U^2).$$

Therefore, the first component can be written as

$$E\left[(w_i^T w_i)^2\right] = 2tr(U^2) + tr(U)^2.$$

Similarly, we can calculate the second component

$$E\left[(w_i^T w_j)^2\right] = Var\left[w_i^T w_j\right] + E\left[w_i^T w_j\right]^2,$$

using the transformation

$$w_i^T w_j = (V^T U^{-\frac{1}{2}} w_i)^T D(V^T U^{-\frac{1}{2}} w_j) = \sum_{m=1}^{p} \lambda_m z_{im} z_{jm}.$$

Since w_i and w_j can be rewritten as

$$w_i = \frac{n-1}{n}(x_i - \mu) - \frac{1}{n}(x_j - \mu) - \frac{1}{n}\sum_{k \neq i,j}(x_k - \mu),$$

$$w_j = -\frac{1}{n}(x_i - \mu) + \frac{n-1}{n}(x_j - \mu) - \frac{1}{n}\sum_{k \neq i,j}(x_k - \mu),$$

the independence of x_i over i implies

$$E\left[w_i w_j^T\right] = -\frac{n-1}{n^2}\Sigma - \frac{n-1}{n^2}\Sigma + \frac{n-2}{n^2}\Sigma = -\frac{1}{n}\Sigma = -\frac{1}{n-1}U,$$

and thus, the first moment of $w_i^T w_j$ and $z_{im} z_{jm}$ can be written as

$$E\left[w_i^T w_j\right] = tr(E\left[w_i w_j^T\right]) = -\frac{1}{n-1}tr(U),$$

$$E\left[z_i z_j^T\right] = E\left[V^T U^{-\frac{1}{2}} w_i w_j^T U^{-\frac{1}{2}} V\right] = V^T U^{-\frac{1}{2}} \left(-\frac{1}{n-1} U\right) U^{-\frac{1}{2}} V = -\frac{1}{n-1} I.$$

For the second moment of $z_{im} z_{jm}$, the formula for multivariate normal distribution implies

$$E\left[(z_{im} z_{jm})^2\right] = Var\left[z_{im}\right] Var\left[z_{jm}\right] + 2Cov\left[z_{im}, z_{jm}\right]^2 = 1 + 2E\left[z_{im} z_{jm}\right]^2 = 1 + \frac{2}{(n-1)^2},$$

$$Var\left[z_{im} z_{jm}\right] = E\left[(z_{im} z_{jm})^2\right] - \left(E\left[z_{im} z_{jm}\right]\right)^2 = 1 + \frac{2}{(n-1)^2} - \frac{1}{(n-1)^2} = 1 + \frac{1}{(n-1)^2}.$$

Note that the joint normal distribution implies independence between $z_{ik} z_{jk}$ and $z_{il} z_{jl}$

$$\begin{bmatrix} z_i \\ z_j \end{bmatrix} \sim N\left(0, \begin{bmatrix} I & -\frac{1}{n-1} I \\ -\frac{1}{n-1} I & I \end{bmatrix}\right) \Rightarrow \begin{bmatrix} z_{ik} \\ z_{jk} \end{bmatrix} \perp\!\!\!\perp \begin{bmatrix} z_{il} \\ z_{jl} \end{bmatrix} \Rightarrow z_{ik} z_{jk} \perp\!\!\!\perp z_{il} z_{ik}, \quad k \neq l.$$

Therefore, the variance and the second moment of the cross-terms are

$$V\left[w_i^T w_j\right] = \sum_{m=1}^{p} \lambda_m^2 V\left[z_{im} z_{jm}\right] = \sum_{m=1}^{p} \lambda_m^2 \left(1 + \frac{1}{(n-1)^2}\right) = \left\{1 + \frac{1}{(n-1)^2}\right\} tr(U^2),$$

$$E\left[(w_i^T w_j)^2\right] = \left\{1 + \frac{1}{(n-1)^2}\right\} tr(U^2) + \frac{1}{(n-1)^2} tr(U)^2.$$

Putting all together, we have

$$\begin{aligned} E\left[tr(S^2)\right] &= \frac{1}{(n-1)^2} E\left[\sum_{i=1}^{n} (w_i^T w_i)^2 + \sum_{i=1, j \neq i}^{n} (w_i^T w_j)^2\right] \\ &= \frac{1}{(n-1)^2} \left[nE\left[(w_i^T w_i)^2\right] + (n^2 - n) E\left[(w_i^T w_j)^2\right]\right] \\ &= \frac{n^3}{(n-1)^3} tr(U^2) + \frac{n^2}{(n-1)^3} tr(U)^2 \\ &= \frac{n}{n-1} tr(\Sigma^2) + \frac{1}{n-1} tr(\Sigma)^2. \end{aligned}$$

For the third equation, the left hand side can be written as

$$E\left[tr(Sdiag(S))\right] = \sum_{m=1}^{p} E\left[(S_{mm})^2\right].$$

The summand can be decomposed into two components.

$$\begin{aligned} E\left[(S_{mm})^2\right] &= E\left[\frac{1}{(n-1)^2}\left(\sum_{i=1}^{n} w_{im}^2\right)^2\right] \\ &= \frac{1}{(n-1)^2} E\left[\sum_{i=1}^{n} w_{im}^4 + \sum_{i=1, j\neq i}^{n} w_{im}^2 w_{jm}^2\right] \\ &= \frac{1}{(n-1)^2}\left(\sum_{i=1}^{n} E\left[w_{im}^4\right] + \sum_{i=1, j\neq i}^{n} E\left[w_{im}^2 w_{jm}^2\right]\right). \end{aligned}$$

From the normality and the first moment of the cross term

$$w_{im} \sim N\left(0, \frac{n-1}{n}\Sigma_{mm}\right), \quad E\left[w_{im} w_{jm}\right] = -\frac{\Sigma_{mm}}{n},$$

we can obtain

$$E\left[w_{im}^4\right] = 3\left(\frac{n-1}{n}\right)^2 (\Sigma_{mm})^2,$$

$$E\left[w_{im}^2 w_{jm}^2\right] = \left(\frac{n-1}{n}\right)^2 (\Sigma_{mm})^2 + 2\left(\frac{\Sigma_{mm}}{n}\right)^2 = \frac{n^2-2n+3}{n^2}(\Sigma_{mm})^2.$$

Substituting these expression gives

$$E\left[(S_{mm})^2\right] = \frac{n+1}{n-1}(\Sigma_{mm})^2.$$

Therefore,

$$E\left[tr(Sdiag(S))\right] = \sum_{m=1}^{p} E\left[(S_{mm})^2\right] = \frac{n+1}{n-1} tr(diag(\Sigma)^2).$$

■

The second result of the theorem follows by substituting $T = diag(S)$ and the lemma.

$$\begin{aligned}
\rho &= \frac{E\left[tr(\Sigma T) - tr(\Sigma S) - tr(ST) + tr(S^2)\right]}{E\left[tr(S^2) - 2tr(ST) + tr(T^2)\right]} \\
&= \frac{tr(diag(\Sigma)^2) - tr(\Sigma^2) - \frac{n+1}{n-1}tr(diag(\Sigma)^2) + \frac{n}{n-1}tr(\Sigma^2) + \frac{1}{n-1}tr(\Sigma)^2}{\frac{n}{n-1}tr(\Sigma^2) + \frac{1}{n-1}tr(\Sigma)^2 - 2(\frac{n+1}{n-1})tr(diag(\Sigma)^2) + (\frac{n+1}{n-1})tr(diag(\Sigma)^2)} \\
&= \frac{-\frac{2}{n-1}tr(diag(\Sigma)^2) + \frac{1}{n-1}tr(\Sigma^2) + \frac{1}{n-1}tr(\Sigma)^2}{\frac{n}{n-1}tr(\Sigma^2) + \frac{1}{n-1}tr(\Sigma)^2 - (\frac{n+1}{n-1})tr(diag(\Sigma)^2)} \\
&= \frac{-2tr(diag(\Sigma)^2) + tr(\Sigma^2) + tr(\Sigma)^2}{ntr(\Sigma^2) + tr(\Sigma)^2 - (n+1)tr(diag(\Sigma)^2)} \\
&= \frac{tr(\Sigma^2) + tr(\Sigma)^2 - 2tr(diag(\Sigma)^2)}{tr(\Sigma^2) + tr(\Sigma)^2 - 2tr(diag(\Sigma)^2) + (n-1)\left\{tr(\Sigma^2) - tr(diag(\Sigma)^2)\right\}} \\
&= \frac{1}{1 + (n-1)\phi(\Sigma)}
\end{aligned}$$

where

$$\phi(\Sigma) = \frac{tr(\Sigma^2) - tr(diag(\Sigma)^2)}{tr(\Sigma^2) + tr(\Sigma)^2 - 2tr(diag(\Sigma)^2)}.$$

$\phi \in [0, 1)$ and $\rho \in (0, 1]$ follow by noting

$$tr(\Sigma^2) = tr(\Sigma^T\Sigma) = \sum_{k,l}(\Sigma_{kl})^2 \geq \sum_m(\Sigma_{mm})^2 = tr(diag(\Sigma)^2),$$

$$tr(\Sigma)^2 = \left(\sum_m \Sigma_{mm}\right)^2 > \sum_m(\Sigma_{mm})^2 = tr(diag(\Sigma)^2),$$

where the strict inequality follows from the positive definiteness of Σ.

B Proof of Theorem 2

Substituting $\Sigma_j = (1-\rho_j)S + \rho_j diag(S)$ and a direct calculation lead to

$$\begin{aligned}
\rho_{j+1} &= \frac{-2tr(diag(\Sigma_j)^2) + tr(\Sigma_j S) + tr(\Sigma_j)^2}{ntr(\Sigma_j S) + tr(\Sigma_j)^2 - (n+1)tr(diag(\Sigma_j)^2)} \\
&= \frac{-2tr(diag(S)^2) + tr(\Sigma_j S) + tr(S)^2}{ntr(\Sigma_j S) + tr(S)^2 - (n+1)tr(diag(S)^2)} \\
&= \frac{-2tr(diag(S)^2) + tr(\{(1-\rho_j)S + \rho_j diag(S)\}S) + tr(S)^2}{ntr(\{(1-\rho_j)S + \rho_j diag(S)\}S) + tr(S)^2 - (n+1)tr(diag(S)^2)} \\
&= \frac{\rho_j \{tr(diag(S)^2) - tr(S^2)\} - 2tr(diag(S)^2) + tr(S^2) + tr(S)^2}{\rho_j n \{tr(diag(S)^2) - tr(S^2)\} - (n+1)tr(diag(S)^2) + ntr(S^2) + tr(S)^2} \\
&= \frac{1-\rho_j\phi}{1-\rho_j n\phi + (n-1)\phi},
\end{aligned}$$

where ϕ is

$$\phi = \frac{tr(S^2) - tr(diag(S)^2)}{tr(S^2) + tr(S)^2 - 2tr(diag(S)^2)}.$$

Similarly to $\phi(\Sigma)$, $\phi \in [0,1)$ because

$$tr(S^2) = tr(S^T S) = \sum_{k,l}(S_{kl})^2 \geq \sum_m (S_{mm})^2 = tr(diag(S)^2),$$

$$tr(S)^2 = \left(\sum_m S_{mm}\right)^2 > \sum_m (S_{mm})^2 = tr(diag(S)^2),$$

where the strict inequality is due to the assumption that the sample variances are positive $S_{mm} > 0$. If $\phi = 0$, $(n\phi)^{-1} = \infty$ and $\rho_j = 1$ for all j, so the statement is proved. Suppose $\phi \in (0,1)$. One can see $\rho_j \in (0,1)$ for all j by noting

$$\rho_{j+1} = \frac{1-\rho_j\phi}{1-\rho_j\phi + (n-1)\phi(1-\rho_j)}, \quad \rho_0 \in (0,1).$$

If $n\phi < 1$, $\rho_j < 1 < (n\phi)^{-1}$ for all j, so the following change of variable is well-defined

$$b_j := \frac{1}{\rho_j - \frac{1}{n\phi}} \Leftrightarrow \rho_j = \frac{1}{b_j} + \frac{1}{n\phi},$$

and the updating equation can be simplified to the following recursion

$$b_{j+1} = \frac{\phi(n-1)}{1-\phi} b_j - \frac{n\phi}{1-\phi} \Leftrightarrow b_{j+1} - \frac{n\phi}{n\phi - 1} = \frac{\phi(n-1)}{1-\phi}\left(b_j - \frac{n\phi}{n\phi-1}\right).$$

The statement is proved by noting

$$n\phi < 1 \Leftrightarrow \frac{\phi(n-1)}{1-\phi} < 1 \Rightarrow b_j \to \frac{n\phi}{n\phi-1} \Rightarrow \rho_j \to 1,$$

and that the convergence is monotonic.

If $n\phi = 1$, the same change of variable proves the statement.

$$b_{j+1} = b_j - \frac{1}{1-\phi} \to -\infty \Rightarrow \rho_j \to \frac{1}{n\phi} = 1.$$

Finally, suppose $n\phi > 1$. If $\rho_j = (n\phi)^{-1}$ for some j, $\rho_{j'} = (n\phi)^{-1}$ for all $j' \geq j$, so the statement is proved. Otherwise, the same change of variable gives a well-defined b_j

$$b_{j+1} - \frac{n\phi}{n\phi-1} = \frac{\phi(n-1)}{1-\phi}\left(b_j - \frac{n\phi}{n\phi-1}\right).$$

By noting

$$n\phi > 1 \Leftrightarrow \frac{\phi(n-1)}{1-\phi} > 1, \quad \rho_j < 1 \Rightarrow b_j > \frac{n\phi}{n\phi-1},$$

one can see the convergence is monotonic and

$$b_j \to \infty \Rightarrow \rho_j \to \frac{1}{n\phi}.$$

Therefore,

$$\rho_{OASD} = \min\left\{\frac{1}{n\phi}, 1\right\}.$$

C Proof of Theorem 3

The proof is a simpler version of Appendix A and B. We first establish the following lemma.

Lemma 2 *When $x_i \sim N(0, \Sigma)$ is i.i.d., the following equations hold.*

$$E\left[tr(\Sigma diag(S))\right] = tr(diag(\Sigma)^2),$$

$$E\left[tr(S^2)\right] = \frac{n+1}{n} tr(\Sigma^2) + \frac{1}{n} tr(\Sigma)^2,$$

$$E\left[tr(S diag(S))\right] = E\left[tr(diag(S)^2)\right] = \frac{n+2}{n} tr(diag(\Sigma)^2).$$

Proof. The first equation is a direct calculation.

$$E\left[tr(\Sigma diag(S))\right] = E\left[\sum_{m=1}^{p} \Sigma_{mm} S_{mm}\right] = \sum_{m=1}^{p} (\Sigma_{mm})^2 = tr(diag(\Sigma)^2).$$

For the second equation,

$$\begin{aligned}
E\left[tr(S^2)\right] &= E\left[tr\left(\left\{\frac{1}{n}\sum_{i=1}^{n} x_i x_i^T\right\}^2\right)\right] \\
&= \frac{1}{n^2} tr\left(E\left[\left(\sum_{i=1}^{n} x_i x_i^T\right)^2\right]\right) \\
&= \frac{1}{n^2} tr\left(Var\left[\sum_{i=1}^{n} x_i x_i^T\right] + \left\{E\left[\sum_{i=1}^{n} x_i x_i^T\right]\right\}^2\right) \\
&= \frac{ntr\left(E\left[(x_i x_i^T)^2\right] - E\left[x_i x_i^T\right]^2\right) + n^2 tr\left(E\left[x_i x_i^T\right]^2\right)}{n^2} \\
&= \frac{ntr\left(E\left[(x_i x_i^T)^2\right]\right) + (n^2 - n) tr(\Sigma^2)}{n^2}.
\end{aligned}$$

The first term can be calculated using diagonalization of $\Sigma = V^T D V$ where $VV^T = I$.

$$tr\left(E\left[(x_i x_i^T)^2\right]\right) = E\left[tr(x_i x_i^T x_i x_i^T)\right] = E\left[(x_i^T x_i)^2\right] = V\left[x_i^T x_i\right] + \left(E\left[x_i^T x_i\right]\right)^2.$$

The integrand can be transformed into

$$x_i^T x_i = \left(V\Sigma^{-\frac{1}{2}}x_i\right)^T D \left(V\Sigma^{-\frac{1}{2}}x_i\right) = \sum_{m=1}^{p} \lambda_m z_{im}^2, \quad z_i := V\Sigma^{-\frac{1}{2}}x_i \sim N(0, I).$$

Using the independence of z_{im} across m and the fourth moment of z_{im} under normality,

$$V\left[x_i^T x_i\right] = \sum_{m=1}^{p} \lambda_m^2 V\left[z_{im}^2\right] = \sum_{m=1} \lambda_m^2 \left(E\left[z_{im}^4\right] - E\left[z_{im}^2\right]^2\right) = 2\sum_{m=1}^{p} \lambda_m^2 = 2tr\left(\Sigma^2\right).$$

Thus

$$tr\left(E\left[(x_i x_i^T)^2\right]\right) = 2tr(\Sigma^2) + tr(\Sigma)^2,$$

and

$$E\left[tr(S^2)\right] = \frac{2ntr(\Sigma^2) + ntr(\Sigma)^2 + (n^2 - n)tr(\Sigma^2)}{n^2} = \frac{n+1}{n}tr(\Sigma^2) + \frac{1}{n}tr(\Sigma)^2.$$

For the third equation,

$$E\left[tr(Sdiag(S))\right] = E\left[tr(diag(S)^2)\right] = \sum_{m=1}^{p} E\left[(S_{mm})^2\right].$$

The result follows by noting

$$\begin{aligned}
E\left[(S_{mm})^2\right] &= E\left[\left(\frac{1}{n}\sum_{i=1}^{n} x_{im}^2\right)^2\right] \\
&= V\left[\frac{1}{n}\sum_{i=1}^{n} x_{im}^2\right] + \left(E\left[\frac{1}{n}\sum_{i=1}^{n} x_{im}^2\right]\right)^2 \\
&= \frac{1}{n}V\left[x_{im}^2\right] + (\Sigma_{mm})^2 \\
&= \frac{1}{n}\left(E\left[x_{im}^4\right] - E\left[x_{im}^2\right]^2\right) + (\Sigma_{mm})^2 \\
&= \left(\frac{2}{n} + 1\right)(\Sigma_{mm})^2.
\end{aligned}$$

■

Substituting $T = diag(S)$ and the equations in the above lemma give

$$\begin{aligned}
\rho &= \frac{E\left[tr(\Sigma T) - tr(\Sigma S) - tr(ST) + tr(S^2)\right]}{E\left[tr(S^2) - 2tr(ST) + tr(T^2)\right]} \\
&= \frac{tr(diag(\Sigma)^2) - tr(\Sigma^2) - \frac{n+2}{n}tr(diag(\Sigma)^2) + \frac{n+1}{n}tr(\Sigma^2) + \frac{1}{n}tr(\Sigma)^2}{\frac{n+1}{n}tr(\Sigma^2) + \frac{1}{n}tr(\Sigma)^2 - \frac{n+2}{n}tr(diag(\Sigma)^2)} \\
&= \frac{-\frac{2}{n}tr(diag(\Sigma)^2) + \frac{1}{n}tr(\Sigma^2) + \frac{1}{n}tr(\Sigma)^2}{\frac{n+1}{n}tr(\Sigma^2) + \frac{1}{n}tr(\Sigma)^2 - \frac{n+2}{n}tr(diag(\Sigma)^2)} \\
&= \frac{-2tr(diag(\Sigma)^2) + tr(\Sigma^2) + tr(\Sigma)^2}{(n+1)tr(\Sigma^2) + tr(\Sigma)^2 - (n+2)tr(diag(\Sigma)^2)} \\
&= \frac{1}{1 + n\phi(\Sigma)}.
\end{aligned}$$

The same argument in Appendix A leads to $\phi(\Sigma) \in [0, 1)$ and $\rho \in (0, 1]$. The iteration is specified by

$$\begin{aligned}
\rho_{j+1} &= \frac{-2tr(diag(\Sigma_j)^2) + tr(\Sigma_j S) + tr(\Sigma_j)^2}{(n+1)tr(\Sigma_j S) + tr(\Sigma_j)^2 - (n+2)tr(diag(\Sigma_j)^2)} \\
&= \frac{(1-\rho_j)tr(S^2) + \rho_j tr(Sdiag(S)) - 2tr(diag(S)^2) + tr(S)^2}{(n+1)\left\{(1-\rho_j)tr(S^2) + \rho_j tr(Sdiag(S))\right\} + tr(S)^2 - (n+1)tr(diag(S)^2)} \\
&= \frac{tr(S^2) + tr(S)^2 - 2tr(diag(S)^2) - \left\{tr(S^2) - tr(Sdiag(S))\right\}\rho_j}{(n+1)tr(S^2) + tr(S)^2 - (n+1)tr(diag(S)^2) - (n+1)\left\{tr(S^2) - tr(Sdiag(S))\right\}\rho_j} \\
&= \frac{1 - \phi\rho_j}{1 + n\phi - (n+1)\phi\rho_j},
\end{aligned}$$

where the parameter ϕ is

$$\phi = \frac{tr(S^2) - tr(diag(S)^2)}{tr(S^2) + tr(S)^2 - 2tr(diag(S)^2)}.$$

Note that the updating equation is identical to the one in Appendix B, except that n is replaced by $n + 1$. Thus, following the same argument, the iteration converges to

$$\rho_{OASD} = \min\left\{\frac{1}{(n+1)\phi}, 1\right\}.$$

PUBLICATIONS

High-Dimensional Covariance Matrix Estimation: Shrinkage Toward a Diagonal Target

Working Paper No. WP/2023/257