High-dimensional Covariance Matrix Estimation: Shrinkage Using a Diagonal Target

Sakai Ando* Mingmei Xiao†

October 4, 2024

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. When the diagonal elements of the true covariance matrix exhibit substantial variation, our method reduces the Mean Squared Error, compared with OAS, which targets an average variance. The degree of improvement is higher when the true covariance matrix is sparser. Our method also outperforms other estimators based on a diagonal target under the normality assumption. We further propose an extended estimator that makes use of two targets: the average variance target and the diagonal target. This more flexible estimator improves upon the single-target estimators in all the scenarios discussed. The proposed estimators are applied to the problem of UK inflation forecast reconciliation and minimum variance portfolio selection to compare their performance with other benchmark methods.

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 important econometric methods such as

*International Monetary Fund, Sando@imf.org
†University of Cambridge, mx235@cam.ac.uk

GMM and PCA (See [Hansen, 1982], [Pearson, 1901]), and empirical applications, including financial portfolio selection and macroeconomic forecasting (See [DeMiguel et al., 2009], [Ban et al., 2018], [Ando and Kim, 2022]). 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 exhibits substantial variation, this paper proposes a shrinkage estimator ($OASD$) 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 possesses 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, we see that the degree of improvement is higher when the true covariance matrix is sparser.

As in [Chen et al., 2009], our method is based on optimality under the normal distribution. Compared to [Schäfer and Strimmer, 2005], which also targets 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 nonlinear shrinkage method of [Ledoit and Wolf, 2012].

However, our proposed estimator $OASD$ does not outperform existing methods in all circumstances and should therefore 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 to

estimate the covariance matrix by applying OAS to the correlation matrix and scaling it back by multiplying sample variances. To examine robustness, we perform 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.

$OASD$ is designed for the case where diagonal elements exhibit large variation, while OAS performs better when diagonal variation is small. To have an estimator that works best in both cases, we further extend the method and propose $OASB$, which allows for two targets: the average variance target used in OAS and the diagonal target in $OASD$. The shrinkage weights for the two targets are chosen in a data-driven manner, adjusting according to features of the sample covariance matrix. Having two parameters instead of one allows us to choose different levels of shrinkage for diagonal and off-diagonal entries. This explains its better performance over OAS (which forces the shrinkage level to be the same for diagonal and off-diagonal entries) and $OASD$ (which keeps the sample variances and only shrinks the off-diagonal entries).

Our two empirical applications: forecast reconciliation and portfolio construction, confirms the lessons we learned in the simulation. When the covariance matrix is dense and variables have less dispersion in variation, OAS and LW perform better than $OASD$, as in the forecast reconciliation results. However, when the variables are less correlated or when they differ greatly in scales, as in our portfolio construction example, $OASD$ performs significantly better. $OASB$ tend to have its performance between the two groups of estimators and is ideal for researchers who are unsure about the patterns of the true covariance matrix.

This paper is organized as follows. Section 2 gives an overview of the literature in this area, Section 3 describes the theoretical framework, Section 4 uses simulations to assess performance and evaluate robustness, Section 5 gives the empirical application of UK inflation forecast reconciliation, Section 6 and Section 7 concludes.

2 Literature Review

The literature on the linear shrinkage of covariance matrices begins with [Stein, 1975], who first demonstrated that shrinking the eigenvalues of the sample covariance matrix could improve its estimation. This insight inspired the empirical Bayes estimator by [Haff, 1980] and the minimax estimator of [Dey and Srinivasan, 1985].

Although these estimators are predicated on normality, their performance with Gaussian samples still lags behind [Ledoit and Wolf, 2003]'s LW estimator, as shown through simulations in [Ledoit and Wolf, 2003]. The LW estimator performs a linear combination of the sample covariance matrix S and the identity matrix, effectively shrinking the sample eigenvalues toward their grand mean while keeping the sample eigenvectors intact. Remarkably, the optimality of LW was established without any specific distributional assumptions. This has led to the widespread use of methods based on LW, particularly in portfolio selection (see [DeMiguel et al., 2009] or [Ban et al., 2018]). However, while [Ledoit and Wolf, 2003] provides a consistent estimator of the optimal combination weight under general asymptotics, its finite-sample efficacy remains uncertain.

[Chen et al., 2009] proposed two improvements to LW, assuming normality: $RBLW$ and OAS. Our simulations show that $RBLW$ does not significantly improve upon LW (which aligns with the findings of [Chen et al., 2009]), whereas OAS outperforms both LW-based methods for Gaussian samples. This is primarily due to the iterative method used in OAS, which achieves a more accurate finite-sample approximation of the optimal combination weight.

[Schäfer and Strimmer, 2005] also followed the linear shrinkage strategy but focused on achieving better finite-sample performance and expanding the list of target matrices. In contrast to the iterative approach of [Chen et al., 2009], they replaced components of the optimal weight, which depend on the true covariance matrix, with unbiased estimators as an approximation. Their main proposed estimator, SS, uses the target $\text{diag}(S)$ as a compromise between the constant variance target of [Ledoit and Wolf, 2003] and the constant correlation target of [Elton and Gruber, 1973]. Besides its application to gene association networks, this method has gained popularity in forecast reconciliation literature. For instance, [Wickramasuriya et al., 2019]'s MinT method used SS to estimate the base forecast error covariance matrix, and it has since been adopted in other probabilistic forecast reconciliation studies [Panagiotelis et al., 2023]. Despite its popularity, the estimator's distribution-free property comes with a cost. As it is a ratio of unbiased components, the resulting estimator remains biased and inconsistent. Our simulations show that this estimator performs well only under highly sparse settings. Moreover, manual clipping of the estimated weights between 0 and 1 is necessary to ensure the estimated covariance matrix is invertible, which further distorts its finite-sample performance.

In recent years, new methods have been developed for high-dimensional covariance matrix estimation, including the factor model approach by [Fan et al., 2008] and the non-linear

shrinkage method by [Ledoit and Wolf, 2012]. The former assumes a factor structure rather than sparsity in the true covariance matrix. The latter asymptotically bias-corrects the sample eigenvalues using the [Marchenko and Pastur, 1967] equation while keeping the sample eigenvectors intact. However, [Ledoit and Wolf, 2012] have shown that when $\frac{p}{n}$ is large or the dispersion of eigenvalues is small, the non-linear shrinkage estimator does not significantly outperform LW. Our own verification supports this finding.

This paper builds on the linear shrinkage literature, particularly in empirical situations where using a diagonal target seems appropriate (e.g., when variables have vastly different variances) and when $p \gg n$. Through simulation studies, we demonstrate that our proposed estimator, $OASD$, performs best among competing estimators when the population covariance matrix exhibits considerable variation. Moreover, our extended estimator, $OASB$, performs the best across all simulation scenarios.

3 Theoretical Framework

Suppose that the data $\{x_i\}_{i=1}^n$ are *i.i.d.* and has $p \geq 2$ dimensions. In a high-dimensional environment $p > n$, 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 the 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 the OAS targets the average mean $T = \frac{tr(S)}{p} I$. 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

$$a'\hat{S}(\rho)\,a = (1-\rho)\underbrace{a'Sa}_{\geq 0} + \rho\underbrace{a'Ta}_{>0} > 0, \quad \forall a \neq 0, \quad \rho \in (0,1]\,. \tag{3}$$

When the true covariance matrix $\Sigma = V(x_i)$ is known, the shrinkage parameter ρ can be pinned down by minimizing the MSE from the true covariance matrix

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

where the resulting shrinkage parameter ρ_{OD} is called an Oracle shrinkage 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*

$$\rho_{OD}(\Sigma, T) = \frac{E\left[tr(\Sigma - S)(T - S)\right]}{E\left[\|T - S\|^2\right]}. \tag{5}$$

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

$$\rho_{OD}(\Sigma) = \frac{tr(\Sigma^2) - 2tr(diag(\Sigma)^2) + tr(\Sigma)^2}{ntr(\Sigma^2) - (n+1)tr(diag(\Sigma)^2) + tr(\Sigma)^2}. \tag{6}$$

Proof. See Appendix 7.1. ■

The oracle shrinkage parameter of (6) is optimal but infeasible, since it is based on the true covariance matrix Σ. To approximate (6), our proposed method, which we call Oracle Approximating Shrinkage with Diagonal target ($OASD$), uses the limit of the following iteration indexed by j

$$\Sigma_j = (1-\rho_j)S + \rho_j diag(S), \tag{7}$$

$$\rho_{j+1} = \frac{tr(\Sigma_j S) - 2tr(diag(\Sigma_j)^2) + tr(\Sigma_j)^2}{ntr(\Sigma_j S) - (n+1)tr(diag(\Sigma_j)^2) + tr(\Sigma_j)^2}. \tag{8}$$

Note that (8) 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 $\Sigma_j S$. In this way, the system remains tractable since ρ_j^2 does not show up.

The following main theorem shows that the iteration converges to a unique limit for any 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 (7) and (8) converges to*

$$\rho_{OASD} = \min\left\{\frac{1}{n\phi}, 1\right\}, \quad \phi = \frac{tr(S^2) - tr(diag(S)^2)}{tr(S^2) + tr(S)^2 - 2tr(diag(S)^2)} \in [0,1). \tag{9}$$

The shrinkage parameter satisfies $\rho_{OASD} \in (0,1]$, and thus, the covariance estimator $S_{OASD} = \hat{S}(\rho_{OASD})$ is positive definite.

Proof. See Appendix 7.2 ■

3.1 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 [Schäfer and Strimmer, 2005]), and thus allows us to compare the performance of different methods, although it is less useful in practice than the general setup with unknown mean.

It turns out that the resulting formula replaces n in (6) and (9) 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*

$$S = \frac{1}{n}\sum_{i=1}^{n} x_i x_i^T. \tag{10}$$

Then, the Oracle and OASD shrinkage estimators (6) and (9) are replaced by

$$\rho_{OD}(\Sigma) = \frac{tr(\Sigma^2) - 2tr(diag(\Sigma)^2) + tr(\Sigma)^2}{(n+1)tr(\Sigma^2) - (n+2)tr(diag(\Sigma)^2) + tr(\Sigma)^2}, \tag{11}$$

$$\rho_{OASD} = \min\left\{\frac{1}{(n+1)\phi}, 1\right\}, \quad \phi = \frac{tr(S^2) - tr(diag(S)^2)}{tr(S^2) + tr(S)^2 - 2tr(diag(S)^2)} \in [0,1). \tag{12}$$

The shrinkage parameter satisfies $\rho_{OASD} \in (0,1]$, so the covariance estimator $S_{OASD} = \hat{S}(\rho_{OASD})$ is positive definite.

Proof. See Appendix 7.3 ■

As in Theorem 2, the shrinkage parameter ρ_{OASD} contains min operator, but this is a result of the convergence and is not manually imposed.

3.2 Extension: Two Targets

Instead of manually selecting either to use the diagonal target as in $OASD$ or the average variance target as in OAS, we propose the updated shrinkage estimator

$$\tilde{S}(\theta, \alpha) = (1-\theta) S + \theta (\alpha T_{OASD} + (1-\alpha) T_{OAS}). \tag{13}$$

where $T_{OASD} = diag(S)$ and $T_{OAS} = \frac{tr(S)}{p} I$. Similar to (4), the shrinkage parameters θ and α can be found by minimizing the MSE from the true covariance matrix

$$\theta_{OB}(\Sigma), \alpha_{OB}(\Sigma) = \arg\min_{\rho \in \mathbb{R}} E\left[\left\|\tilde{S}(\theta, \alpha) - \Sigma\right\|^2\right]. \tag{14}$$

where the resulting θ_{OB} and α_{OB} are called the Oracle shrinkage estimators with two targets. This problem (14) is again quadratic in both parameters and we can get the following closed-form solution.

Theorem 4 *Suppose that x_i is i.i.d. and follows a joint normal distribution $N(\mu, \Sigma)$. S is the sample covariance matrix (1). $T_{OASD} = diag(S)$ a diagonal matrix that shares the same diagonal elements as S and $T_{OAS} = \frac{tr(S)}{p} I$ is a diagonal matrix with each element equal to the averaged sample variances. The optimal shrinkage parameters that solves (14) is*

$$\theta_{OB}(\Sigma) = \frac{tr(\Sigma)^2 + tr(\Sigma^2) - 2tr(diag(\Sigma)^2)}{ntr(\Sigma^2) + tr(\Sigma)^2 - (n+1)tr(diag(\Sigma)^2)}, \tag{15}$$

$$\alpha_{OB}(\Sigma) = 1 - \frac{1}{\theta_{OB}} \frac{2ptr(diag(\Sigma)^2) - 2tr(\Sigma^2)}{p(n+1)tr(diag(\Sigma)^2) - 2tr(\Sigma^2) - (n-1)tr(\Sigma)^2}. \tag{16}$$

Proof. See Appendix 7.4. ■ This oracle estimator is again infeasible, as it depends on the true covariance matrix Σ. A convenient feature of the above theorem is that θ_{OB} coincides with ρ_O and therefore we can use ρ_{OASD} as an estimator for θ_{OB}. For α_{OB}, we can adopt the oracle approximating strategy used in Theorem 2 and estimate α_{OB} using the limit of

the following iteration indexed by j (using θ_{OASB} in place of θ)

$$\Sigma_j = (1-\theta_{OASB})S + \theta_{OASB}(\alpha_j T_{OASD} + (1-\alpha_j)T_{OAS}), \tag{17}$$

$$\alpha_{j+1} = 1 - \frac{1}{\theta_{OASB}} \frac{2ptr(diag(\Sigma_j)diag(S)) - 2tr(\Sigma_j S)}{p(n+1)tr(diag(\Sigma_j)diag(S)) - 2tr(\Sigma_j S) - (n-1)tr(\Sigma_j)^2}. \tag{18}$$

The following theorem shows that the above iteration converges to a unique limit regardless of the initial value α.

Theorem 5 *For any initial value α_0 , the sequence $\{\alpha_j\}_j$, specified by (17) and (18) converges to*

$$\alpha_{OASB} = \begin{cases} \dfrac{\theta_{OASB}-1}{\theta_{OASB}}, & if \left| \dfrac{\tau_3}{(\tau_1+\tau_2)\theta_{OASB}+1-\tau_1} \right| < 1 \\ \dfrac{\theta_{OASB}(\tau_3-\tau_2)-1}{\theta_{OASB}(\tau_1+\tau_3)}, & if \left| \dfrac{\tau_3}{(\tau_1+\tau_2)\theta_{OASB}+1-\tau_1} \right| \geq 1, \end{cases} \tag{19}$$

where we use the following

$$\theta_{OASB} = \min\left\{\frac{1}{n\phi}, 1\right\}, \quad \phi = \frac{tr(S^2) - tr(diag(S)^2)}{tr(S^2) + tr(S)^2 - 2tr(diag(S)^2)}, \tag{20}$$

$$\tau_1 = \frac{(p-1)\left[tr(diag(S)^2) - \frac{tr(S)^2}{p}\right]}{ptr(diag(S)^2) - tr(S^2)}, \tag{21}$$

$$\tau_2 = \frac{tr(S^2) - tr(diag(S)^2) - (p-1)\left[tr(diag(S)^2) - \frac{tr(S)^2}{p}\right]}{ptr(diag(S)^2) - tr(S^2)}, \tag{22}$$

$$\tau_3 = \frac{(n-1)p\left[tr(diag(S)^2) - \frac{tr(S)^2}{p}\right]}{2(ptr(diag(S)^2) - tr(S^2))}. \tag{23}$$

The shrinkage parameters now satisfy $\theta_{OASB} \in (0,1]$ and $\alpha_{OASB} \in [1 - \frac{1}{\theta_{OASB}}, 1)$, and we can establish that the covariance estimator $S_{OASB} = \tilde{S}(\theta_{OASB}, \alpha_{OASB})$ is positive definite.

Proof. See Appendix 7.5. ■

3.2.1 Special Case: Two Targets with Known Mean

Theorem 6 *Suppose that x_i is i.i.d. and follows a joint normal distribution $N(0,\Sigma)$. S is now replaced by the sample covariance matrix assuming known mean(10). The oracle*

shrinkage parameters in 16 can now be replaced by

$$\theta_{OB}(\Sigma) = \frac{tr(\Sigma)^2 + tr(\Sigma^2) - 2tr(diag(\Sigma)^2)}{(n+1)tr(\Sigma^2) + tr(\Sigma)^2 - (n+2)tr(diag(\Sigma)^2)}, \tag{24}$$

$$\alpha_{OB}(\Sigma) = 1 - \frac{1}{\theta_{OB}} \frac{2ptr(diag(\Sigma)^2) - 2tr(\Sigma^2)}{p(n+2)tr(diag(\Sigma)^2) - 2tr(\Sigma^2) - ntr(\Sigma)^2}. \tag{25}$$

The approximating OASB shrinkage estimators are specified as

$$\theta_{OASB} = \min\left\{\frac{1}{(n+1)\phi}, 1\right\}, \quad \phi = \frac{tr(S^2) - tr(diag(S)^2)}{tr(S^2) + tr(S)^2 - 2tr(diag(S)^2)}, \tag{26}$$

$$\alpha_{OASB} = \begin{cases} \dfrac{\theta_{OASB} - 1}{\theta_{OASB}}, & \text{if } \left|\dfrac{\tau_3}{(\tau_1+\tau_2)\theta_{OASB} + 1 - \tau_1}\right| < 1 \\ \dfrac{\theta_{OASB}(\tau_3 - \tau_2) - 1}{\theta_{OASB}(\tau_1 + \tau_3)}, & \text{if } \left|\dfrac{\tau_3}{(\tau_1+\tau_2)\theta_{OASB} + 1 - \tau_1}\right| \geq 1, \end{cases} \tag{27}$$

where τ_1 and τ_2 are as specified in 21 and 22 and τ_3 is defined as follows

$$\tau_3 = \frac{np\left[tr(diag(S)^2) - \frac{tr(S)^2}{p}\right]}{2(ptr(diag(S)^2) - tr(S^2))}. \tag{28}$$

The shrinkage parameters now satisfy $\theta_{OASB} \in (0, 1]$ and $\alpha_{OASB} \in [1 - \frac{1}{\theta_{OASB}}, 1)$, and we can establish that the covariance estimator $S_{OASB} = \tilde{S}(\theta_{OASB}, \alpha_{OASB})$ is positive definite.

Proof. See Appendix 7.6. ■

4 Simulation

This section uses simulations to assess the performance of S_{OASD} and S_{OASB} in a high-dimensional environment with large variation in the diagonal elements of the true covariance matrix Σ. Simulations consider different degrees of variation and sparsity of the true correlation matrix, as well as different sample sizes. The S_{OASD} and S_{OASB} both perform reasonably well with S_{OASB} having a small advantage over S_{OASD} in these settings.

4.1 Setting

To conduct simulations in a high-dimensional environment, we 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 $\Gamma_{kl} = \gamma^{|k-l|}$, 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 a different unit,

$$\Sigma = \Lambda\Gamma\Lambda, \quad \Lambda = \Lambda^T = \begin{bmatrix} 1 & & & & & 0 \\ & \ddots & & & & \\ & & 1 & & & \\ & & & sd & & \\ & & & & \ddots & \\ 0 & & & & & sd \end{bmatrix}, \tag{29}$$

where the parameter sd varies from 1 to 20. Large variations of 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 $B = 1000$ times. The performance criterion is the percentage relative improvement in average loss (PRIAL), defined as

$$PRIAL(\hat{S}) = \left(1 - \frac{\sum_{b=1}^{B} \left\| \hat{S}^{(b)} - \Sigma \right\|^2}{\sum_{b=1}^{B} \left\| S^{(b)} - \Sigma \right\|^2}\right) \times 100, \tag{30}$$

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 of $\hat{S}$ relative to the sample covariance matrix S.

To assess the performance of S_{OASD} and S_{OASB}, we compare them under the known mean assumption, as derived in (3 and 6), with three methods in the literature, most of which

[1] We set $\Gamma_{kl} = 1$ when $\gamma = 0$ and $k = l$.

also assume the known mean in their derivations [2]. First, we denote by LW the estimator proposed by [Ledoit and Wolf, 2004]

$$S_{LW} = (1-\rho_{LW})S + \rho_{LW}\frac{tr(S)}{p}I, \quad \rho_{LW} = \min\left\{\frac{\sum_{i=1}^{n}\left\|x_i x_i^T - S\right\|^2}{n^2\left[tr(S^2) - \frac{tr(S)^2}{p}\right]}, 1\right\}. \tag{31}$$

Second, we denote by OAS and $RBLW$ the estimators proposed by [Chen et al., 2009][3]

$$S_{OAS} = (1-\rho_{OAS})S + \rho_{OAS}\frac{tr(S)}{p}I, \quad \rho_{OAS} = \min\left\{\frac{\left(1-\frac{2}{p}\right)tr(S^2) + tr(S)^2}{\left(n+1-\frac{2}{p}\right)\left[tr(S^2) - \frac{tr(S)^2}{p}\right]}, 1\right\}. \tag{32}$$

$$S_{RBLW} = (1-\rho_{RBLW})S + \rho_{RBLW}\frac{tr(S)}{p}I, \quad \rho_{LW} = \min\left\{\frac{\frac{n-2}{n}tr(S^2) + tr(S)^2}{(n+2)[tr(S^2) - \frac{tr(S)^2}{p}]}, 1\right\}. \tag{33}$$

Third, we denote by SS the estimator proposed by [Schäfer and Strimmer, 2005]

$$S_{SS} = (1-\rho_{SS})S + \rho_{SS}diag(S), \quad \rho_{SS} = \min\left\{\frac{\sum_{m\neq k}\widehat{Var}(r_{mk})}{\sum_{m\neq k} r_{mk}^2}, 1\right\}, \tag{34}$$

where r_{mk} is the (m,k) element of the sample correlation matrix and $\widehat{Var}(r_{mk})$ is the sample variance estimator of r_{ij}. 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 S_{OASD} and S_{OASB} with their respective Oracle estimators $S_{OD} = \hat{S}(\rho_{OD})$ and $S_{OB} = \tilde{S}\left(\theta_{OASB}, \alpha_{OASB}\right)$.

In summary, we compare 8 estimators, $\{S_{OASD}, S_{OD}, S_{OASB}, S_{OB}, S_{LW}, S_{RBLW}, S_{OAS}, S_{SS}\}$, varying the three parameters $\{n, sd, \gamma\}$ that control the sample size, the scale differences of 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.

4.2 Simulation results

The following subsections demonstrate that, compared to other methods, S_{OASB} and S_{OASD} exhibit a higher $PRIAL$ and that both the shrinkage parameters α_{OASB} and ρ_{OASD} tracks

[2] with the exception of SS, which we derived their known mean version and used it in the comparison

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

the infeasible Oracle estimator ρ_{OB} and ρ_{OD} close in all three dimensions $\{n, sd, \gamma\}$. Note that because $\theta_{OASB} = \rho_{OASD}$, we only included α_{OASB} (denoted as $OASB$) in the plot for the shrinkage parameters.

4.2.1 Variation of scales

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

For most regions of the variation parameter sd except for areas with small sd, the $OASB$ and $OASD$ exhibit higher $PRIAL$ than the average variance methods LW, $RBLW$ and OAS. This is not surprising since the larger the variation of scales, the closer the diagonal target is to the true covariance matrix compared with the average variance target. Interestingly, when the variation of scales sd is small, the shrinkage parameter is similar to the average variance methods.

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

$OASB$ is the best among all methods and improves on $OASD$ by around 1%. The improvement comes from the use of the additional average variance target. As can be seen on the right chart of Figure 1, the weight on the diagonal target is negative when sd is close to 1 but increases to 0.75 when sd is 3, demonstrating the necessity of having a diagonal target when the variation of diagonal elements is large.

All methods exhibit a lower $PRIAL$ as the variation of scales increases. This is because the off-diagonal elements of the true covariance matrix Σ are larger, and therefore the approximation of 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.

4.2.2 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 $OASB$ and $OASD$ exhibit a higher $PRIAL$ than all other methods. The improvement compared to the average variance targets, LW and OAS, can be as large as 10% when the true covariance matrix Σ is sparser. One way to understand this comparative static is to consider the limit case $\gamma \to 0$, where the true covariance matrix Σ is diagonal. Both $OASB$ and $OASD$ 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 6%. The difference in $PRIAL$ is similar when the true covariance matrix Σ is sparse, but the difference increases as the sparsity decreases. This can be attributed to the better approximation of the shrinkage parameter ρ_{OASD} to the oracle weight ρ_{OD} compared to ρ_{DD}, as can be seen in the right chart of Figure 2.

$OASB$ has up to 1% improvement over $OASD$ in this experiment[4]. The difference between the two can be explained by the decrease of α_{OASB} as γ increases. From the right chart for shrinkage parameters, we can see that the reduction in α_{OASB} is more slowly than that of θ_{OASB}. This means the as we shrink the off-diagonal elements less, we also shrink the diagonal terms less, albeit at a slower rate. This is due to a combination of less shrinkage needed as the true covariance is denser, and the increased estimation burden with a limited sample size.

4.2.3 Sample Size

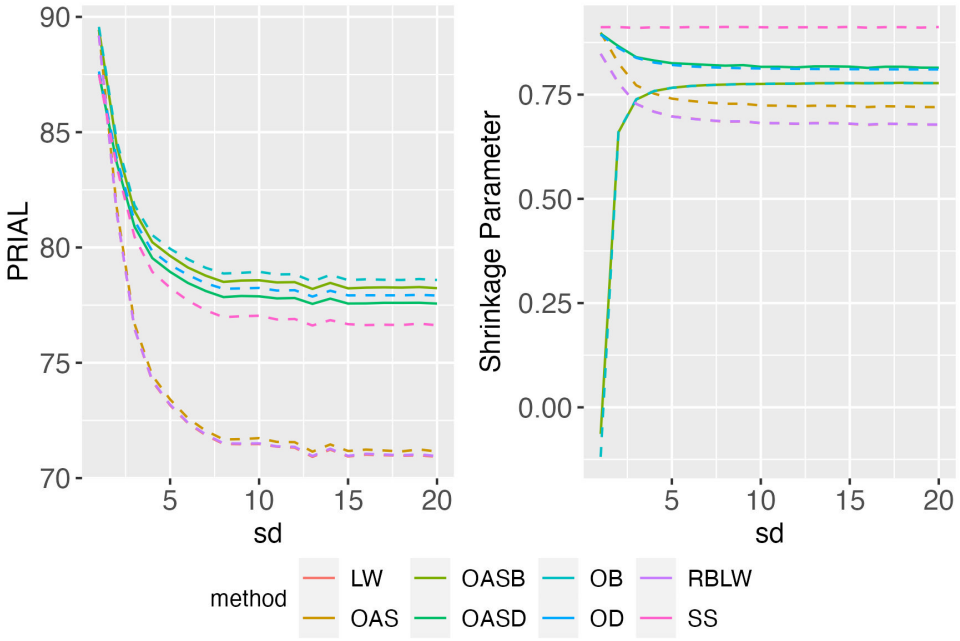
Figure 3 shows the $PRIAL$ on the left and average shrinkage parameters on the right for each method over the sample size n.

When we increase n while fixing p at 100, we find from Figure 3 that all methods show a decreasing $PRIAL$. This is because the sample variance is converging to the true covariance matrix and dampening the additional benefits from shrinkage methods.

$OASB$ and $OASD$ perform best over all sample size n. On average, the $PRIAL$ of $OASD$ is 5% higher than LW and OAS. The difference increases as the sample size n increases. This is because as sample size increases, we can afford to estimate more parameters accurately and thus a diagonal target would be closer to the true covariance matrix than an average

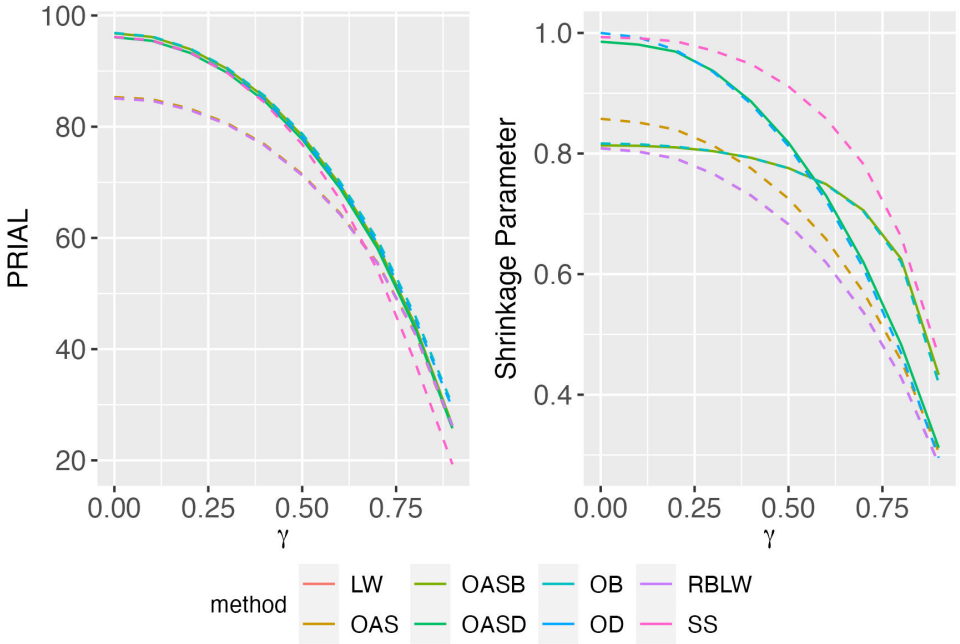
[4]This isn't very noticeable in the chart due to the scale of the y axis

Figure 1: $PRIAL$ and shrinkage parameters with different variation of scales sd



*Note the above results are generated under $p = 100$, $n = 18$, $\gamma = 0.5$

Figure 2: $PRIAL$ and shrinkage parameters with different correlation sparsity γ



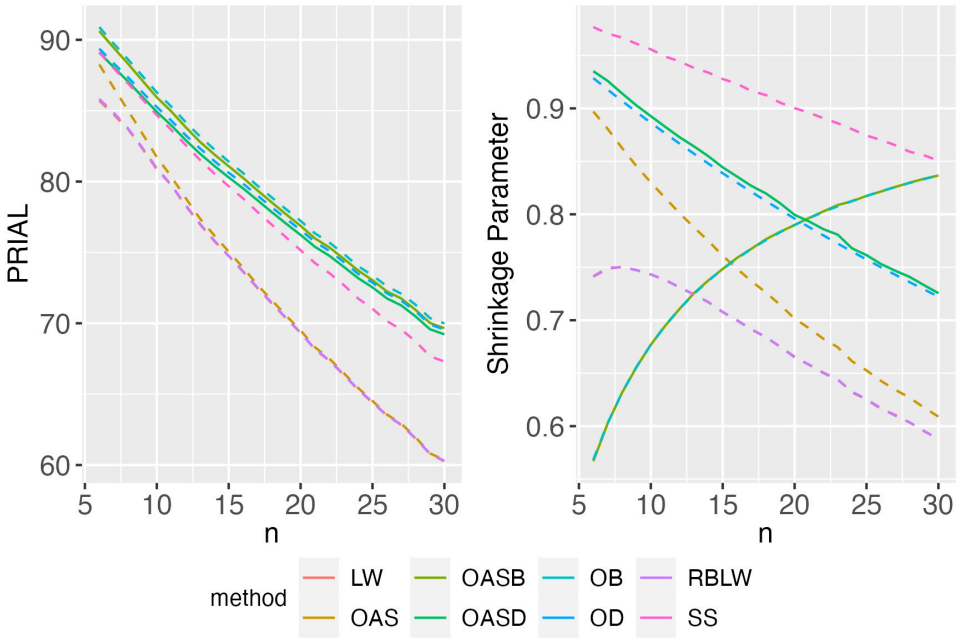
*Note the above results are generated under $p = 100$, $n = 18$, $sd = 10$

variance target.

Compared with SS, the $PRIAL$ of $OASD$ is on average 2 percent higher and the difference also widens as n increases. This can be explained by the worsened approximation of SS to the oracle weight, as can be seen in the right chart of Figure 3 while $OASD$ tracks the oracle weight ρ_{OD} closer for all sample sizes.

$OASB$ is better than $OASD$, with larger difference at small sample sizes. This is a result of the flexibility of shrink the diagonal terms as well as the off-diagonal terms when sample size is small. As n increases, the weight on the diagonal target increases as can be seen from the right chart of Figure 3, and the additional advantage of $OASB$ over $OASD$ decreases.

Figure 3: $PRIAL$ and shrinkage parameters with different sample sizes



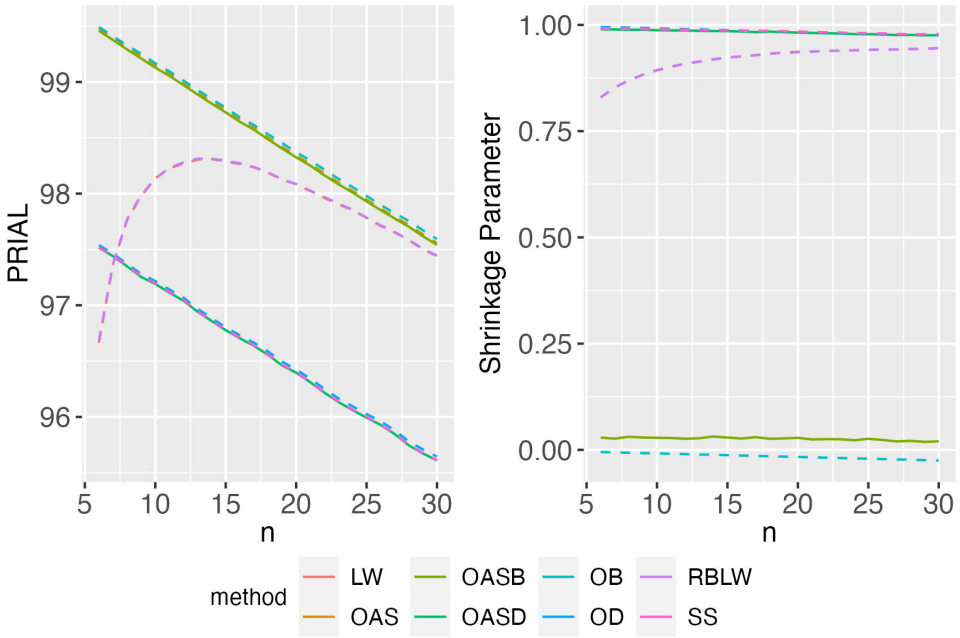
*Note the above results are generated under $p = 100$, $sd = 10$, $\gamma = 0.5$

4.2.4 Scenario for the better performance of the average variance target

$OASD$ works better than estimators using an average variance target under the above discussed scenarios. However, it is expected that when the variation of scales is small and the true correlation matrix is sparse, there's no trade-off between shrinking diagonal and off-diagonal elements and therefore the latter group of estimators should perform better. It can be seen from Figure 4 that this is indeed the case. Under this scenario, our $OASB$

resembles the performance of OAS, both of which are similar to LW-based methods when n is large. $OASB$'s $PRIAL$ is higher than that of $OASD$ by 2% for all sample sizes. The shrinkage parameters plot shows that the parameters of $OASB$ rightly chose to put almost all weight on the average variance target.

Figure 4: $PRIAL$ and shrinkage parameters with different sample sizes (for small variation in scale and large correlation sparsity)



*Note the above results are generated under $p = 100$, $sd = 1$, $\gamma = 0.2$

4.3 Validity of the iterative approach

In Figure 5, we demonstrate that our derived analytical limit is indeed the converging limit when we keep iterating the optimal value for $\hat{\rho}_j$ following the updating rule defined in Equation 8. The blue lines represent the RMSE of $\hat{\rho}_j$ at a round of iteration j, under different parameter settings, with darker color representing its value after more iterations. The red line represents the derived $\hat{\rho}_{oas_v}$ under different parameter settings. We can notice that the convergence happen rather quickly, with around 10 iterations we already get pretty close to the limiting value. This can serve as motivations for extending the iterative appraoch to other useful shrinking targets without the need to derive their analytical limits.

4.4 Eigenvalues of OASD

It's beneficial to reduce the dispersion of the sample eigenvalues for the following three reasons. First, the sample covariance matrix when $p > n$ is non-invertible and has $\lambda_{min}(S) = 0$. This is problematic, as its inverse is intractable. Therefore, we would like to increase the minimum eigenvalue in our combined estimator. Moreover, as demonstrated in [Ledoit and Wolf, 2004] Lemma 2.1, the eigenvalues of the sample covariance matrix suffer from overdispersion due to the following decomposition:

$$E\left[||S-\mu||_F^2\right] = E\left[||S-\Sigma||_F^2\right] + ||\Sigma-\mu I||_F^2,$$

where $\mu = \frac{tr(\Sigma)}{p}$. This leads to:

$$E\left[\Sigma_{i=1}^p(\lambda_i(S)-\mu)^2\right] = E\left[\|S-\Sigma\|_F^2\right] + \Sigma_{i=1}^p(\lambda_i(\Sigma)-\mu)^2.$$

Thus to reducing the dispersion of sample eigenvalues allows us to better approximate the true eigenvalues. Lastly, it is known that if a matrix is better conditioned (ie $\frac{\lambda_{max}}{\lambda_{min}}$ is smaller), inverting the estimated sample covariance matrix will lead to less estimation error for the precision matrix (Σ^{-1}). The following comment proves that our combined estimator reduces the range of eigenvalues compared with sample covariance matrix and, therefore, is invertible, offers better approximation to population eigenvalues, and is better conditioned.

Comment 1

For any estimator S_c in the form of $(1-\rho)S + \rho diag(S)$, where $\rho \in (0, 1]$. This estimator S_c has the following property under mild conditions:

$$\lambda_{min}(S) < \lambda_{min}(S_c) < \lambda_{max}(S_c) < \lambda_{max}(S)$$

By shrinking the sample covariance target towards $diag(S)$, we effectively make it invertible, better conditioned, and potentially closer to the true eigenvalues now that they are less dispersed. Unlike shrinking all eigenvalues towards their grand mean as achieved by a target of $\frac{tr(S)}{p}$ in [Ledoit and Wolf, 2003], our level of shrinkage is milder and would be useful for scenarios where we expect a lot of variation in true eigenvalues (for example, the case of different units of measurement)[5].

[5]This level of shrinkage isn't generally achievable by varying the shrinkage coefficient in [Ledoit and Wolf, 2003] as their linear shrinkage estimator assumes common shrinkage for all sample eigenvalues

4.5 Inverse of the covariance estimators

In this section, we compare the performances of the inverse of each covariance estimator in estimating the inverse of the true covariance matrix. This is relevant as most empirial analysis directly uses inverse of the covariance matrix (precision matrix) estimator rather than the covariance matrix itself. Our benchmark is the widely-adopted Moore-Penrose Inverse and therefore to faciliate comparison, we define our metric $PRIAL_{INV}$ as the percentage relative improvement in average loss against the MSE of the Moore-Penrose Inverse Φ_{MP}.

$$PRIAL_{INV}(\hat{S}) = \left(1 - \frac{\sum_{b=1}^{B} \left\| \hat{S}^{(b)^{-1}} - \Sigma^{-1} \right\|^2}{\sum_{b=1}^{B} \left\| \Phi_{MP}^{(b)} - \Sigma^{-1} \right\|^2}\right) \times 100, \tag{35}$$

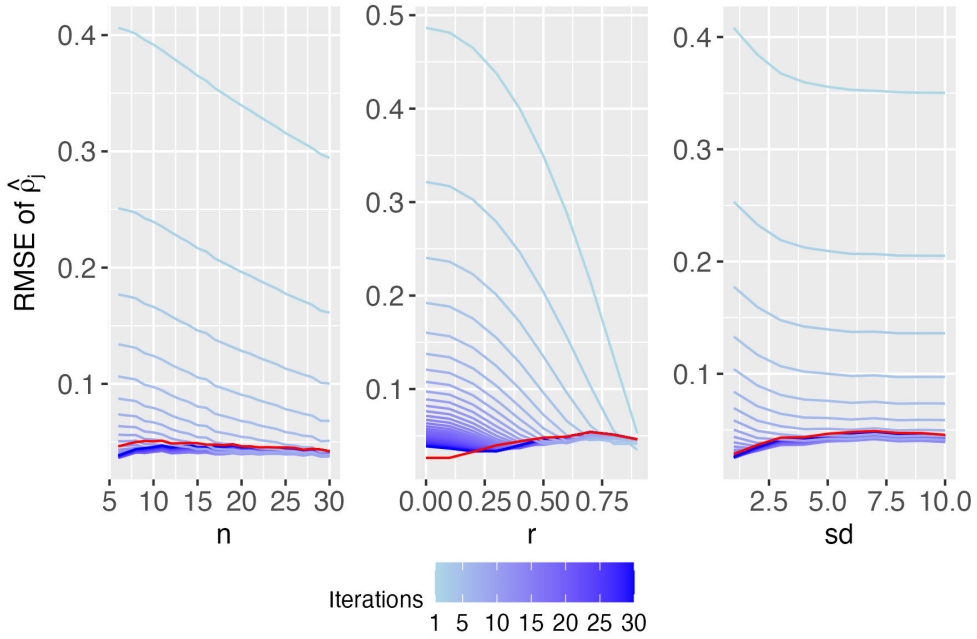
Repeating the three sets of experiment above, we can see that $OASB$ and $OASD$ improve the estimation error of the inverse of the covariance matrix the most.

Figure 6 shows that the inverse of $OASD$ tends to perform better than other methods. Intuitively, suppose that 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. This may also be related to the observation in the previous section that $OASD$ is well-conditioned.

4.5.1 Alternative methods based on shrinking correlation matrix

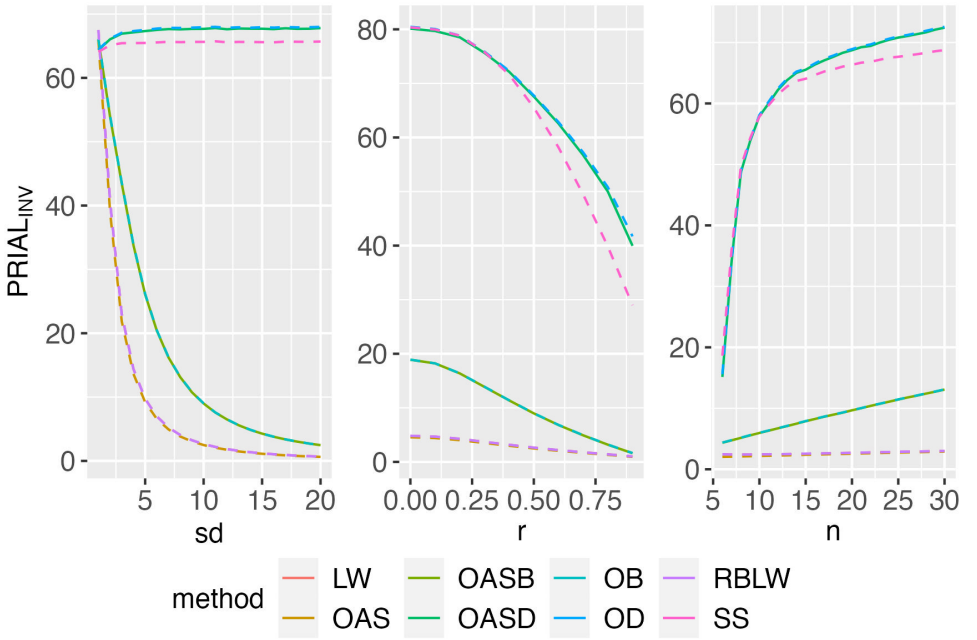
One notable problem with using an average variance target is the neglience of different variable scales. This motivates scaling the variables first, or equivalently applying shrinkage only to the correlation matrix before multiplying back the sample variances. This is expected to have finite-sample problems because unless the sample variances are highly accurate, approximating the true covariances with sample variances does not give the same optimal weight as approximating the true correlation with sample correlation. This optimality of directly minimizing the distance to the true covariance matrix is shown in Figure 7. $OASB$ is still the best and approximates OD well in all 3 dimensions. The general patterns when compared with other methods are similar to those discussed in the previous section. However, the adjustments in scaling make the methods using an average-variance target perform closer to those using a diagonal target.

Figure 5: Iterated MSE of $\hat{\rho}_j$ and convergence to the MSE of ρ_{OASD}



*Note the above results are generated by varying one dimension and keeping the other two dimensions at their median. The red line shows MSE of $\hat{\rho}_{oas_v}$

Figure 6: $PRIAL_{INV}$ of all methods along each dimension



5 Application 1: Inflation nowcasting using forecast reconciliation

Inflation nowcasting is crucial for the everyday decision-making processes of policymakers, market practitioners, and consumers. Central banks aim for price stability, requiring them to monitor short-term inflation movements for more effective monetary policy decisions. Market participants and consumers also frequently adjust their investment and consumption plans based on up-to-date inflation rates and their expectations. This is particularly relevant in periods of economic volatility, where prices may change rapidly.

The all-items inflation rate is the headline metric widely followed by markets. Its sub-components, however, are equally important, guiding business decisions in specific industries—such as housing, energy, and food—while offering policymakers insights into potential distributional effects and the development of targeted policy initiatives.

Macroeconomic institutions are tasked with releasing forecasts for these variables, working toward two main objectives. Firstly, they aim to ensure that the forecasts for each variable are as accurate as possible, using appropriate models and predictors. Secondly, they seek to maintain consistency between the forecasts, ensuring coherence among various sub-components and the aggregate inflation rate, as well as between variables at different frequencies. Achieving this balance is essential for presenting a coherent and accurate picture of inflationary trends to policymakers, investors, and the public.

To meet both objectives, we adopt a two-step forecasting procedure. In the first step, a forecast model and set of predictors are chosen for each variable based on historical performance. In the second step, a cross-temporal forecast reconciliation method from [Di Fonzo and Girolimetto, 2023] is adopted to ensure that the forecasts are coherent across components and frequencies. This method minimizes information loss during the aggregation or disaggregation of forecasts for reconciliation purposes, allowing each variable to be forecasted independently and reconciled at the end.

Beyond satisfying practical constraints, this approach has the added benefit of leveraging the information extracted from one variable to improve the forecast performance of another during reconciliation. Previous literature has shown that predictive models for disaggregated series capture data heterogeneity and pick up different dynamics in seasonality and short-term changes [Bermingham and D'Agostino, 2014], [Espasa et al., 2002], [Boaretto and Medeiros, 2023], [Capistrán et al., 2010], and [Ibarra, 2012]. Time series data

for variables with higher frequencies are also much longer than those of lower-frequency aggregated variables, leading to better forecast model estimation. Therefore, disaggregated series can potentially improve forecasts of aggregated series. Information may also flow in the opposite direction: aggregated series, such as the all-items inflation rate, respond to broader economic trends and tend to be less noisy than their sub-components. Annual inflation rates smooth out transitory fluctuations in monthly rates, capture lagged effects, and incorporate slower adjustments in wage and price variables. As a result, they may show a stronger relationship with other macroeconomic variables, as suggested by theories like the Phillips curve.

The key parameter in implementing forecast reconciliation is the covariance matrix of base forecast errors, which guides the decision of how much deviation from the base forecasts is optimal. When a variable is predicted accurately with a small forecast error variance, we tend to adjust it less and instead focus on modifying variables with larger forecast errors to satisfy the constraints. Estimating the covariance matrix poses a challenge due to the large number of variables and the small sample sizes. This feature renders the sample covariance matrix unsuitable, and past literature has resorted to shrinkage methods, predominantly SS [Panagiotelis et al., 2023], [Wickramasuriya et al., 2019], [Di Fonzo and Girolimetto, 2023]. The consensus has been to adopt a diagonal target due to the large variation in scale between different levels of aggregation, but few alternatives for covariance estimation using a diagonal target exist. Therefore, we propose using our proposed S_{OASD} and S_{OASB} estimators in this empirical application to test their forecast MSE.

5.1 Data

5.1.1 Variables

We work with monthly UK CPI data from 1988 to 2021 from ONS. This dataset includes the aggregated All-items CPI (00_IX) and its 12 subcomponents (their detailed definitions are provided in 1). Our goal is to generate base and reconciled forecasts (explained later) for future monthly and annual inflation rates of the current year, and compare their performances with the IMF's official World Economic Outlook (WEO) forecast for the inflation rate. To ensure a fair comparison, we generate our forecasts based on the information IMF economists have when they make their biannual forecasts in April and October[6]. Other

[6]For the April forecast, we use information until February, and for the October forecast, we use information until August each year.

than inflation-related variables, we collect conventional monthly predictor variables from FRED, representing information from price and money supply, production and sales, employment, interest rates, exchange rates, and business and consumer confidence. It's worth noting that we also use two unconventional predictors of inflation rates: annual commodity price projections estimated with futures data from the IMF (GAS) and monthly 5, 20, and 30-year breakeven inflation rates. Both measures represent implicit public opinions of price movements. We aim to include private forecasts like these because they may account for information that is not available to us [Faust and Wright, 2013].

5.1.2 Constraints

Due to the nature of our data (i.e., a mixture of All-items inflation and its subgroups, as well as inflation at different time frequencies), we face cross-temporal constraints. The temporal constraints work as expected, with annual prices being simple averages of their monthly counterparts. However, the way ONS constructs the All-items price index from its subgroups is not as straightforward. Because the UK consumption basket is updated twice a year[7], the weights of different subgroups of the price index must account for the new weights and the change in the price base period used in the construction of the weights. The final relationship between the aggregate All-items index and its components forms our cross-sectional constraints.

Formulating these cross-temporal constraints gives the following:

Let I_y denote the All-items price index for year y and its monthly counterpart as $I_{y,m}$ for month $m \in \{1, \dots, 12\}$. The component weight and index are $W_{y,m,i}$ and $I_{y,m,i}$ for component $i \in I$. The annual component index is denoted $I_{y,i}$.

The temporal constraints are:

$$I_y = \frac{1}{12} \sum_{m=1}^{12} I_{y,m}, \quad I_{y,i} = \frac{1}{12} \sum_{m=1}^{12} I_{y,m,i}. \tag{36}$$

The cross-sectional constraints are: When $m = 1$:

$$I_{y,1} = \sum_i w_{y,1,i} I_{y,1,i}, \quad w_{y,1,i} = \frac{I_{y-1,12}}{I_{y-1,12,i}} \frac{W_{y,1,i}}{\sum_j W_{y,1,j}}. \tag{37}$$

[7]Once in December to conform with the regulatory update of the COICOP weights, and once in January to agree with the price reference period of RPI. See [ONS,]

When $m \geq 2$:

$$I_{y,m} = \sum_i w_{y,m,i} I_{y,m,i}, \quad w_{y,m,i} = \frac{I_{y,1}}{I_{y,1,i}} \frac{W_{y,m,i}}{\sum_j W_{y,m,j}}. \tag{38}$$

Since we are working with inflation rather than CPI data, we need to take care of the transformation to growth rates in our constraints as follows (using i_y and $i_{y,m}$ to denote the annual and monthly All-items inflation rate, $i_{y,i}$ and $i_{y,m,i}$ to denote the annual and monthly inflation rate for component i):

$$i_y = \frac{I_y - I_{y-1}}{I_{y-1}}, \quad i_{y,m} = \frac{I_{y,m} - I_{y-1,m}}{I_{y-1,m}}, \quad i_{y,m,i} = \frac{I_{y,m,i} - I_{y-1,m,i}}{I_{y-1,m,i}}. \tag{39}$$

Based on the constraints on the price indices, the inflation rates satisfy the following constraints:

$$i_y - \sum_{m=1}^{12} \frac{1}{12} \frac{I_{y-1,m}}{I_{y-1}} i_{y,m} = 0, \tag{40}$$

$$i_{y,i} - \sum_{m=1}^{12} \frac{1}{12} \frac{I_{y-1,m,i}}{I_{y-1,i}} i_{y,m,i} = 0, \tag{41}$$

$$i_{y,m} - \sum_i \frac{w_{y,m} I_{y-1,m,i}}{I_{y-1,m}} i_{y,m,i} = \sum_i \frac{I_{y-1,m,i}}{I_{y-1,m}} \left(w_{y,m,i} - w_{y-1,m,i} \right), \tag{42}$$

where the last equation is derived from

$$\begin{aligned} I_{y,m} - I_{y-1,m} &= \sum_i w_{y,m,i} I_{y,m,i} - \sum_i w_{y-1,m,i} I_{y-1,m,i} \\ &= \sum_i w_{y,m,i} \left(I_{y,m,i} - I_{y-1,m,i} \right) + \sum_i \left(w_{y,m,i} - w_{y-1,m,i} \right) I_{y-1,m,i}. \end{aligned}$$

To generate forecasts that are coherent with the cross-temporal constraints discussed above, we first generate base forecasts for each variable. These forecasts are selected from a range of candidate forecast models and predictors using cross-validation. We then reconcile these forecasts using an iterative approach, as suggested in [Di Fonzo and Girolimetto, 2023]. In the following sections, we discuss each step in detail.

5.1.3 Base Forecast Methods

We have different candidate models for annual and monthly variables, primarily due to two reasons: the length of the sample size and the availability of concurrent data. We have 34 years of annual data, and the last 5 years are used for evaluating forecast performance. This means that our training sample size for annual data is as small as 29, which restricts the estimation of some models. The advantage of predicting annual variables is the availability of some higher-frequency variables for the current year, allowing us to use these already-available aggregates as predictors.

We group our variables according to whether they are monthly or annual variables. For monthly variables, we forecast 10 steps ahead if predicting in April (using information until February) and 4 steps ahead if predicting in October (using information until August). For annual variables, we forecast 1 step ahead.

In the following sections, we use $\pi_t^A = [\pi_t^{A\_00}, \pi_t^{A\_01}, \ldots, \pi_t^{A\_12}]^T$, where $t \in [1988, 2021]$, to denote annual inflation rates for all-items and its 12 sub-components. We use $\pi_t^M = [\pi_t^{M\_00}, \pi_t^{M\_01}, \ldots, \pi_t^{M\_12}]^T$, where $t \in [1988\_m1, 2021\_m12]$, to denote monthly all-items inflation rates and its sub-components for all months. To enhance clarity in the following sections, we drop the superscripts A and M, which indicate the level of aggregation.

Random Walk (for both monthly and annual inflation rates) We use the current inflation rate as the forecast for h steps ahead:

$$\hat{\pi}_{t+h|t}^{RW} = \pi_t \tag{43}$$

Rolling Historical Mean (for both monthly and annual inflation rates) For predictions h steps ahead, we use the historical average inflation rates from S periods ago until the current period. S is determined by the shortest sample size at the first forecast point. For example, for 1-step-ahead annual forecasts over the last 5 years, we take S to be 29 (the training sample size for 2017):

$$\hat{\pi}_{t+h|t}^{HM} = \frac{1}{S} \sum_{s=t-S+1}^{t} \pi_s \tag{44}$$

VAR(p) (for monthly inflation rates) We estimate h-step-ahead forecasts recursively using a VAR(p) model, where the order p is chosen by the Bayesian Information Criterion

(BIC). The coefficient matrix ϕ_l is estimated using Ordinary Least Squares (OLS):

$$\hat{\pi}^{AR}_{t+h|t} = \hat{\mu} + \sum_{l=1}^{p} \hat{\phi}_l \tilde{\pi}_{t+h-l} \tag{45}$$

where

$$\tilde{\pi}_{t+h-l} = \begin{cases} \pi_{t+h-l}, & \text{if } t+h-l \leq t \\ \hat{\pi}_{t+h-l|t}, & \text{if } t+h-l > t \end{cases} \tag{46}$$

Augmented VAR (for monthly inflation rates) We include seasonal dummies to account for possible seasonal effects in monthly observations. The following augmented AR model is estimated using OLS:

$$\hat{\pi}^{Aug\_AR}_{t+h|t} = \hat{\mu} + \sum_{l=1}^{p} \hat{\phi}_l \tilde{\pi}_{t+h-l} + \sum_{m=1}^{11} \hat{\delta}_m d_{m,t+h} \tag{47}$$

where

$$\tilde{\pi}_{t+h-l} = \begin{cases} \pi_{t+h-l}, & \text{if } t+h-l \leq t \\ \hat{\pi}_{t+h-l|t}, & \text{if } t+h-l > t \end{cases} \tag{48}$$

and $d_{m,t}$ is a seasonal dummy for month m, with δ_m as the associated seasonal effect coefficient.

Hybrid Philips Curve (for annual inflation rates) Following the expectation-augmented Philips curve in [Galı and Gertler, 1999], and including the additional predictor of exchange rates (which are important for import prices), we get the following hybrid Philips curve forecasting model:

$$\hat{\pi}^{PC}_{t+1|t} = \hat{\mu} + \hat{\eta} \pi^{e}_{t+1|t} + \hat{\psi}_1 g_{t+1} \tag{49}$$

where $\pi^{e}_{t+1|t}$ is the largest principal component of inflation expectation-related variables, including the previous year's inflation and available monthly breakeven inflation rates for the current year. The variable g_{t+1} is the largest principal component of available monthly GDP growth rates for the current year, and e_{t+1} is the largest principal component of the available monthly GBP to USD exchange rate for the current year. Principal components are used to reduce the number of parameters in the regression, given the small sample size for annual data.

Dynamic Factor Model (for monthly inflation rates) We assume all monthly inflation rates follow a dynamic factor model, as specified in [Bańbura and Modugno, 2014] and [Banbura et al., 2011]:

$$y_t = \Lambda f_t + \epsilon_t \tag{50}$$

$$f_t = A_1 f_{t-1} + \cdots + A_p f_{t-p} + u_t \tag{51}$$

$$y_t = [\pi_t^T, x_{GAS}^T, x_t^T]^T \tag{52}$$

where y_t is a vector of standardized endogenous variables, including all inflation rates for month t, and principal components of GAS variables that explain 90% of the variation for this month[8]. Λ represents factor loadings, A_i are the AR coefficients, and f_t are the unobserved factors.

We assume a block structure for factor loadings, with 2 Global factors shared by all variables, and 7 themed factors (e.g., rates, production, confidence, labor market, price and money, inflation expectation) to group variables. The Global factors follow an AR(2), while the themed factors follow an AR(1) based on our previous VAR model estimates. We assume $u_t \sim N(0, Q)$ and follows an AR(1). The EM algorithm is used to estimate the model, and h-step-ahead predictions are made using the estimated parameters:

$$\hat{\pi}_{t+h}^{DFM} = [1_{1\times 13} \quad 0]\hat{\Lambda}\hat{f}_{t+h|h} \tag{53}$$

Shrinkage and Machine Learning Methods (for annual inflation rates) For the following methods, we use the same set of standardized predictor variables and estimate a model for each component of the annual inflation rate to generate 1-step-ahead forecasts. The predictor variables include principal components of GAS variables explaining 90% of the variation, previous year's averaged monthly predictors, and the available monthly predictors for the current year. We use z_t to represent all predictor variables for period t:

$$\pi_{it} = \mu_i + \beta_i^T z_t + \epsilon_{it} \tag{54}$$

Ridge We estimate the ridge coefficients as follows:

$$(\hat{\mu}_i^{Ridge}, \hat{\beta}_i^{Ridge}) = \underset{\mu_i,\beta_i}{\arg\min} \left\{ \frac{1}{T} \sum_{t=1}^{T} (\pi_{it} - \mu_i - \beta_i z_t)^2 + \lambda_i \beta_i^T \beta_i \right\} \tag{55}$$

[8]Since GAS variables have an annual frequency, we repeat them 12 times a year to convert them into monthly values. We also reduced the dimension of the GAS variables because many are highly correlated.

where the regularization parameter λ_i is chosen using 5-fold cross-validation[9]. The 1-step-ahead forecast is:

$$\hat{\pi}_{it+1}^{Ridge} = \hat{\mu}_i^{Ridge} + (\hat{\beta}_i^{Ridge})^T z_{t+1} \tag{56}$$

Elastic Net (for annual inflation rates) We estimate the coefficients using the elastic net method of [Zou and Hastie, 2005]:

$$(\hat{\mu}_i^{EN}, \hat{\beta}_i^{EN}) = \underset{\mu_i,\beta_i}{\arg\min} \left\{ \frac{1}{T} \sum_{t=1}^{T} (\pi_{it} - \mu_i - \beta_i z_t)^2 + \alpha l_1 ||\beta_i||_1 + 0.5\alpha(1 - l_1)||\beta_i||_2^2 \right\} \tag{57}$$

where α and l_1 are chosen by 5-fold cross-validation[10]. Elastic net retains correlated predictors, which would otherwise be dropped by lasso, leading to a 1-step-ahead forecast:

$$\hat{\pi}_{it+1}^{EN} = \hat{\mu}_i^{EN} + (\hat{\beta}_i^{EN})^T z_{t+1} \tag{58}$$

Adalasso (for annual inflation rates) We estimate the coefficients using the adaptive lasso method:

$$(\hat{\mu}_i^{ada}, \hat{\beta}_i^{ada}) = \underset{\mu_i,\beta_i}{\arg\min} \left\{ \frac{1}{T} \sum_{t=1}^{T} (\pi_{it} - \mu_i - \beta_i z_t)^2 + \lambda_i \sum_{j=1}^{p} \omega_j |\beta_{ij}| \right\} \tag{59}$$

where λ are chosen by 5-fold cross-validation. The Adaptive Lasso modifies the Lasso by introducing adaptive weights for each coefficient, leading to a 1-step-ahead forecast:

$$\hat{\pi}_{it+1}^{EN} = \hat{\mu}_i^{ada} + (\hat{\beta}_i^{ada})^T z_{t+1} \tag{60}$$

Random Forest and Gradient Boosting (for annual inflation rates) We also use random forest and gradient boosting, along with the same pool of predictors. [Breiman, 2001] developed the random forest as a method for averaging over multiple regression trees. Each tree approximates a nonlinear function and partitions the predictor space into local regions.

[9]We select the candidate hyperparameter from a grid of 100 values, evenly spaced on a logarithmic scale from 10^{-6} to 10^{6}.

[10]Candidate l_1 values are selected from a grid of 10 values from 0.5 to 1, evenly spaced on a square root scale. For each l_1, α values range from a maximum defined by setting all parameters to zero, to a minimum of 0.001 of the maximum.

Predictions for each tree are the average value of the corresponding region:

$$\pi_{it+1} = \sum_{k=1}^{K} c_k I_k(z_{t+1} \in R_k) \tag{61}$$

where R_k denotes the k^{th} region partitioned by the algorithm, and c_k is the average inflation rate of the region. Usually, block bootstrapping on the observation dimension and random subsets of predictors are drawn to construct trees, reducing correlations between trees. Given the short sample size, we only randomize the subsets of features, while using the entire sample for each tree. We choose the number of trees and maximum depth via 5-fold cross-validation[11]. Final forecasts are based on averaging over the trees:

$$\hat{\pi}_{it+1}^{RF} = \frac{1}{B} \sum_{b=1}^{B} \sum_{k=1}^{K} c_{k,b} I_{k,b}(z_{t+1} \in R_{k,b}) \tag{62}$$

where $R_{k,b}$ represents the k^{th} region of the b^{th} tree. Gradient boosting, on the other hand, extends on the regression trees in that it builds the forecasting model sequentially, eg. starting with a constant estimator, and iteratively fitting a new regression tree to the residuals of the previous forecast model. The final additive model is use to make the prediction for the target variable

$$\hat{\pi}_{it+1,m}^{GB} = \hat{\pi}_{it+1,m-1}^{GB} + v h_m(z_{t+1}) \tag{63}$$

where v is the learning rate that controls the contribution of each model.

5.1.4 Cross-temporal forecast reconciliation

As we have a cross-temporal constraint structure, we use the cross-temporal forecast reconciliation proposed in [Di Fonzo and Girolimetto, 2023] to reconcile both inflation rate of various frequencies and of different categories of goods and services according to the constraints specified in Section 5.1.2. The main idea behind forecast reconciliation is to project the base forecasts onto the (linear) constraint space, with the distance metrics being a function of the error covariance matrix of the base forecast errors. To be specific, our reconciled forecast for inflation rates $\tilde{\pi}_{i,t+1}$ is the solution of the following constrained minimization problem:

$$\tilde{\pi} = \arg\min_{\pi} (\pi - \hat{\pi})^T W^{-1} (\pi - \hat{\pi}) \quad s.t. C\pi = 0. \tag{64}$$

[11] Candidate values for the number of trees range from 2 to 100, and for the maximum depth, from 1 to 9

The solution is then given by

$$\tilde{\pi} = \hat{\pi} - WC^T(CWC^T)^{-1}C\hat{\pi} \tag{65}$$

What [Di Fonzo and Girolimetto, 2023] suggested is that we can iteratively perform this reconciliation step cross-sectionally for different inflation categories, and temporally across different frequencies. In our experience, it only takes around 5 iterations for the the iteration to converge and both dimensions of constraints satisfied.

5.2 Results

The results for our forecasting exercise is shown in Figure 8 and 9. The former shows the how many times each candidate base forecast model gets selected through cross-validation for each forecasting period. From the stacked bar chart on the left, we can see that adalasso seem to perform the best in predicting annual variables while other methods seem to perform similarly. From the performance for monthly variables on the right of Figure 8, we can see that dynamic factor model performs the best, surprisingly followed by random walk [12]. In terms of the second stage forecast presented in Figure 9, we can see that the cross-temporal forecast reconciliation generally improved the first-stage forecast and get to similar level as the WEO forecast for the first half of the sample. However, towards the beginning of Covid, our data-driven forecast methods became highly inaccurate while WEO forecast stayed accurate, leading to large difference in overall performance. From the right side plot comparing different covariance matrix estimators, we can see that estimators using identity target seem to work better in this case than those using diagonal target, resulting from a less sparse covariance matrix with small variability on the diagonal. $OASB$ in this case correctly puts more weight on the identity target and performs well.

6 Application 2: Minimum variance portfolio

We also applied our covariance matrix estimator to the construction of a mimumim variance portfolio. We first selected 100 most uncorrelated instruments from Bloomberg US Equity and Fixed-income indices so that they are representative of different aspects of the market (They still have a reasonable amount of correlation. Then we construct our minimum

[12]The reason April has more variables to be predicted than October is that the forecaster needs to forecast all remaining monthly variables of the year, which is more than those in October

variance portfolio based on the following weights x:

$$x(W) = \frac{W^{-1}\mathbf{1}}{\mathbf{1}^T W^{-1}\mathbf{1}} = \arg\min_x x^T W x \quad s.t. \quad x^T\mathbf{1} = 1. \tag{66}$$

The weight estimated from a training sample will be used to form portfolios for several periods' ahead, depending on the holding period. We then compute the realized volatility of the portfolio's holding period return. Different covariance matrix results in different volatility for the portfolio, and the true covariance matrix should deliver the lowest volatility. This gives us a clean environment to test the performance of different covariance matrix estimators by comparing the realized volatility of the portfolios constructed using weights calculated from Equation 66. We work with monthly data and experiment with training sizes ranging from 10 months to 40 months and holding periods of 6, 12, 24, and 60 months, which are typical in practice.

The results for the realized standard deviation of porfolios constructed with different covariance matrix estimators are shown in Figure 10, with Figure 11 zooming in on the best performing estimators. The results show that $OASD$ performs the best among all candidate covariance matrix estimators while estimators based on an identity target performs much worse. $OASB$'s performance again is between the two groups of estimators with different targets.

7 Conclusion

This paper has proposed a novel covariance matrix estimator $OASD$ that achieve a smaller MSE than the existing methods when the variation in variable scales is large. It is useful, for example, when different variables have different units. We further went on to propose $OASB$ that adapts the target choice depending on the patterns of the true covariance matrix and is shown to perform better under more scenarios. A forecast reconciliation and a portfolio construction applications were conducted to demonstrate the usefulness of the methods proposed. We confirmed that $OASD$ typically works better when the true covariance matrix is sparser and variables exhibit higher dispersion in scales. $OASB$'s performance is usually between OAS and $OASD$, best for when the researcher is unsure about the features of the true covariance matrix.

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 deviates substantially from normality.

References

[ONS,] Assessing the impact of methodological improvements on the consumer prices index. https://www.ons.gov.uk/economy/inflationandpriceindices/methodologies/assessingtheimpactofmethodologicalimprovementsontheconsumerpricesindex. Accessed: 2023-10-19.

[Ando and Kim, 2022] 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 et al., 2018] Ban, G.-Y., El Karoui, N., and Lim, A. E. (2018). Machine learning and portfolio optimization. Management Science, 64(3):1136–1154.

[Banbura et al., 2011] Banbura, M., Giannone, D., and Reichlin, L. (2011). Nowcasting with daily data. European Central Bank, Working Paper, page 18.

[Bańbura and Modugno, 2014] Bańbura, M. and Modugno, M. (2014). Maximum likelihood estimation of factor models on datasets with arbitrary pattern of missing data. Journal of applied econometrics, 29(1):133–160.

[Bermingham and D'Agostino, 2014] Bermingham, C. and D'Agostino, A. (2014). Understanding and forecasting aggregate and disaggregate price dynamics. Empirical Economics, 46(2):765–788.

[Boaretto and Medeiros, 2023] Boaretto, G. and Medeiros, M. C. (2023). Forecasting inflation using disaggregates and machine learning. arXiv preprint arXiv:2308.11173.

[Breiman, 2001] Breiman, L. (2001). Random forests. Machine learning, 45:5–32.

[Capistrán et al., 2010] Capistrán, C., Constandse, C., and Ramos-Francia, M. (2010). Multi-horizon inflation forecasts using disaggregated data. Economic Modelling, 27(3):666–677.

[Chen et al., 2009] 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 et al., 2009] 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.

[Dey and Srinivasan, 1985] Dey, D. K. and Srinivasan, C. (1985). Estimation of a covariance matrix under stein's loss. The Annals of Statistics, pages 1581–1591.

[Di Fonzo and Girolimetto, 2023] Di Fonzo, T. and Girolimetto, D. (2023). Cross-temporal forecast reconciliation: Optimal combination method and heuristic alternatives. International Journal of Forecasting, 39(1):39–57.

[Eldar and Chernoi, 2008] 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.

[Elton and Gruber, 1973] Elton, E. J. and Gruber, M. J. (1973). Estimating the dependence structure of share prices–implications for portfolio selection. The Journal of Finance, 28(5):1203–1232.

[Espasa et al., 2002] Espasa, A., Senra, E., and Albacete, R. (2002). Forecasting inflation in the european monetary union: A disaggregated approach by countries and by sectors. The European Journal of Finance, 8(4):402–421.

[Fan et al., 2008] Fan, J., Fan, Y., and Lv, J. (2008). High dimensional covariance matrix estimation using a factor model. Journal of Econometrics, 147(1):186–197.

[Faust and Wright, 2013] Faust, J. and Wright, J. H. (2013). Forecasting inflation. In Handbook of economic forecasting, volume 2, pages 2–56. Elsevier.

[Galı and Gertler, 1999] Galı, J. and Gertler, M. (1999). Inflation dynamics: A structural econometric analysis. Journal of monetary Economics, 44(2):195–222.

[Haff, 1980] Haff, L. (1980). Empirical bayes estimation of the multivariate normal covariance matrix. The Annals of Statistics, 8(3):586–597.

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

[Ibarra, 2012] Ibarra, R. (2012). Do disaggregated cpi data improve the accuracy of inflation forecasts? Economic Modelling, 29(4):1305–1313.

[Ledoit and Wolf, 2003] Ledoit, O. and Wolf, M. (2003). Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. Journal of empirical finance, 10(5):603–621.

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

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

[Marchenko and Pastur, 1967] Marchenko, V. A. and Pastur, L. A. (1967). Distribution of eigenvalues for some sets of random matrices. Matematicheskii Sbornik, 114(4):507–536.

[Million, 2007] Million, E. (2007). The hadamard product. Course Notes, 3(6):1–7.

[Panagiotelis et al., 2023] Panagiotelis, A., Gamakumara, P., Athanasopoulos, G., and Hyndman, R. J. (2023). Probabilistic forecast reconciliation: Properties, evaluation and score optimisation. European Journal of Operational Research, 306(2):693–706.

[Pearson, 1901] 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.

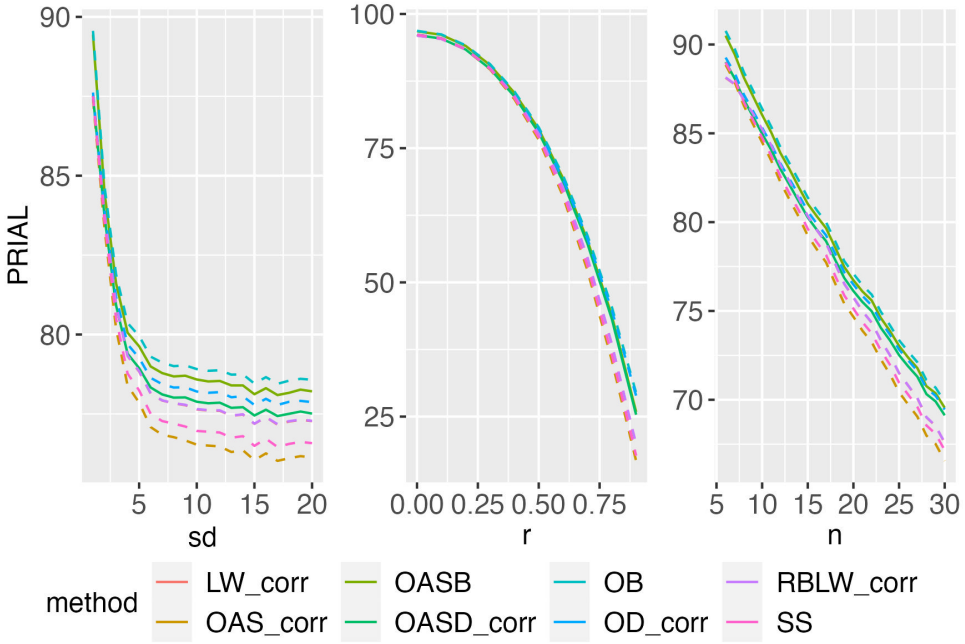
[Schäfer and Strimmer, 2005] Schäfer, J. and Strimmer, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. Statistical applications in genetics and molecular biology, 4(1).

[Stein, 1975] Stein, C. (1975). Estimation of a covariance matrix. In 39th Annual Meeting IMS, Atlanta, GA, 1975.

[Wickramasuriya et al., 2019] Wickramasuriya, S. L., Athanasopoulos, G., and Hyndman, R. J. (2019). Optimal forecast reconciliation for hierarchical and grouped time series through trace minimization. Journal of the American Statistical Association, 114(526):804–819.

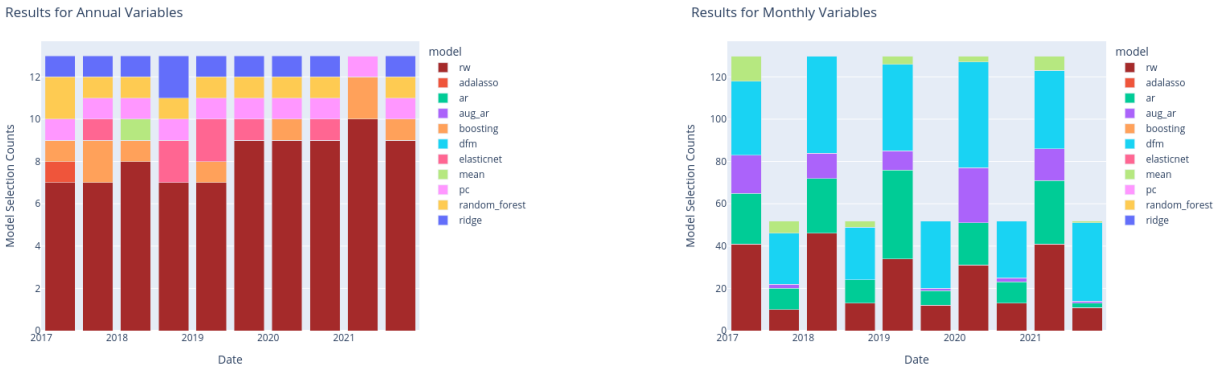
[Zou and Hastie, 2005] Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society Series B: Statistical Methodology, 67(2):301–320.

Figure 7: $PRIAL$ comparison with correlation-based methods



COICOP Symbol	Description
01_IX	Food and non-alcoholic beverages
02_IX	Alcohol and tobacco
03_IX	Clothing and footwear
04_IX	Housing and household services
05_IX	Furniture and household goods
06_IX	Health
07_IX	Transport
08_IX	Communication
09_IX	Recreation and culture
10_IX	Education
11_IX	Restaurants and hotels
12_IX	Miscellaneous goods and services

Table 1: COICOP subgroups of CPI



(a) Annual variables (b) Monthly variables

Figure 8: Model selection for different covariance matrix estimators

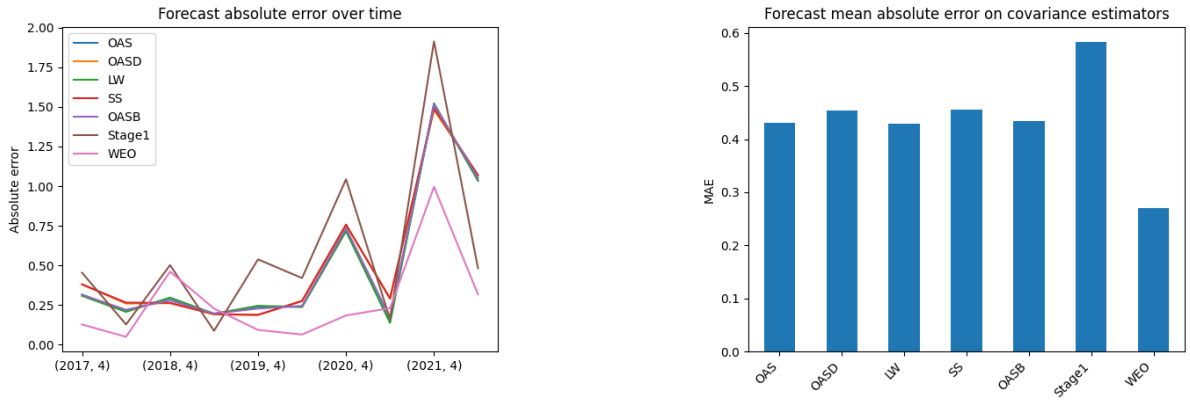


Figure 9: Forecast absolute error for different covariance matrix estimators

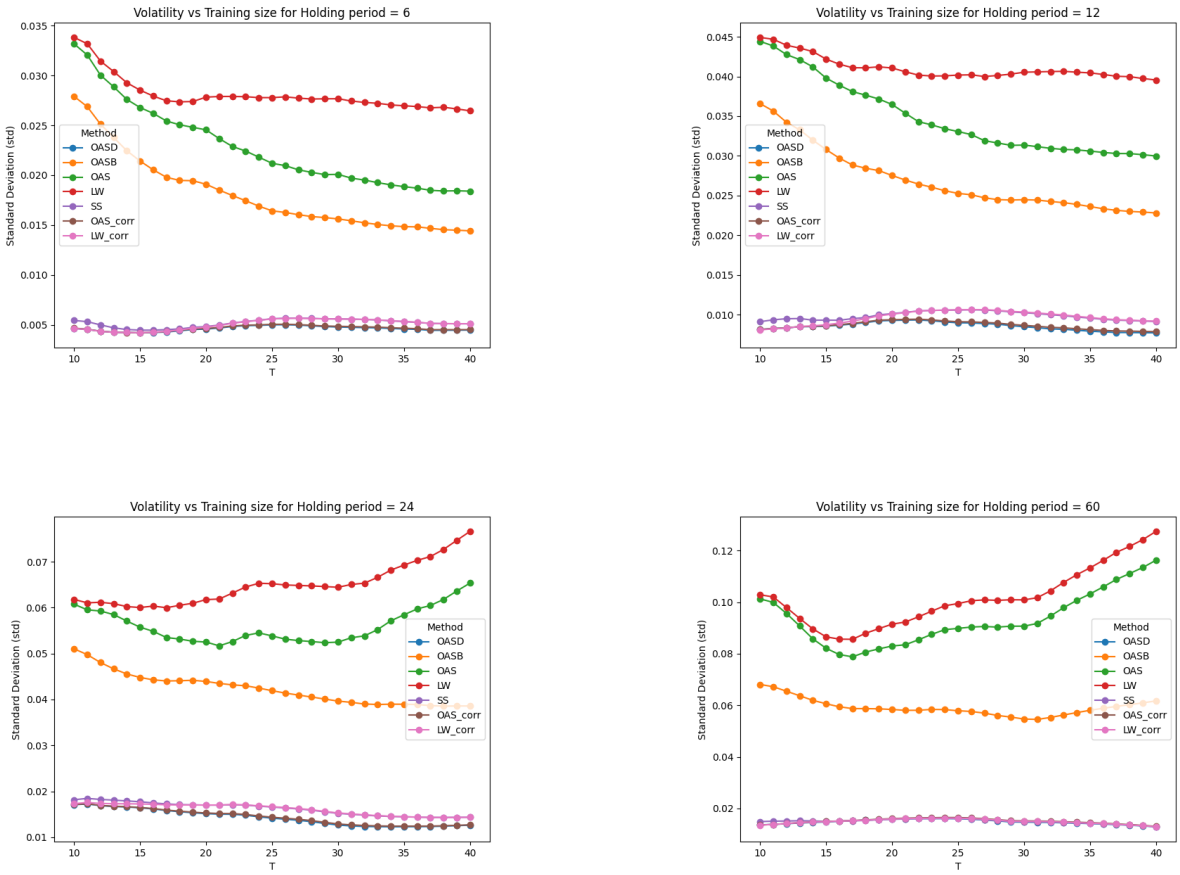


Figure 10: Portfolio standard deviation for various covariance matrix estimators

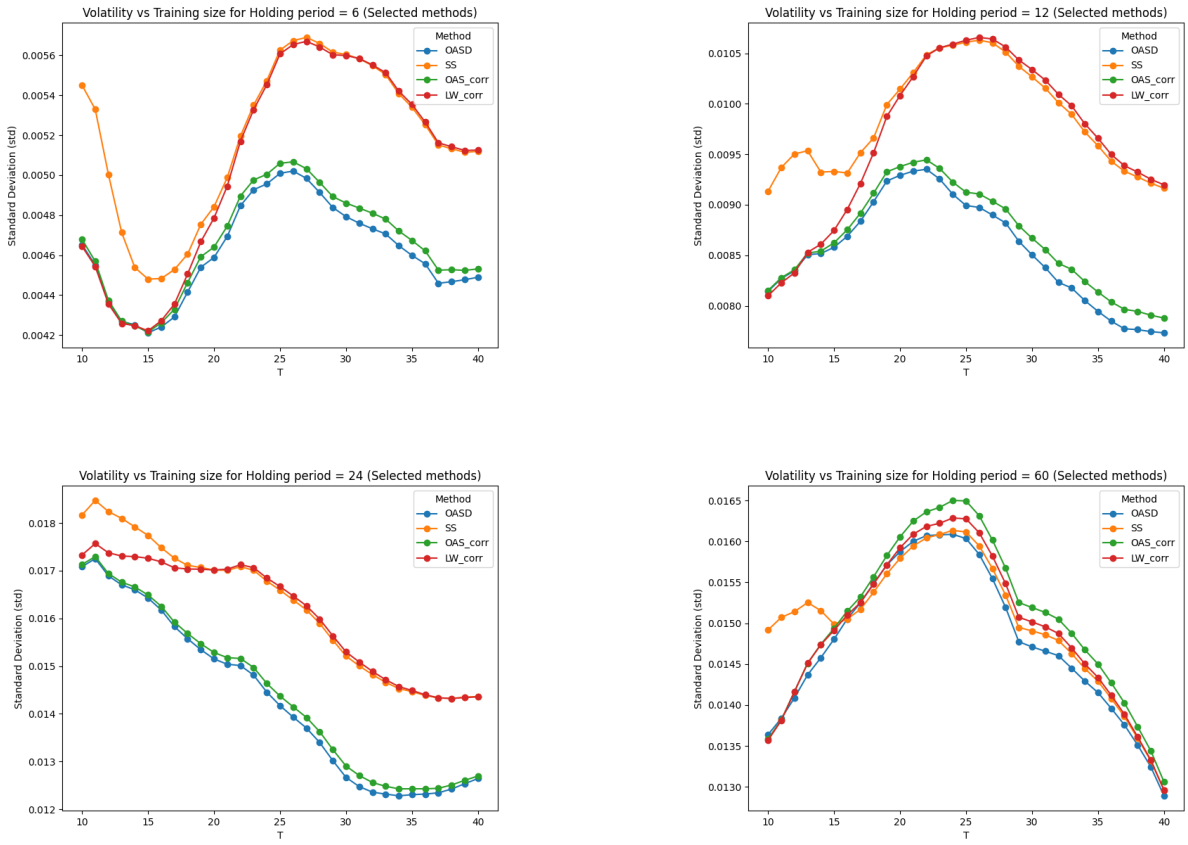


Figure 11: Portfolio standard deviation for selected covariance matrix estimators

Appendix

7.1 Proof of theorem 1

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

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

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.

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 $z_{il} z_{jl}$

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

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

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

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.

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

7.2 Proof of theorem 2

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

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

where ϕ is defined as

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

Note that $\phi \in [0, 1)$ because

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

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

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, then 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

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

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 $j' \geq j$, then the statement is proved. Otherwise, the same change of variable gives a well-defined b_j

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

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

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

Therefore,

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

7.3 Proof of theorem 3

The proof is a simpler version of Appendix 7.1 and 7.2. We first establish the following lemma. 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,

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

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

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

The integrand can be transformed into

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

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

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

Thus

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

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 gives

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

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 7.2, except that n is replaced by $n+1$. Thus, following the same argument,

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

7.4 Proof of theorem 4

Note from 13,

$$\tilde{S}(\theta, \alpha) = (1-\theta)S + \theta(\alpha T_{OASD} + (1-\alpha)T_{OAS})$$

Similar to the result 5 in Theorem 1, by minimizing the MSE criterion in 14, we can get the oracle shrinkage parameter θ_{OB} and α_{OB}. (To simplify notation, use $< M, N >$ to denote $tr(M^T N)$:

$$\theta_{OB} = \frac{E[< \Sigma - S, \alpha_{OB} T_{OASD} + (1-\alpha_{OB})T_{OAS} - S >]}{E\left[\|S - (\alpha_{OB} T_{OASD} + (1-\alpha_{OB})T_{OAS})\|^2\right]}$$

$$\alpha_{OB} = 1 - \frac{1}{\theta_{OB}} \frac{E[< S - \Sigma, T_{OASD} - T_{OAS} >]}{E\left[\|T_{OASD} - T_{OAS}\|^2\right]}$$

Substituting the equation for α_{OB} into the equation for θ_{OB} gives

$$\theta_{OB} = \frac{E[< S - \Sigma, S - T_{OASD}) >]}{E[< S - T_{OAS}, S - T_{OASD}) >]}$$

Using the lemma derived in the proof of Theorem 1, we have

$$\begin{aligned}
\theta_{OB} &= \frac{E[< S - \Sigma, S - T_{OASD}) >]}{E[< S - T_{OAS}, S - T_{OASD} >]} \\
&= \frac{\frac{tr(\Sigma)^2 + tr(\Sigma^2) - 2tr(diag(\Sigma)^2)}{n-1}}{\frac{n}{n-1}tr(\Sigma^2) + \frac{1}{n-1}tr(\Sigma)^2 - \frac{n+1}{n-1}tr(diag(\Sigma)^2)} \\
&= \frac{tr(\Sigma)^2 + tr(\Sigma^2) - 2tr(diag(\Sigma)^2)}{ntr(\Sigma^2) + tr(\Sigma)^2 - (n+1)tr(diag(\Sigma)^2)}
\end{aligned}$$

which is the same as the optimal value for ρ_{OD} when the target is only $diag(S)$ in Theorem 6. To simplify α_{OB}, we need to use the lemma derived in Theorem 6 as well as an additional lemma, which we derive below. When $x_i \sim N(\mu, \Sigma)$ is *i.i.d.*, the following equation holds

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

Proof. Following the notation for proof of Theorem 6, we have

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

The first quantity in the curly bracket is proven in Theorem 6 as

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

For the second quantity, we first use the equation $w_i^T w_i = \sum_{m=1}^p \lambda_m z_{im}^2$ to derive the following

$$\begin{aligned} E\left[w_i^T w_i w_j^T w_j\right] &= E\left[\sum_{m=1}^p \sum_{k=1}^p \lambda_m \lambda_k z_{im}^2 z_{jk}^2\right] \\ &= \sum_{m=1}^p \sum_{k=1}^p \lambda_m \lambda_k E\left[z_{im}^2 z_{jk}^2\right] \end{aligned}$$

From the second moment of the multivariate normal distribution and that $E\left[z_i z_j^T\right] = -\frac{1}{n-1} I$, we can get the second moment of $z_{im} z_{jk}$

$$\begin{aligned} E\left[z_{im}^2 z_{jk}^2\right] &= Var\left[z_{im}\right] Var\left[z_{jk}\right] + 2Cov\left[z_{im}, z_{jk}\right]^2 \\ &= \begin{cases} 1 + \frac{2}{(n-1)^2}, & \text{if } m = k \\ 1, & \text{if } m \neq k \end{cases} \end{aligned}$$

Therefore we have

$$\begin{aligned} \sum_{m=1}^p \sum_{k=1}^p \lambda_m \lambda_k E\left[z_{im}^2 z_{jk}^2\right] &= \sum_{m=1}^p \lambda_m^2 \left[1 + \frac{2}{(n-1)^2}\right] + \sum_{m\neq k}^p \lambda_m \lambda_k \\ &= \left[1 + \frac{2}{(n-1)^2}\right] tr(U^2) + \left[tr(U)^2 - tr(U^2)\right] \\ &= \frac{2}{(n-1)^2} tr(U^2) + tr(U)^2 \end{aligned}$$

When we sum over all $i \neq j$, we get the second quantity as

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

Putting both quantities together

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

Now we use the definition $U = \frac{n-1}{n}\Sigma$ to get

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

■Using the lemma from Theorem 6 and the one just derived, α_O can be simplified to

$$\begin{aligned}
\alpha_{OB} &= 1 - \frac{1}{\theta_{OB}}\frac{E\left[< S-\Sigma, T_{OASD} - T_{OAS} >\right]}{E\left[\|T_{OASD} - T_{OAS}\|^2\right]} \\
&= 1 - \frac{1}{\theta_{OB}}\frac{\frac{2}{n-1}tr(diag(\Sigma)^2) - \frac{2}{(n-1)p}tr(\Sigma^2)}{\frac{n+1}{n-1}tr(diag(\Sigma)^2) - \frac{2}{(n-1)p}tr(\Sigma^2) - \frac{1}{p}tr(\Sigma)^2} \\
&= 1 - \frac{1}{\theta_{OB}}\frac{2ptr(diag(\Sigma)^2) - 2tr(\Sigma^2)}{p(n+1)tr(diag(\Sigma)^2) - 2tr(\Sigma^2) - (n-1)tr(\Sigma)^2}
\end{aligned}$$

7.5 Proof of theorem 5

To simplify notation, we write θ_{OASB} as θ in this subsection and define the following variables

$$A = \frac{tr(S)^2}{p} \quad B = tr(S^2) - tr(diag(S)^2) \quad C = tr(diag(S)^2) - \frac{tr(S)^2}{p}$$

Note that by Cauchy Schwartz and properties of the trace of a matrix, we have the following inequalities:

$$tr(S)^2 \geq tr(S^2) > tr(diag(S)^2) \geq \frac{tr(S)^2}{p} \geq \frac{tr(S^2)}{p}$$

where the first and last inequality would be strict unless all the sample correlations are 1. The second inequality is strict because we assume that all the sample variances are positive. The third inequality is strict unless all sample variances are equal. Therefore, A and B defined above are positive and C is non-negative and only 0 when all the sample variances are equal. Substituting $\Sigma_j = (1-\theta_{OASB})S + \theta_{OASB}(\alpha_j T_{OASD} + (1-\alpha_j)T_{OAS})$ into the iteration and a direct calculation lead to

$$\begin{aligned}\alpha_{j+1} &= 1 - \frac{1}{\theta}\frac{2(ptr(diag(S)^2) - tr(S^2)) + 2\theta\alpha_j(p-1)C + 2\theta(B-(p-1)C)}{2(ptr(diag(S)^2) - tr(S^2)) + (n-1)pC + \theta\alpha_j(np+p-2)C + \theta((2-np-p)C + 2B)} \\ &= 1 - \frac{1}{\theta}\frac{1 + \tau_1\theta\alpha_j + \tau_2\theta}{1 + \tau_3 + (\tau_1+\tau_3)\theta\alpha_j + (\tau_2-\tau_3)\theta},\end{aligned}$$

where we defined

$$\tau_1 = \frac{(p-1)C}{ptr(diag(S)^2) - tr(S^2)} = \frac{(p-1)C}{(p-1)A + (p-1)C - B}$$

$$\tau_2 = \frac{B-(p-1)C}{ptr(diag(S)^2) - tr(S^2)} = \frac{B-(p-1)C}{(p-1)A + (p-1)C - B}$$

$$\tau_3 = \frac{(n-1)pC}{2(ptr(diag(S)^2) - tr(S^2))} = \frac{(n-1)pC}{2[(p-1)A + (p-1)C - B]}$$

assuming all quantities are well-defined. In the case where $ptr(diag(S)^2) = tr(S^2)$, we have $\alpha_j = 1 - \frac{1}{\theta}$ for all j and the theorem is proved. If $C = 0$ thus τ_1 and τ_3 are both 0, we again have $\alpha_j = 1 - \frac{1}{\theta}$ for all j and the theorem is proved. If now τ_1 and τ_3 are both non-zero, but if $\alpha_j = \frac{\theta(\tau_3-\tau_2)-1}{\theta(\tau_1+\tau_3)}$ at any step of the iteration process, substituting it into the iteration results in $\alpha_{j'} = \frac{\theta(\tau_3-\tau_2)-1}{\theta(\tau_1+\tau_3)}$ for all $j' > j$ and the theorem is proved. For the general case where τ_1, τ_2, and τ_3 are all well-defined, $C \neq 0$, and $\alpha_j \neq \frac{\theta(\tau_3-\tau_2)-1}{\theta(\tau_1+\tau_3)}$ for all j, we can do the following change of variable

$$c_j := \frac{1}{\alpha_j - \frac{\theta(\tau_3-\tau_2)-1}{\theta(\tau_1+\tau_3)}} \Leftrightarrow \alpha_j = \frac{1}{c_j} + \frac{\theta(\tau_3-\tau_2)-1}{\theta(\tau_1+\tau_3)}$$

The updating equation now becomes the following recursion

$$c_{j+1} = \frac{\tau_3}{(\tau_1+\tau_2)\theta + 1 - \tau_1}c_j + \frac{(\tau_1+\tau_3)\theta}{(\tau_1+\tau_2)\theta + 1 - \tau_1}$$

and we can get the limit of this linear dynamic system

$$\left|\frac{\tau_3}{(\tau_1+\tau_2)\theta + 1 - \tau_1}\right| < 1 \Rightarrow c_j \to \frac{(\tau_1+\tau_3)\theta}{(\tau_1+\tau_2)\theta + 1 - \tau_1 - \tau_3} \Leftrightarrow \alpha_j \to \frac{\theta-1}{\theta}$$

$$\left|\frac{\tau_3}{(\tau_1+\tau_2)\theta+1-\tau_1}\right| \geq 1 \Rightarrow |c_j| \to \infty \Leftrightarrow \alpha_j \to \frac{\theta(\tau_3-\tau_2)-1}{\theta(\tau_1+\tau_3)}$$

Therefore, we have the following converging limit

$$\alpha_{OASB} = \begin{cases} \frac{\theta-1}{\theta}, & \text{if } \left|\frac{\tau_3}{(\tau_1+\tau_2)\theta+1-\tau_1}\right| < 1 \\ \frac{\theta(\tau_3-\tau_2)-1}{\theta(\tau_1+\tau_3)}, & \text{if } \left|\frac{\tau_3}{(\tau_1+\tau_2)\theta+1-\tau_1}\right| \geq 1 \end{cases}$$

Because $\theta = \rho_{OASD}$, it follows that $\theta \in (0, 1]$. For α_{OASB}, we have if

$$\left|\frac{\tau_3}{(\tau_1+\tau_2)\theta+1-\tau_1}\right| \geq 1$$

then we have

$$\alpha_{OASB} - \frac{\theta-1}{\theta} = \frac{\theta(\tau_3-\tau_2)-1}{\theta(\tau_1+\tau_3)} - \frac{\theta-1}{\theta} = \frac{\tau_3+\tau_1-1-\theta(\tau_1+\tau_2)}{\theta(\tau_1+\tau_3)} \geq 0$$

Notice that $\tau_1 + \tau_2 > 0$ and thus

$$\frac{\theta(\tau_3-\tau_2)-1}{\theta(\tau_1+\tau_3)} - 1 = \frac{-\theta(\tau_1+\tau_2)-1}{\theta(\tau_1+\tau_3)} < 0$$

Similar results holds trivially for the case when $\alpha_{OASB} = \frac{\theta-1}{\theta}$. Therefore $\alpha_{OASB} \in [1-\frac{1}{\theta}, 1)$ Despite the fact that α_{OASB} is no longer restricted to be between 0 and 1, the property of the final estimator being positive definite still holds because of the following claim For any nonzero vector x and any weighting scaler α, the linear combination

$$\alpha diag(S) + (1-\alpha)\frac{tr(S)}{p}$$

is positive definite and therefore any convex combination of the above quantity with the sample covariance matrix results in a positive definite covariance matrix estimator. **Proof.**

$$\begin{aligned} \alpha x^T diag(S)x + (1-\alpha)x^T\frac{tr(S)}{p}x &\geq \alpha\lambda_{min}(diag(S))x^Tx + (1-\alpha)\lambda_{min}(S)x^Tx \\ &\geq \lambda_{min}(S)x^Tx > 0 \end{aligned}$$

where the second to last inequality follows from the property of $diag(S)$ derived in Appendix 7.7 and the last inequality follows from our assumption that the sample variances are all positive. ■

7.6 Proof of theorem 6

The proof is a simpler version of Appendix 7.4 and 7.5. The proof is the same as Appendix 7.4 until we get the following expressions for θ_{OB} and α_{OB}

$$\theta_{OB} = \frac{E\left[< S - \Sigma, S - T_{OASD}) >\right]}{E\left[< S - T_{OAS}, S - T_{OASD}) >\right]}$$

$$\alpha_{OB} = 1 - \frac{1}{\theta_{OB}} \frac{E\left[< S - \Sigma, T_{OASD} - T_{OAS} >\right]}{E\left[\|T_{OASD} - T_{OAS}\|^2\right]}$$

In order to evaluate the expectations, we make use of the lemmas proved in Appendix 7.3 and a simpler version of the additional lemma we proved in Appendix 16.

When $x_i \sim N(0, \Sigma)$ is *i.i.d.*, the following equation holds

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

Proof.

$$\begin{aligned}
E\left[tr(S)^2\right] &= \frac{1}{n^2} E\left[\left\{\sum_i tr(x_i x_i^T)\right\}^2\right] \\
&= \frac{1}{n^2} V\left[\sum_i tr(x_i x_i^T)\right] + \frac{1}{n^2}\left\{E\left[\sum_i tr(x_i x_i^T)\right]\right\}^2 \\
&= \frac{1}{n} V\left[tr(x_i x_i^T)\right] + tr(\Sigma)^2 \\
&= \frac{1}{n} V\left[x_i^T x_i\right] + tr(\Sigma)^2 \\
&= \frac{2}{n} tr(\Sigma^2) + tr(\Sigma)^2
\end{aligned}$$

where the last equality used the expression for $V\left[x_i^T x_i\right]$ we derived in Appendix 12 ■ Now using the lemma in Appendix 12 and the additional one derived above, we can simplify θ_{OB} to be

$$\begin{aligned}
\theta_{OB} &= \frac{E\left[< S - \Sigma, S - T_{OASD}) >\right]}{E\left[< S - T_{OAS}, S - T_{OASD} >\right]} \\
&= \frac{tr(\Sigma)^2 + tr(\Sigma^2) - 2tr(diag(\Sigma)^2)}{(n+1)tr(\Sigma^2) + tr(\Sigma)^2 - (n+2)tr(diag(\Sigma)^2)}
\end{aligned}$$

which is the same as the optimal value for ρ_{OD} when the target is only $diag(S)$ in Theorem 3 α_{OB} therefore simplifies to

$$\begin{aligned}\alpha_{OB} &= 1 - \frac{1}{\theta_{OB}} \frac{\frac{2}{n} tr(diag(\Sigma)^2) - \frac{2}{np} tr(\Sigma^2)}{\frac{n+2}{n} tr(diag(\Sigma)^2) - \frac{2}{np} tr(\Sigma^2) - \frac{1}{p} tr(\Sigma)^2} \\ &= 1 - \frac{1}{\theta_{OB}} \frac{2ptr(diag(\Sigma)^2) - 2tr(\Sigma^2)}{p(n+2)tr(diag(\Sigma)^2) - 2tr(\Sigma^2) - ntr(\Sigma)^2}\end{aligned}$$

Note again θ_{OB} is the same as ρ_{OD} under the known mean case so we take $\theta_{OASB} = \rho_{OASD}$. For α_{OASB}, we use the limit of the following iteration to approximate the oracle

$$\alpha_{j+1} = 1 - \frac{1}{\theta_{OASB}} \frac{2ptr(diag(\Sigma_j)diag(S)) - 2tr(\Sigma_j S)}{p(n+2)tr(diag(\Sigma_j)diag(S)) - 2tr(\Sigma_j S) - ntr(\Sigma_j)^2}$$

Notice that this updating equation is identical to 18, except that n is replaced by $n+1$. Thus, following the same argument, we get

$$\alpha_{OASB} = \begin{cases} \dfrac{\theta_{OASB} - 1}{\theta_{OASB}}, & \text{if } \left| \dfrac{\tau_3}{(\tau_1 + \tau_2)\theta_{OASB} + 1 - \tau_1} \right| < 1 \\ \dfrac{\theta_{OASB}(\tau_3 - \tau_2) - 1}{\theta_{OASB}(\tau_1 + \tau_3)}, & \text{if } \left| \dfrac{\tau_3}{(\tau_1 + \tau_2)\theta_{OASB} + 1 - \tau_1} \right| \geq 1 \end{cases}$$

where we used the θ_{OASB} in 12, τ_3 is adjusted to be

$$\tau_3 = \frac{np\left[tr(diag(S)^2) - \frac{tr(S)^2}{p}\right]}{2(ptr(diag(S)^2) - tr(S^2))}$$

and τ_1 and τ_2 are as defined in Appendix 7.5 since they don't involve n.

7.7 Proof of Comment 1

Given S is a real symmetric matrix, we can use Theorem 2.1 of [Million, 2007] to get the following results for $diag(S)$:

$$diag(S) \times 1 = \begin{pmatrix} s_{11} \\ s_{22} \\ ... \\ s_{pp} \end{pmatrix} = \Gamma \circ \Gamma \begin{pmatrix} \lambda_1(S) \\ \lambda_2(S) \\ ... \\ \lambda_p(S) \end{pmatrix}$$

where Γ is the eigenvector matrix of S with each column being an eigenvector and 1 is a vector of ones. Since the diagonal entries of $diag(S)$ are its eigenvalues, we consider the dispersion of $\begin{pmatrix} s_{11} \\ s_{22} \\ ... \\ s_{pp} \end{pmatrix}$ relative to that of $\begin{pmatrix} \lambda_1(S) \\ \lambda_2(S) \\ ... \\ \lambda_p(S) \end{pmatrix}$. From the orthogonality of Γ, we know

$$(\Gamma \circ \Gamma) \times 1 = 1$$

We also know that all entries in $\Gamma \circ \Gamma$ are nonnegative. Therefore, all entries in $\Gamma \circ \Gamma$ are between 0 and 1 and each row sums to be 1. This means that any s_{ii} is a linear convex combination of $\lambda_1(S), ...\lambda_p(S)$.This leads to the following:

$$\lambda_{min}(S) \leq \lambda_{min}(diag(S)) \leq \lambda_{max}(diag(S)) \leq \lambda_{max}(S)$$

The first equality would only hold if the row in $\Gamma \circ \Gamma$ that corresponds to $\lambda_{min}(S)$ happens to put all all the weight on this value and similarly for the last equality. Now since we have

$$\lambda_{max}(A+B) = \sup_{|v|=1} v^T(A+B)v \leq \lambda_{max}(A) + \lambda_{max}(B)$$

$$\lambda_{min}(A+B) = \inf_{|v|=1} v^T(A+B)v \geq \lambda_{min}(A) + \lambda_{min}(B)$$

where the equality only holds when all $A+B$, A, B share the same eigenvector that corresponds to the largest/smallest eigenvalues. Therefore, in general (assuming that we are not under the special cases mentioned above), for a combined estimator in the form of $S_c = (1-\rho)S + \rho diag(S)$, we would have

$$\lambda_{min}(S_c) > (1-\rho)\lambda_{min}(S) + \rho\lambda_{min}(diag(S)) > \lambda_{min}(S)$$

Similarly, we have

$$\lambda_{max}(S_c) < (1-\rho)\lambda_{max}(S) + \rho\lambda_{max}(diag(S)) < \lambda_{max}(S)$$

Therefore, S_c is a better conditioned estimator compared to S.