Pomona College

Senior Thesis in Mathematics

Shrinkage Estimators for High-Dimensional Covariance Matrices

Author:
Brian Williamson

Advisor:
Dr. Jo Hardin

Submitted to Pomona College in Partial Fulfillment
of the Degree of Bachelor of Arts

April 4, 2014

SHRINKAGE ESTIMATORS FOR HIGH-DIMENSIONAL COVARIANCE MATRICES

BRIAN WILLIAMSON

Abstract. As high-dimensional data becomes ubiquitous, standard estimators of the population covariance matrix become difficult to use. Specifically, in the case where the number of samples is small (large p small n) the sample covariance matrix is not positive definite. In this paper we explore some recent estimators of sample covariance matrices in the large p, small n setting - namely, shrinkage estimators. Shrinkage estimators have been shown to be positive definite and well-conditioned, two key properties to a good estimate of the population covariance matrix. We test how well the estimators preserve the qualitites of the population covariance matrix and how much information is retained from the sample covariace matrix. We also perform a simulation study to measure the difference between the estimators and the population covariance and compare the different estimators.

Contents

Date: April 4, 2014.

1. Introduction

In many sciences - including spectroscopy, functional magnetic resonance imaging, text retrieval, and gene arrays - high-dimensional data is becoming the norm. As more variables can be measured with the advent of new computing techniques, it becomes difficult to estimate relationships between the p variables using standard covariance and correlation matrices. For example, say that you are a researcher studying the correlation between various genes in humans. You are able to measure 10,000 genes, but only have 200 participants in your study. The resulting data matrix is $200 \times 10,000$. If you wish to study all pairwise correlations, then the covariance matrix that is calculated is $10,000 \times 10,000$. If $n < p$ the sample covariance matrix will never be positive definite (in fact it will be positive semidefinite), and thus is not necessarily a good estimate for the population covariance matrix. We need an estimate which exhibits the same characteristics as the population covariance as well as being positive definite. The problem is finding a reliable and accurate estimate that also performs well when the data has only a few observations in relation to the number of parameters ($n << p$).

A computationally simple estimator for the population covariance Σ is the maximum likelihood estimate, given by

$$(\mathbf{S}^{ML})_{ij} = \frac{1}{n}\sum_{k=1}^{n}(x_{ki} - \overline{x}_i)(x_{kj} - \overline{x}_j), \tag{1}$$

where $\overline{x}_i = \frac{1}{n}\sum_{k=1}^{n} x_{ki}$, and x_{ki} is the kth observation of the vector X_i. Another equally simple estimator is the unbiased sample covariance,

$$\mathbf{S} = \frac{n}{n-1}\mathbf{S}^{ML}. \tag{2}$$

However, both of these estimators exhibit problems in the $n << p$ case where a large number of eigenvalues become zero and the sample covariance loses its full rank. As we will show in the next section, this makes both the maximum likelihood estimate and the unbiased sample covariance undesirable as estimates of the population covariance in the $n << p$ case.

Many estimators for the population covariance assume that it is sparse, or that there are many marginal independencies in the population. While this is intuitive, the data may not behave in exactly the same way. For instance, take the estimator where all of the off-diagonal elements are zero and the diagonals are kept as estimates of the variance of the element, which preserves some information from the sample covariance matrix (namely the variances) but loses all of the information on the off-diagonal. Clearly we need an estimator which preserves more information from the sample covariance matrix but also performs better as an estimate for the population covariance.

One way of quantifying the issue of "better" is calculating the Frobenius norm between the estimate and a created population covariance matrix via a simulation study. The Frobenius norm measures "closeness" to Σ, and so minimal Frobenius norm points to a "better" estimate. A researcher can thus determine the best estimator for different combinations of p and n or for different methods of calculating pairwise correlation (Spearman versus Pearson). These studies are particularly revealing as we enter the large p small n case, in which finding a good estimator affects a growing number of problems.

Also consider the calculation for the mean squared error of the sample covariance,

$$MSE(\mathbf{S}) = Bias(\mathbf{S})^2 + Var(\mathbf{S}). \tag{3}$$

We know that $Bias(\mathbf{S}) = 0$ (since we have defined it as the unbiased estimator) and thus to make a more accurate estimator from the sample covariance we need to reduce its variance. Shrinkage techniques using the sample covariance produce an estimator with reduced variance and provide us with a much improved estimator for the population covariance. The shrinkage estimators we explore are all convex combinations of the sample covariance matrix (2) and a target matrix $\mathbf{T}$, which is a positive definite matrix such as the identity and is generally structured (like the identity with 1's along the diagonal and 0's elsewhere). The estimator is of the form

$$\Sigma^* = \lambda_1\mathbf{S} + \lambda_2\mathbf{T}. \tag{4}$$

We will show that these optimal weights are λ^* and $1-\lambda^*$. However, this comes at a tradeoff, since changing the sample covariance (or using a convex combination of the sample covariance matrix and another matrix) will introduce bias. We need to consider MSE in its entirety, then, in dealing with these estimators.

In the next section we will go through some background of the problem and define some necessary terms. In the third section we will explore a few shrinkage estimators and go through their derivation and properties. In the fourth section, we will present a simulation study of the accuracy of these estimators. In the last section, we will discuss the results of the simulation study and their implications.

2. Background

The most desirable estimate of the population covariance will both be *positive definite* and *well-conditioned*, as well as having a minimized MSE.

Definition 2.1. *Let A be a $k \times k$ symetric matrix. Then A is* positive semidefinite *if* $0 \leq x^T Ax \ \forall x \neq \mathbf{0}$. *If* $0 < x^T Ax \ \forall x \neq \mathbf{0}$, *then A is* positive definite.

Fulfilling the criteria for positive definiteness allows us to quickly check if our matrix is invertible. However, first we need to introduce some notation that will be helpful moving forward.

Remark 2.2. *Let A be a $k \times k$ symmetric matrix with spectral decomposition $A = \sum_{i=1}^k \lambda_i e_i e_i^T$. Let the normalized vectors be the columns of another matrix $P = [e_1, e_2, ...]$. Then*

$$A_{(k\times k)} = P_{(k\times k)}\Lambda_{(k\times k)}P^T_{(k\times k)}, \tag{5}$$

where

$$PP^T = I \Rightarrow P^T = P^{-1}$$

and

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 & ... & 0 \\ 0 & \lambda_2 & ... & 0 \\ . & . & . & . \\ 0 & ... & ... & \lambda_k \end{bmatrix} \tag{6}$$

Theorem 2.1. *A real symmetric matrix A is positive definite if all of the eigenvalues of A are positive.*

Proof. Let A be a real symmetrix matrix. Then if A is a positive definite matrix, we can write

$$0 < x^T Ax = x^T P\Lambda P^T x,$$

where P and Λ are described as in 2.2 and $x \neq 0$. Then we can write

$$x^T P\Lambda P^T x = (P^T x)^T \Lambda (P^T x) = y^T \Lambda y,$$

where $y = P^T x$. Thus we have

$$\begin{bmatrix} y_1 & y_2 & ... & y_k \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 & ... & 0 \\ 0 & \lambda_2 & ... & 0 \\ . & . & . & . \\ 0 & ... & ... & \lambda_k \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ ... \\ y_k \end{bmatrix} = \lambda_1 y_1^2 + \lambda_2 y_2^2 + ... + \lambda_k y_k^2,$$

where $\lambda_i \in \mathbb{R}$. If one of these λ_i is negative (or zero), then we can choose x such that the $y_i = 0$ except for y_i corresponding to the negative λ_i, which is a contradiction. Thus we must have that $\lambda_i > 0$. □

Example 2.1.

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

is positive definite, since for any non-zero column vector $\boldsymbol{x} = [a\, b]^T$ we have

$$\boldsymbol{x}^T A\boldsymbol{x} = \begin{bmatrix} a & b \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = a^2 + b^2 > 0.$$

Recall that our estimator is a linear combination of two matrices. In fact, we will show in Section 3 that it is a linear combination of a positive definite matrix and a positive semidefinite matrix. Thus it follows that the sum is positive definite.

Theorem 2.2. *Let A be a real, symmetric positive definite matrix and B be a real, symmetric positive semidefinite matrix. Let $0 < \alpha < 1$. Then the linear combination*

$$\alpha A + (1-\alpha)B \tag{7}$$

is positive definite.

Proof. We want to show that

$$x^T(\alpha A + (1-\alpha)B)x > 0 \tag{8}$$

for all $x \neq 0$. Thus we need to show

$$\alpha x^T Ax + (1-\alpha)x^T Bx > 0 \tag{9}$$

for all $x \neq 0$. By definition we have $\alpha x^T Ax > 0$ for all $x \neq 0$, and similarly by definition $(1-\alpha)x^T Bx \geq 0$ for all x. Thus we are guaranteed that Equation (8) is satisfied for all $x \neq 0$, guaranteeing the positive definiteness of the sum and completing the proof. □

Also, recall that our estimator will always involve the sample covariance matrix. We now show that the sample covariance matrix is guaranteed to be positive semidefinite.

Theorem 2.3. *The sample covariance matrix $\boldsymbol{S} = \frac{1}{n-1}\mathbb{E}\left[(x-\overline{x})(x-\overline{x})^T\right]$ is positive semidefinite.*

Proof. For all vectors $u \neq 0$,

$$u^T\mathbf{S}u = u^T\left(\frac{1}{n-1}\mathbb{E}\left[(x-\overline{x})(x-\overline{x})^T\right]\right)u \tag{10}$$

$$= \frac{1}{n-1}\mathbb{E}\left[u^T(x-\overline{x})(x-\overline{x})^T u\right] \tag{11}$$

$$= \frac{1}{n-1}\mathbb{E}\left[\left(u^T(x-\overline{x})\right)^2\right] \tag{12}$$

$$= \left(u^T(x-\overline{x})\right)^2 \geq 0. \tag{13}$$

□

A second necessary condition for our estimator is that it is well-conditioned. Since we are exclusively working with real numbers, we simplify the definition of well-conditioned to the case dealing only in the reals.

Definition 2.3. *The **condition number** of a nonsingular matrix A with respect to norm $||\circ||$ is $K(A) = ||A||||A^{-1}||$. For any nonsingular matrix A and natural norm $||\circ||$,*

$$1 = ||I|| = ||AA^{-1}|| \leq ||A||||A^{-1}|| = K(A).$$

*A matrix A is **well-conditioned** if $K(A)$ is close to 1, and **ill-conditioned** if $K(A)$ is significantly greater than 1.*

Good-conditioning guarantees that the estimator is computationally feasible. Since a computer doesn't differentiate between true zero and 10^{-100}, for example, a good condidtion number tells us that we can actually compute our estimator. This is necessary for data analysis.

The class of shrinkage estimators, as we will prove in the next section, is always positive definite and well-conditioned, and since it is defined to be the convex combination which minimizes MSE for the associated target matrix, it exhibits all of the characteristics of a good estimate for Σ. In the next section we will cover the properties of some of the more widely used shrinkage estimators.

3. Shrinkage Estimators

As we covered in the previous section, ideal estimators are both positive definite and well-conditioned, and retain the information from the sample covariance matrix. In particular, some estimators exhibit all of this behavior and are distribution-free (i.e. do not make any assumptions about the underlying distribution of the data). Ledoit and Wolf's ([7], [6], [8]) proposed shrinkage estimator is distribution-free, and they provide a procedure for finding the optimal shrinkage intensity. Schäfer and Strimmer [11] take the Ledoit-Wolf estimator and apply it to six different target matrices (recall from (4) that a shrinkage estimate is a convex combination of the sample covariance matrix and a target matrix). Chen et al.[2] improve on the Ledoit-Wolf method to create the Rao-Blackwell Ledoit-Wolf (RBLW) method, which they prove outperforms the Ledoit-Wolf method when the data is Gaussian. Thus the assumption that the underlying data are Gaussian is best used if applying the RBLW method. They also create an oracle approximation shrinkage (OAS) estimator, both of which they show to outperform the Ledoit-Wolf estimator with Gaussian data. Chen et al.[3] improve on both LW and the OAS method assuming data from the elliptical family (e.g. Gaussian, Student's t, multivariate Cauchy, multivariate Exponential) by starting from a robust covariance estimator and iterating until convergence, which is distribution free within the elliptical family. The last of the shrinkage estimators we will consider is that of Chen et al.[1], which combines the shrinkage estimator with a tapering estimator. Tapering is a soft form of banding (used by Rothman et al.[10]). While banding sets the entries far away from the diagonal to be zero and keeps the entries within the band unchanged, tapering also gradually shrinks the off-diagonal entries within the band to zero. Banding only makes sense to use if there is a natural ordering to the variables.

3.1. **Ledoit-Wolf Estimator.**

The Ledoit-Wolf estimator (LW) enjoys properties which make it a good estimator. First, it is guaranteed to be well-conditioned (as we will see later, it is the weighted average of $\mathbf{S}$ and a structured estimator created to be well-conditioned, and thus the LW estimator inherits the good conditioning). Second, it is proven to be more accurate asymptotically than the sample covariance matrix. Also, as we mentioned above, it is distribution-free - no asumptions are made about the underlying distribution of the data - which is especially useful in cases where the data come from an unknown source or from a new technique. The LW estimator is easy to compute and interpret, since it is the asymptotically optimal convex linear combination of the sample covariance matrix and the identity matrix.

To guarantee that the LW estimator is both well-conditioned and accurate, the sample covariance matrix is shrunk to a structured matrix (in this case the identity, though we will see later that other structured matrices can be used) with an optimal weight identified by minimizing a quadratic loss function. The resulting matrix has smaller MSE asymptotically than the sample covariance matrix. The identity is chosen for the target matrix in order to inherit its good conditioning and to gain accuracy. Also, through the lemmas of Ledoit and Wolf [8], we can find a consistent estimator for the optimal weights λ_1 and λ_2 in Equation (4).

The full derivation of the LW estimator can be found in [8]. First, we start with the finite sample case. Let $X_{p\times n}$ be a data matrix distributed with mean zero and covariance Σ. Define the Frobenius norm to be $||A|| = \sqrt{\frac{tr(AA^T)}{p}}$, thus setting the norm of the identity to be one. Now denote the sample covariance matrix by $\mathbf{S}$, and the identity by I. We wish to find the estimators given in Equation (4) which minimizes $\mathbb{E}[||\Sigma^* - \Sigma||^2]$. We now set $\mathbf{T} = I$ and look for λ_1 and λ_2. In order to find Σ^* we need four parameters related to Σ. The dependence on Σ is not realistic (since we do not know Σ), so after defining Σ^* (i.e. finding λ_1 and λ_2) our task will be to find $\mathbf{S}^*$ which estimates Σ^* and has the same properties as Σ^* asymptotically. First define

$$< A_1, A_2 >= \frac{tr(A_1 A_2^T)}{p}. \tag{14}$$

The four variables necessary to define Σ^* are: $\mu =< \Sigma, I >, \alpha^2 = ||\Sigma - \mu I||^2$, $\beta^2 = \mathbb{E}[||\mathbf{S} - \Sigma||^2]$, and $\delta^2 = \mathbb{E}[||\mathbf{S} - \mu I||^2]$. Now we derive some lemmas and theorems which help to bring us to Σ^*.

Lemma 3.1. ***Ledoit and Wolf*** [8] ***Lemma 2.1, page 368.*** $\alpha^2 + \beta^2 = \delta^2$.

Proof.

$$\delta^2 = \mathbb{E}[||\mathbf{S} - \mu I||^2] = \mathbb{E}[||\mathbf{S} - \Sigma + \Sigma - \mu I||^2] \tag{15}$$

$$= \mathbb{E}[||\mathbf{S} - \Sigma||^2] + \mathbb{E}[||\Sigma - \mu I||^2] + 2\mathbb{E}[< \mathbf{S} - \Sigma, \Sigma - \mu I >] \tag{16}$$

$$= \mathbb{E}[||\mathbf{S} - \Sigma||^2] + \mathbb{E}[||\Sigma - \mu I||^2] + 2 < \mathbb{E}[\mathbf{S} - \Sigma], \Sigma - \mu I > . \tag{17}$$

Now since $\mathbb{E}[\mathbf{S}] = \Sigma$, the last term cancels and we get

$$\mathbb{E}[||\mathbf{S} - \Sigma||^2] + \mathbb{E}[||\Sigma - \mu I||^2] = \mathbb{E}[||\mathbf{S} - \Sigma||^2] + ||\Sigma - \mu I||^2 = \alpha^2 + \beta^2$$

by definition. □

Using the four variables defined above and Lemma 3.1, we can now write down Σ^*.

Theorem 3.2. ***Ledoit and Wolf*** [8] ***Theorem 2.1, page 368.***

$$\Sigma^* = \frac{\beta^2}{\delta^2}\mu I + \frac{\alpha^2}{\delta^2}\boldsymbol{S} \tag{18}$$

minimizes

$$\mathbb{E}[||\Sigma^* - \Sigma||^2] = \frac{\alpha^2\beta^2}{\delta^2} \tag{19}$$

for all Σ^* of the form $\Sigma^* = \lambda_1 I + \lambda_2 \boldsymbol{S}$.

Proof. Let $\lambda\nu = \lambda_1$ and $(1 - \lambda) = \lambda_2$. Then

$$\mathbb{E}[||\Sigma^* - \Sigma||^2] = \mathbb{E}[||\lambda\nu I + (1 - \lambda)\mathbf{S} - \Sigma||^2] \tag{20}$$

$$= \lambda^2||\Sigma - \nu I||^2 + (1 - \lambda)^2\mathbb{E}[||\mathbf{S} - \Sigma||^2], \tag{21}$$

where we can replace the expected value term by β^2.
First we restrict ourselves to minimizing $||\Sigma - \nu I||^2$ with respect to ν. Since $||I|| = 1$ by definition,

$$||\Sigma - \nu I||^2 = ||\Sigma||^2 - 2\nu < \Sigma, I > +\nu^2.$$

We differentiate with respect to ν and solve:

$$0 = -2 < \Sigma, I >= \nu^2 \Rightarrow \nu =< \Sigma, I >= \mu. \tag{22}$$

Now we can rewrite our previous condition as

$$\mathbb{E}[||\Sigma^* - \Sigma||^2] = \lambda^2||\Sigma - \mu I||^2 + (1 - \lambda)^2\mathbb{E}[||\mathbf{S} - \Sigma||^2] = \lambda^2\alpha^2 + (1 - \lambda)^2\beta^2. \tag{23}$$

In order to minimize MSE, we take the derivative with respect to λ and set equal to zero:

$$2\lambda\alpha^2 - 2(1 - \lambda)\beta^2 = 0 \Rightarrow \lambda = \frac{\beta^2}{\alpha^2 + \beta^2} = \frac{\beta^2}{\delta^2}, \tag{24}$$

and

$$1 - \lambda = \frac{\alpha^2}{\delta^2}. \tag{25}$$

Then

$$\mathbb{E}\left[||\Sigma^* - \Sigma||^2\right] = \left(\frac{\beta^2}{\delta^2}\right)^2\alpha^2 + \left(\frac{\alpha^2}{\delta^2}\right)^2\beta^2 = \frac{\alpha^2\beta^2}{\delta^2}. \tag{26}$$

Thus $\lambda_1 = \frac{\beta^2}{\delta^2}$ and $\lambda_2 = \frac{\alpha^2}{\delta^2}$, and the condition on the expectation is met. □

In the proof of Theorem 3.2 we notice that $\frac{\beta^2}{\delta^2}$ is the *shrinkage intensity*, and μI is the shrinkage target. If we now think in terms of MSE, we see that the expression $\mathbb{E}[||\Sigma^* - \Sigma||^2]$ can be broken into $\mathbb{E}[||\Sigma^* - \mathbb{E}[\Sigma^*]||^2] + ||\mathbb{E}[\Sigma^*] - \Sigma||^2$, where the first term denotes the variance and the second term denotes the bias (squared). We know that I has no variance, and thus its only contribution to the MSE is to increase the bias. We also know that because $\mathbf{S}$ is the MLE, it has no bias. Thus the contribution from $\mathbf{S}$ is only variance. Therefore, the linear combination Σ^* is the optimal tradeoff in terms of variance and bias. This tradeoff (and the balance achieved by the LW estimator) is shown in Figure 1. Here we see the MSE achieved by the optimal shrinkage intensity (the black dot along the MSE curve). We also see the variance curve (that is, the contribution from $\mathbf{S}$) and the squared bias curve (the contribution from I). Notice that as we increase the shrinkage intensity past the optimal point, we are lowering variance but increasing bias. Also, as we decrease the shrinkage intensity, we are lowering bias but increasing variance. This tradeoff is a delicate balance, and is illustrated in Figure 1.

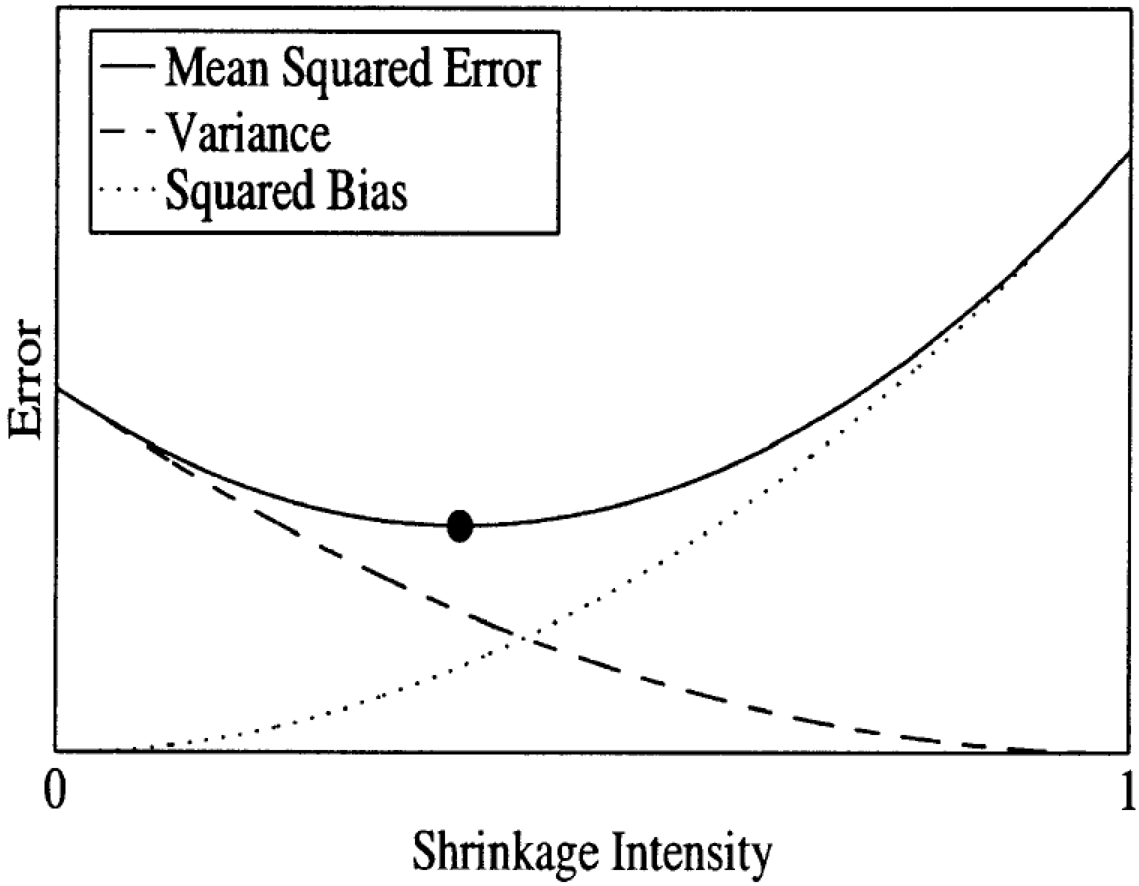


FIGURE 1. Tradeoff Between Variance and Bias in MSE Calculation for the LW Estimator [8].

However, Σ^* relies on μ, α^2, β^2, and δ^2, which are all scalar functions of the unknown Σ. Thus we must find the asymptotically consistent estimators for each of these four functions, which will in turn give us a consistent estimator of Σ^*. Usually in studying the asymptotic properties of an estimator, we would fix p and let $n \to \infty$. However, since our problems involve p on the same order as n or even larger, we can't fix p and let $n \to \infty$. Thus we let $p \to \infty$ at the same rate as $n \to \infty$. This is called *general asymptotics.* Then the optimal shrinkage intensity λ becomes a constant that we can estimate consistently, using the lemmas and theorems in Ledoit and Wolf (2004).

In the general asymptotics case, let $n = 1, 2, ...$ be the index in a sequence of models. Then for each n, X_n is a $p_n \times n$ matrix with n observations on each of p_n random variables. We assume the matrix X_n has been centered to have mean 0 and covariance Σ_n (the population covariance has a subscript merely to denote that the dimensions are set by n). The sample covariance be written as

$$\mathbf{S}_n = \frac{X_n X_n^T}{n}, \tag{27}$$

and now our four parameters from above become: $\mu_n =< \Sigma_n, I_n >$, $\alpha_n^2 = ||\Sigma_n - \mu_n I_n||^2$, $\beta_n^2 = \mathbb{E}[||\mathbf{S}_n - \Sigma_n||^2]$, and $\delta_n^2 = \mathbb{E}[||\mathbf{S}_n - \mu_n I_n||^2]$. We introduce new notation here with the n subscript to denote work in asymptotics, and for taking limits. We find the value of each of the four parameters for each fixed n, and work as $n \to \infty$.

Here we present the main theorems of Ledoit and Wolf, leaving the proofs to the dedicated reader who wishes to read the original paper. The proofs rely on a few assumptions - not normality of the data, however - and a deeper understanding of general asymptotics than we need to understand the results as they relate to our goals.

Lemma 3.3. ***Ledoit and Wolf*** [8] ***Lemma 3.1, page 377.*** *μ_n, α_n^2, β_n^2, and δ_n^2 remain bounded as $n \to \infty$.*

The boundedness of these four values is important for the results in the asymptotic case, since we wish μ_n, α_n^2, β_n^2, and δ_n^2 to be consistent. We decompose our covariance matrix into eigenvalues and eigenvectors

$$\Sigma_n = \Gamma_n \Lambda_n \Gamma_n^T,$$

where Λ_n is the diagonal matrix of eigenvalues and Γ_n is a rotation matrix with the columns the eigenvectors. Now set $Y_n = \Gamma_n^T X_n$, which is a $p_n \times n$ matrix. Then $(y_{11}^n, ..., y_{p_n 1}^n)^T$ is the first column of Y_n. Notice that Y_n is a matrix of iid observations on a system of p_n uncorrelated random variables. We obtain the following result.

Theorem 3.4. ***Ledoit and Wolf*** [8] ***Theorem 3.1, page 377.*** *Define $\Theta_n^2 = \mathrm{Var}[\frac{1}{p_n}\sum_{i=1}^{p_n}(y_{i1}^n)^2]$. Then Θ_n^2 is bounded as $n \to \infty$, and $\lim_{n\to\infty} \mathbb{E}[||S_n - \Sigma_n||^2] - \frac{p_n}{n}(\mu_n^2 + \Theta_n^2) = 0$.*

Theorem 3.4 tells us that the expected loss of the sample covariance is bounded, but is on the same order as $\frac{p_n}{n}(\mu_n^2 + \Theta_n^2)$. However, this quantity only goes to zero in special cases and thus the sample covariance matrix is only consistent under general asymptotics in only those special cases. This occurs when $\frac{p_n}{n}$ goes to zero and when μ_n^2 and Θ_n^2 go to zero (which implies that most of the random variables have variances which are zero asymptotically). Thus any error off the diagonal in the sample covariance matrix cause it not to be consistent under general asymptotics. In order to improve our estimate of the true covariance matrix, we turn to shrinkage. Due to Theorem 3.4 we see that the error of the sample covariance (denoted by β_n^2) goes to $\frac{p_n}{n}\mu_n^2$ as $n \to \infty$. Unless $\frac{p_n}{n}$ is negligable, shrinkage will improve our estimate by lowering the error (because as we showed in Theorem 3.2 shrinkage will achieve minimum MSE).
Now we want to find a consistent estimator for Σ_n^*. We already stated that Σ_n^* is not a practical estimator because it depends on Σ_n, which is unknown. Thus we estimate μ_n, α_n^2, β_n^2, and δ_n^2.

Lemma 3.5. ***Ledoit and Wolf*** [8] ***Lemma 3.2, page 379.*** *Define $m_n =< S_n, I_n >$. Then $\mathbb{E}[m_n] = \mu_n$ for all n, and $m_n - \mu_n$ converges to zero in quartic mean (fourth moment) as n goes to infinity.*

As a bit of notation, let $\underset{q.m.}{\to}$ denote convergence in quartic mean as $n \to \infty$. A consistent estimator for δ_n^2 is also its sample counterpart:

Lemma 3.6. ***Ledoit and Wolf*** [8] ***Lemma 3.3, page 379.*** *Define $d_n^2 = ||S_n - m_n I_n||^2$. Then $d_n^2 - \delta_n^2 \underset{q.m.}{\to} 0$.*

Set the vector x_k^n to be the kth column of X_n, and let $k = 1, 2, ..., n$. Then $S_n = \frac{X_n X_n^T}{n} = \frac{1}{n}\sum_{k=1}^n x_k^n (x_k^n)^T$. The matrices $x_k^n (x_k^n)^T$ are iid, so we can find the error β_n^2 of the average.

Lemma 3.7. ***Ledoit and Wolf*** [8] ***Lemma 3.4, page 380.*** *Define $\bar{b}_n^2 = \frac{1}{n^2}\sum_{k=1}^n ||x_k^n (x_k^n)^T - S_n||^2$ and $b_n^2 = \min(\bar{b}_n^2, d_n^2)$. Then $\bar{b}_n^2 - \beta_n^2 \underset{q.m.}{\to} 0$ and $b_n^2 - \beta_n^2 \underset{q.m.}{\to} 0$.*

The estimator b_n^2 ensures that Lemma 3.1 is satisfied, and that the following estimator is nonnegative.

Lemma 3.8. ***Ledoit and Wolf*** [8] ***Lemma 3.5, page 380.*** *Define $a_n^2 = d_n^2 - b_n^2$. Then $a_n^2 - \alpha_n^2 \underset{q.m.}{\to} 0$.*

Now we can replace the formula for Σ_n^* with our consistent estimators, which yields

$$S_n^* = \frac{b_n^2}{d_n^2} m_n I_n + \frac{a_n^2}{d_n^2} S_n. \tag{28}$$

The next theorems show that the asymptotic properties of the estimator are unchanged from the original.

Theorem 3.9. ***Ledoit and Wolf*** [8] ***Theorem 3.3, page 381.*** *S_n^* is a consistent estimator of Σ_n^*, since $||S_n^* - \Sigma_n^*|| \underset{q.m.}{\to} 0$. Thus $\mathbb{E}[||S_n^* - \Sigma_n||^2] - \mathbb{E}[||\Sigma_n^* - \Sigma_n||^2] \to 0$.*

Due to Theorem 3.9 here and Theorem 3.4 in [8] (page 380), we know that S_n^* is an asymptotically optimal linear shrinkage estimator of the sample covariance matrix (with respect to quadratic loss and under general asymptotics). Also recall that the convex linear combination of a positive definite matrix and a positive semidefinite matrix is positive definite, our result from Theorem 2.2. However, this is only true when the optimal shrinkage intensity is positive. In practice, the optimal shrinkage intensity might be calculated to be a negative number. Thus we constrain it to be positive in all calculations. Now we know that the identity is positive definite. The sample covariance matrix is guaranteed to be positive semidefinite by Theorem 2.3. Thus, we know that S_n^* is positive definite and therefore guaranteed to always be invertible, even when $p_n > n$, which makes it an ideal estimator of the sample covariance matrix. Theorem 3.5 in [8] tells us that S_n^* is generally well-conditioned, meeting the other criterion we established for a good estimator. Ledoit and Wolf also simplify their estimator in [7] to the form

$$\hat{\Sigma} = (1 - \lambda)S + \lambda I, \tag{29}$$

where λ is the optimal shrinkage intensity given in Lemma 3.8 as $\frac{b_n^2}{d_n^2}$. This is the general form equation that we will use to denote $\hat{\Sigma}$ for the rest of our discussion. Also notice that this is the same form as Equation (4) with $\lambda_1 = (1 - \lambda)$ and $\lambda_2 = \lambda$. They go on to say that in order to avoid overshrinkage and undershrinkage, it is acceptable to replace λ in Equation (29) with

$$\lambda^* = \max(0, \min(1, \lambda)). \tag{30}$$

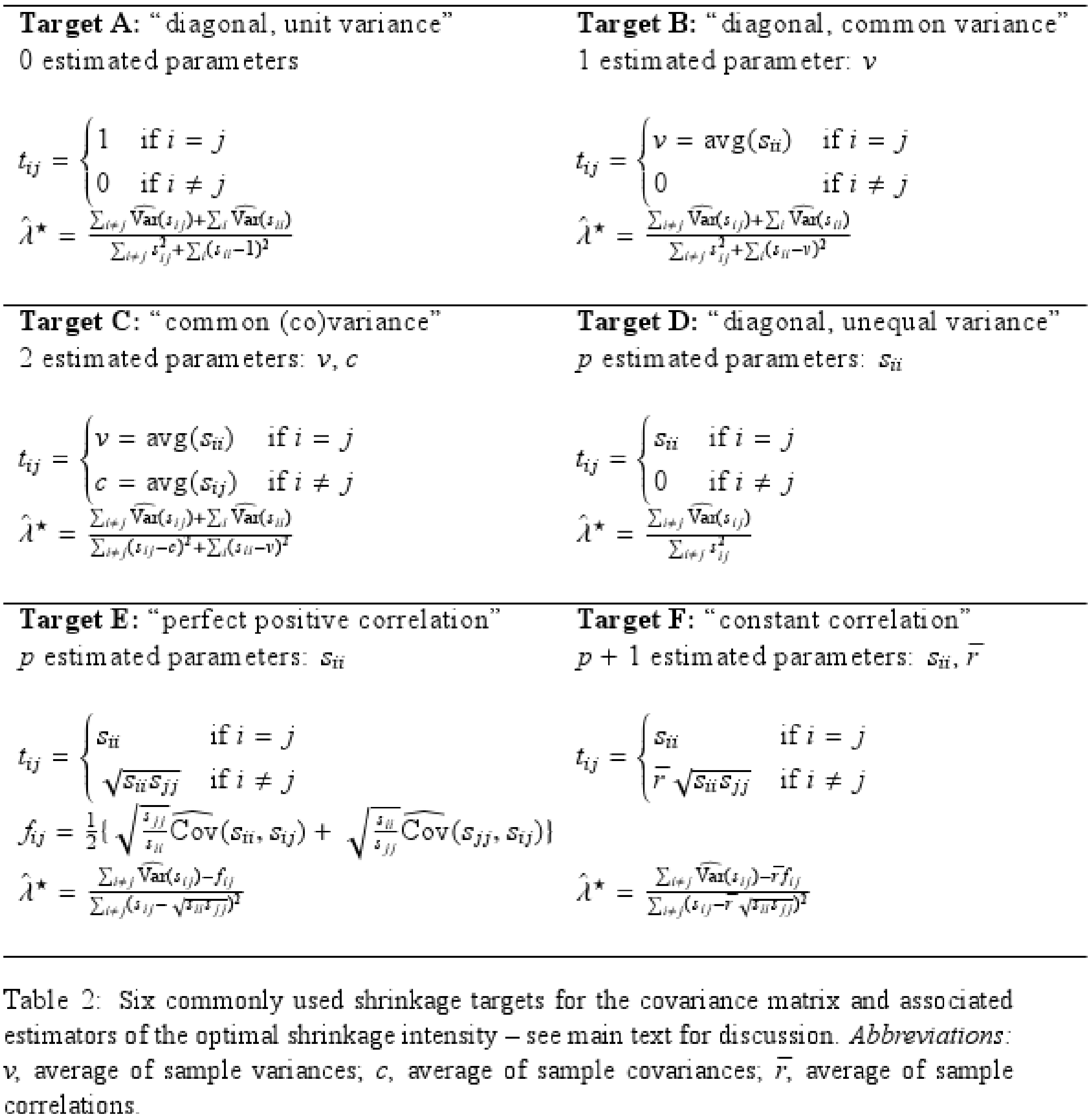
Here we constrain to $\lambda \in [0, 1]$ because in practice the optimal shrinkage intensity can be calculated to be negative. Notice that this solution does not take into account the problem that if the shrinkage intensity is zero, the estimate $\hat{\Sigma}$ is not positive definite. Ledoit and Wolf didn't need to take this into account since they only used one target matrix, but this is dangerous if we use a different target matrix.

3.2. **Target Matrices other than the Identity.** As we mentioned above, Schäfer and Strimmer, [11], take the LW estimator given in Equation (29) and apply it to six different target matrices. They again address the problem of $n << p$, but now instead of merely shrinking towards the identity they choose one of six target matrices, with equations of the form $\hat{\Sigma} = (1 - \lambda)S + \lambda T$, where S is the sample covariance matrix and T is the target matrix. However, choosing new matrices means that the optimal shrinkage intensity varies based on the target matrix chosen. Schäfer and Strimmer write the equation for the optimal shrinkage intensity for target matrix $T = (t_{ij})$ as

$$\lambda^* = \frac{\sum_{i=1}^{p} Var(s_i) - Cov(t_i, s_i)}{\sum_{i=1}^{p} \mathbb{E}[(t_i - s_i)^2]}. \tag{31}$$

In order to estimate λ^*, it is possible to replace all expectations, variances, and covariances by their sample counterparts.

At this point it is important to pick the target matrix. In order to maximize the benefits of shrinkage, the target should be chosen using the presumed structure of the data as a guide. Six commonly used targets (those presented in [11]) are given in Figure 2. Notice that each of these targets has a different equation for calculating the optimal shrinkage intensity λ^*. Target A is the identity, used in LW. Target B is the scalar multiple of the identity. A and B are two natural choices for the target matrix due to their simplicity. One does not have to make many assumptions about the underlying structure of the population covariance matrix to use the identity as the shrinkage target. Target C builds on target B by adding a common covariance. A, B, and C are all low dimensional, which is ideal because they impose a relatively strong structure on the resulting estimate, which then requires little data to fit. Also, estimators created using these targets shrink all elements of the sample covariance matrix (recall from Figure 2 that these three targets are the only ones presented which shrink the diagonal elements of the sample covariance matrix towards the target). The resulting estimator is a linear combination of a scaled version of the sample covariance with a scaled version of the target. Target A is used by Ledoit and Wolf in their papers, while target B is used in a few papers cited in [11]. Targets D, E, and F only shrink the off-diagonal



Target A: "diagonal, unit variance"
0 estimated parameters

$$t_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$

$$\hat{\lambda}^\star = \frac{\sum_{i \neq j} \widehat{\text{Var}}(s_{ij}) + \sum_i \widehat{\text{Var}}(s_{ii})}{\sum_{i \neq j} s_{ij}^2 + \sum_i (s_{ii} - 1)^2}$$

Target B: "diagonal, common variance"
1 estimated parameter: v

$$t_{ij} = \begin{cases} v = \text{avg}(s_{ii}) & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$

$$\hat{\lambda}^\star = \frac{\sum_{i \neq j} \widehat{\text{Var}}(s_{ij}) + \sum_i \widehat{\text{Var}}(s_{ii})}{\sum_{i \neq j} s_{ij}^2 + \sum_i (s_{ii} - v)^2}$$

Target C: "common (co)variance"
2 estimated parameters: v, c

$$t_{ij} = \begin{cases} v = \text{avg}(s_{ii}) & \text{if } i = j \\ c = \text{avg}(s_{ij}) & \text{if } i \neq j \end{cases}$$

$$\hat{\lambda}^\star = \frac{\sum_{i \neq j} \widehat{\text{Var}}(s_{ij}) + \sum_i \widehat{\text{Var}}(s_{ii})}{\sum_{i \neq j} (s_{ij} - c)^2 + \sum_i (s_{ii} - v)^2}$$

Target D: "diagonal, unequal variance"
p estimated parameters: s_{ii}

$$t_{ij} = \begin{cases} s_{ii} & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$

$$\hat{\lambda}^\star = \frac{\sum_{i \neq j} \widehat{\text{Var}}(s_{ij})}{\sum_{i \neq j} s_{ij}^2}$$

Target E: "perfect positive correlation"
p estimated parameters: s_{ii}

$$t_{ij} = \begin{cases} s_{ii} & \text{if } i = j \\ \sqrt{s_{ii} s_{jj}} & \text{if } i \neq j \end{cases}$$

$$f_{ij} = \frac{1}{2} \left\{ \sqrt{\frac{s_{jj}}{s_{ii}}} \widehat{\text{Cov}}(s_{ii}, s_{ij}) + \sqrt{\frac{s_{ii}}{s_{jj}}} \widehat{\text{Cov}}(s_{jj}, s_{ij}) \right\}$$

$$\hat{\lambda}^\star = \frac{\sum_{i \neq j} \widehat{\text{Var}}(s_{ij}) - f_{ij}}{\sum_{i \neq j} (s_{ij} - \sqrt{s_{ii} s_{jj}})^2}$$

Target F: "constant correlation"
$p + 1$ estimated parameters: s_{ii}, $\bar{r}$

$$t_{ij} = \begin{cases} s_{ii} & \text{if } i = j \\ \bar{r} \sqrt{s_{ii} s_{jj}} & \text{if } i \neq j \end{cases}$$

$$\hat{\lambda}^\star = \frac{\sum_{i \neq j} \widehat{\text{Var}}(s_{ij}) - \bar{r} f_{ij}}{\sum_{i \neq j} (s_{ij} - \bar{r} \sqrt{s_{ii} s_{jj}})^2}$$

Table 2: Six commonly used shrinkage targets for the covariance matrix and associated estimators of the optimal shrinkage intensity – see main text for discussion. *Abbreviations:* v, average of sample variances; c, average of sample covariances; $\bar{r}$, average of sample correlations.

Figure 2. Shrinkage Targets and Optimal Shrinkage Intensity [11].

elements. E and F were used in [7] to model stock returns, while D is used in [11].

The choice of D in [11] illustrates the problem of finding the correct target matrix. The authors are interested in inferring gene networks from small sample genomic data. Thus they choose D, which is a compromise between A, B, C (the low dimensional models) and E, F (high dimensional). Target D shrinks the off-diagonal entries toward zero, but keeps the variances unchanged. This separation of variance and covariance assumes that the parameters of the covariance matrix fall into two distinct classes which can be treated separately by shrinkage.

3.3. Rao-Blackwell Ledoit-Wolf Estimator.

One of the desirable properties of the LW estimator was that it was distribution-free. This allows us to approach data with a consistent estimator that performs better than the sample covariance matrix without making many assumptions about the structure of the data. However, if we are confident that the data comes from a Gaussian distribution, then the LW estimator can be improved upon. Applying the Rao-Blackwell theorem to the LW method results in a new estimator that we will call the RBLW estimator. Again our goal is to minimize MSE of the estimator, so we use the format $\hat{\Sigma} = \lambda_1 \mathbf{S} + \lambda_2 \mathbf{T}$, where now $\mathbf{T} = \frac{tr(\mathbf{S})}{p} I$. Chen et al. [2] make this choice for $\mathbf{T}$ because it is well-conditioned. In this case, as with the LW estimator, the expression for $\hat{\Sigma}$ simplifies to

$$\hat{\Sigma} = (1-\lambda^*)\mathbf{S} + \lambda^*\mathbf{T}. \tag{32}$$

As above, $||\circ||$ denotes the Frobenius norm. Applying the LW method to calculate the optimal shrinkage intensity for this target matrix gives

$$\lambda^*_{LW} = \frac{\sum_{i=1}^{n} ||x_i x_i^T - \mathbf{S}||^2}{n^2 \left[tr(\mathbf{S}^2) - \frac{tr^2(\mathbf{S})}{p} \right]}. \tag{33}$$

Then the LW estimator is determined by plugging (33) into (32). Now under the Gaussian assumption on the data, $\mathbf{S}$ is a sufficient statistic for estimating Σ. Simply by looking at the LW estimator, we notice that it is a function of $\mathbf{S}$ and of other statistics (namely $\mathbf{T}$). We now introduce the Rao-Blackwell theorem, as it will help us in improving the LW estimator. First we present the notion of a sufficient statistic.

Definition 3.1. *A statistic $T(X)$ is* sufficient *for an underlying parameter θ if $\mathbb{P}(X = x|T(X) = t, \theta) = \mathbb{P}(X = x|T(X) = t)$.*

Theorem 3.10. [14] *Let $\hat{\theta}$ be an estimator of θ with $\mathbb{E}(\hat{\theta}^2) < \infty$ for all θ. Suppose that T is sufficient for θ, and set $\theta^* = \mathbb{E}(\hat{\theta}|T)$. Then for all θ*

$$\mathbb{E}(\theta^* - \theta)^2 \le \mathbb{E}(\hat{\theta} - \theta)^2. \tag{34}$$

This theorem tells us that we can improve on our original estimator, or at least stay constant. Now we apply the Rao-Blackwell theorem to the LW estimator to obtain the RBLW estimator. The proof of Theorem 3.11 can be found in [2].

Theorem 3.11. ***Chen et al.*** [2] ***Theorem 2, page 5018.*** *Let $X_{p\times n}$ be a matrix consisting of n independent p-dimensional Gaussian vectors, and covariance matrix Σ. Let $\boldsymbol{S}$ be the sample covariance matrix. Then the conditional expectation of the LW estimator is*

$$\hat{\Sigma}_{RBLW} = \mathbb{E}[\hat{\Sigma}_{LW}|\boldsymbol{S}] = (1-\lambda^*_{RBLW})\boldsymbol{S} + \lambda^*_{RBLW}\boldsymbol{T}, \tag{35}$$

where

$$\lambda^*_{RBLW} = \frac{\frac{(n-2)}{n} tr(\boldsymbol{S}) + tr^2(\boldsymbol{S})}{(n+2)\left[tr(\boldsymbol{S}^2) - \frac{tr^2(\boldsymbol{S})}{p} \right]}. \tag{36}$$

This estimator satisfies that

$$\mathbb{E}[||\hat{\Sigma}_{RBLW} - \Sigma||^2] \le \mathbb{E}[||\hat{\Sigma}_{LW} - \Sigma||^2] \tag{37}$$

for all Σ.

Similarly to the LW estimator, the shrinkage intensity is modified to avoid overshrinkage to be

$$\lambda^*_{RBLW} = min(1, \lambda^*_{RBLW}). \tag{38}$$

The RBLW estimator inherits the property of achieving the asymptotically minimum MSE from the LW estimator (since the RBLW and LW estimators are equivalent asymptotically [2]). However, as is the case with the LW estimator, for very small n the asymptotics do not hold and we are not guaranteed to have minimum MSE.

3.4. The OAS Estimator.

The justification for the oracle approximating shrinkage (OAS) estimator is developing an estimator that works for small n. Rather than considering asymptotic solutions, we employ an iterative process. Consider the case of $n = 2$: λ^*_{RBLW} and λ^*_{LW} are close to 1, and thus $\hat{\Sigma} \approx \mathbf{S}$ for each. Ideally, we would be closer to the target matrix. Thus, the iterative approach. We take an initial guess at Σ, $\hat{\Sigma}_0$, which is the sample covariance or any other symmetric nonnegative definite estimator. Then the iterative process generates a new estimate for $\hat{\Sigma}$, and this continues until convergence. The authors of [2] denote the solution as $\hat{\Sigma}_{OAS}$. The iterative process is:

$$\lambda_{j+1} = \frac{\left(\frac{1-2}{p}\right) tr(\hat{\Sigma}_j \mathbf{S}) + tr^2(\hat{\Sigma}_j)}{\left(\frac{n+1-2}{p}\right) tr(\hat{\Sigma}_j \mathbf{S}) + \left(\frac{1-n}{p}\right) tr^2(\hat{\Sigma}_j)} \tag{39}$$

$$\hat{\Sigma}_{j+1} = (1 - \lambda_{j+1})\mathbf{S} + \lambda_{j+1}\mathbf{T}. \tag{40}$$

However, these equations lead to the following theorem which gives us a solution for the optimal shrinkage intensity and the estimator itself:

Theorem 3.12. ***Chen et al.*** [2] ***Theorem 3, page 5019.*** *Let* $\lambda_0 \in [0,1]$ *be an initial guess of the shrinkage intensity. Then the iterations specified in* (39) *and* (40) *converge as* $j \to \infty$ *to*

$$\lambda^*_{OAS} = min\left(\frac{\left(\frac{1-2}{p}\right) tr(\boldsymbol{S}^2) + tr^2(\boldsymbol{S})}{\left(\frac{n+1-2}{p}\right) \left[tr(\boldsymbol{S}^2) - \frac{tr^2(\boldsymbol{S})}{p}\right]}, 1 \right) \tag{41}$$

and

$$\hat{\Sigma}_{OAS} = (1 - \lambda^*_{OAS})\, \boldsymbol{S} + \lambda^*_{OAS}\, \boldsymbol{T}. \tag{42}$$

The proof of Theorem 3.12 is in [2]. Notice that under asymptotic conditions ($p \to \infty$ and $n \to \infty$) the OAS solution and RBLW solution converge to each other, and are equivalent to the LW solution [2]. However, in small sample situations (take the $n = 2$ extreme case, for example) the OAS and RBLW estimators behave entirely differently - $\lambda^*_{OAS} \approx 1$ while $\lambda^*_{RBLW} \approx 0$. Thus choosing an estimator becomes very important especially in small sample cases. When the samples are truly Gaussian, the authors in [2] show that the RBLW and OAS estimators perform better than the LW estimator. However, violations of the Gaussian assumption change the small sample behavior, creating a need for estimators which are robust to this assumption.

3.5. **Chen Estimator.** In order to be distribution-free (to an extent - no assumptions must be made as long as the data comes from the elliptical family, so fewer assumptions are made than any estimator besides the LW estimator), Chen et al. [3] work in the elliptical class of distributions. These include the Gaussian, multivariate t, and many others.

Definition 3.2. [3] *Let x be a $p \times 1$ random vector generated by $x = \nu u$, where ν is a positive real random variable and u is a $p \times 1$ real Gaussian random vector with mean zero and positive definite covariance Σ. Then x is* elliptically distributed *and the pdf of x can be written as*

$$p(x) = \phi\left(x^T \Sigma^{-1} x\right), \tag{43}$$

where ϕ is the characteristic function related to the pdf of ν.

Here the characteristic function is defined by Gabriel Frahm [4]. For example, the characteristic function of the centered multivariate normal is $e^{-\frac{1}{2}t^T \Sigma t}$ Again, an iterative process determines the estimator. Recall that the sample covariance is defined in Equation (2). In order to work around the problems with the sample covariance in high dimension, Chen et al. use Tyler's method [13] and work with the normalized samples $s_i = \frac{x_i}{||x_i||_2}$, where $||\circ||_2$ is the 2-norm, and $||\circ||$ still denotes the Frobenius norm. They state that the maximum likelihood estimate of Σ will be of the form

$$\hat{\Sigma} = \frac{p}{n} \sum_{i=1}^{n} \frac{s_i s_i^T}{s_i^T \hat{\Sigma}^{-1} s_i}. \tag{44}$$

To find this solution, we iterate between

$$\tilde{\Sigma}_{j+1} = (1 - \lambda)\frac{p}{n} \sum_{i=1}^{n} \frac{s_i s_i^T}{s_i^T \hat{\Sigma}_j^{-1} s_i} + \lambda I \tag{45}$$

and

$$\hat{\Sigma}_{j+1} = \frac{p\tilde{\Sigma}_{j+1}}{tr\left(\tilde{\Sigma}_{j+1}\right)} \tag{46}$$

until convergence. Assuming we know Σ (and therefore λ, the optimal shrinkage intensity) we can write down the "estimator" (since we have no real knowledge of Σ, this estimator is useless to us until we can approximate it)

$$\tilde{\Sigma} = (1-\lambda_O)\frac{p}{n}\sum_{i=1}^{n}\frac{s_i s_i^T}{s_i^T\hat{\Sigma}_j^{-1}s_i} + \lambda_O I \tag{47}$$

where λ_O minimizes MSE as

$$\lambda_O = arg\ min_{\lambda}\mathbb{E}\left[||\tilde{\Sigma}-\Sigma||^2\right]. \tag{48}$$

The following theorem provides a closed form solution to Equation (48):

Theorem 3.13. ***Chen et al.*** [3]***, page 4099.*** *Let $tr(\Sigma) = p$. Then for iid elliptical samples, the solution to Equation* (48) *is*

$$\lambda_O = \frac{p^2 + \frac{1-2}{p}tr(\Sigma^2)}{(p^2 - np - 2n) + \left(n + 1 + \frac{2(n-1)}{p}\right)tr(\Sigma^2)}. \tag{49}$$

The proof of this theorem is contained in the Appendix of [3]. Since we do not know Σ, an estimate of λ_O is proposed where each instance of Σ is replaced with $\hat{M}$, a consistent estimator of Σ. An example of $\hat{M}$ is the LW estimator, while another (used by the authors of [3]) is

$$\hat{R} = \frac{p}{n}\sum_{i=1}^{n} s_i s_i^T, \tag{50}$$

the trace-normalized sample covariance matrix. Chen et al. use the trace-normalized sample covariance to keep constant with their earlier use of Tyler's method. Then we can use fixed-point iteration applied to Equations (45) and (46) to find the estimator.

3.6. **Shrinkage-to-Tapering Estimator.** All of the previous estimators we have considered thus far rely on p and n going to ∞ at roughly the same rate. However, what if this is not the case? In some situations (Chen et al. [1] apply this estimator to breast cancer gene expression data), p goes to ∞ much faster than n. To deal with this situation, Chen et al. [1] combine the strengths of both the shrinkage estimator and the tapering estimator. The estimator has the same form as the general shrinkage estimator in (29) but now the shrinkage target is a tapered version of the sample covariance matrix. A *tapering* estimator - as stated by T. Cai et al. [12] - takes each entry in the proposed matrix and multiplies it by a weight. This estimator has been shown to be consistent when the dimensionality p grows at any sub-exponential rate of n, which allows us to consider much higher-dimensional matrices [1]. For instance, starting with the sample covariance matrix, we could have weight $\frac{1}{s_{ij}}$ along the diagonal and zero elsewhere to create the identity. The weights can be chosen so that the tapering is done however one desires. A more formal definition is provided by Chen et. al [1] (with the notation that S is the set of all $p \times p$ symmetric matrices and $A \circ B = (a_{ij}b_{ij})$):

Definition 3.3. ***Chen et al.*** [1] ***Definition II.1, page 5641.*** *Let S be the set of $p \times p$ smmetric matrices. A* covariance matrix taper *$A \in S$ satisfies*

$$\sum_{j=1}^{p}\nu_j(A\circ B) \leq \sum_{j=1}^{p}\nu_j(B) \tag{51}$$

for all $B \in S$, where ν_i denotes the ith eigenvalue.

Thus, element-wise multiplication with any covariance matrix taper is guaranteed to decrease the average eigenvalue. Then we can define a *tapering estimator*:

$$\hat{\Sigma}_{taper} = W \circ \mathbf{S}, \tag{52}$$

with W a covariance matrix taper.

Now define W as

$$w_{ij} = \begin{cases} 1, & \text{for } |i-j| \leq k/2 \\ 2 - \frac{|i-j|}{k/2}, & \text{for } k/2 < |i-j| < k \\ 0, & \text{else} \end{cases} \tag{53}$$

Notice that since $diag(W) = 1$, W is a covariance matrix taper. Thus

$$\sum_{j=1}^{p} \nu_j(W \circ \Sigma) = tr(W \circ \Sigma) = tr(\Sigma) = \sum_{j=1}^{p} \nu_j(\Sigma). \tag{54}$$

Also, the way we have defined W sets k as its *bandwidth*, since any entry at least $k-1$ off of the diagonal is set to zero. Last, we must modify $W \circ \Sigma$ since it is not necessarily positive definite [1]. If we diagonalize $W \circ \Sigma$ and then replace the negative eigenvalues with zeros, we will ensure the positive semi-definiteness of the new estimate. According to [1], the optimal bandwidth of W under Frobenius risk (which we have considered previously) is $n^{\frac{1}{2(\alpha+1)}}$, where $\alpha > 0$ is a smoothing parameter which specifies the rate of decay of the off-diagonal elements in Σ [1]. However, this taper relies on a natural ordering of the variables, which is very dependent on the situation. Thus we have to assume order to use this estimator.

The proposed estimator now is a shrinkage estimator following the format of (29), but the shrinkage target is the tapered version of the sample covariance matrix. That is, we have a *shrinkage-to-tapering oracle* (STO) estimator

$$\hat{\Sigma}_{STO} = (1 - \lambda^*_{STO})\mathbf{S} + \lambda^*_{STO}(W \circ \mathbf{S}). \tag{55}$$

The optimal shrinkage intensity λ^*_{STO} is the solution to

$$\hat{\lambda}^*_{STO} = \arg\ \min_{\lambda \in [0,1]} \mathbb{E}\left[||\hat{\Sigma}_{STO} - \Sigma||^2 \right]. \tag{56}$$

In [1] we learn that for any covariance matrix the STO estimator can improve on both the tapering and shrinkage oracle estimators. Then we can determine the optimal shrinkage intensity.

Theorem 3.14. *Chen et al.* [1] ***Theorem III.1, page 5645.*** *The optimal shrinkage intensity of the STO estimator under minimum Frobenius risk is*

$$\hat{\lambda}^*_{STO} = \frac{\mathbb{E}\left(||\boldsymbol{S}||^2 - ||V \circ \boldsymbol{S}||^2\right) - \left(||\Sigma||^2 - ||V \circ \Sigma||^2\right)}{\mathbb{E}\left(||\boldsymbol{S}||^2\right) + \mathbb{E}\left(||W \circ \boldsymbol{S}||^2\right) - 2\mathbb{E}\left(||V \circ \boldsymbol{S}||^2\right)}, \tag{57}$$

where $V = (v_{ij}) = \sqrt{w_{ij}}$.

However, as in many of out previous discussions, this estimate involves Σ which is unknown. Chen et al. then suggest an approximating algorithm. We initialize with $\hat{\Sigma}_0 = \mathbf{S}$, and replace Σ by $\hat{\Sigma}_0$ in (57). Then we get the next estimate $\hat{\Sigma}_1$. This process is iterated until convergence. The optimal bandwidth k^* is determined by cross-validation to minimize prediction error on test data.

4. Simulations

We now run simulations using the Ledoit-Wolf estimator (Target A in Schäfer and Strimmer), the rest of the Target Matrices, a new estimator devised by Schäfer et al. [11] [9] for an R package, and my implementation of their method. The first simulation takes all of the estimators and for fixed n and p (starting at 200 and 200), simulates a true covariance matrix using a method devised by Hardin et al. [5], and then calculates the estimator using data generated from the multivariate normal first and then the multivariate normal with outliers. The multivariate normal used has mean 5 and standard deviation 1 for each of the variables, and follows Σ for its covariance structure. To simulate outliers, we calculate the sample covariance matrix using 90% normal data and 10% multivariate normal data with mean 50 and standard deviation 1. Then for each estimator $\hat{\Sigma}$ we calculate $||\hat{\Sigma} - \Sigma||^2$, and repeat this process 100 times. These steps are performed first using the Pearson method of correlation calculation and then with the Spearman method.

Next, we fix p at 500 and then increment $n = (10, 100, 400)$. We again calculate MSE as we did in the first simulation, and add the "corpcor" estimate. Here we only work with the Pearson correlation due to issues in computation time and the fact that the "corpcor" estimate does not allow calculation with Spearman correlation. We also add the estimator calculated by the "corpcor" package in R developed by Schäfer et al. [11] [9]. The estimator calculated in the R package is different from their paper, in that it calculates only one target. The shrinkage intensity is calculated as

$$\lambda^* = \frac{\sum_{k=1}^{p} \mathbb{V}(s_{kk})}{\sum_{k=1}^{p} (s_{kk} - median(s))^2}, \tag{58}$$

where $median(s)$ denotes the median of the sample variances. This is the last estimator that we track. We also track the calculated shrinkage intensity as we increment n. We plot the calculated shrinkage intensity for the targets A-F as well as "corpcor" and the "median" value calculated in Equation (58). Even though using the "corpcor" function "estimate.lambda.var" under the base conditions should result in the same thing as the "median", there is enough blackboxed by the "corpcor" package that it is an interesting exercise to plot both. On both figures each data point represents the average of 10 iterations of the simulation.

Last, with p fixed at 500, we increment $n = (5, 10, 20, 50, 100, 200, 300, 400, 500)$ and again calculate MSE as we did in the previous simulation. The data is first distributed normally according to the multivariate distribution described above, and then according to the outlier model. Here each data point represents the average of 100 iterations of the simulation.

4.1. Pearson vs Spearman Correlation and the Six Target Matrices.

Looking at Figure 3, notice that in (a) Target D performs best, though only slightly. As we move to (b) and the Spearman correlation, notice that not only does the median value of $MSE(\hat{\Sigma} - \Sigma)$ double on the log scale (from around 11.6 to about 22) but Targets A, B, and F outperform all other targets by a large margin. In (c), working with outliers and the Pearson correlation, all targets perform the same on average, but it is much worse than in (a), where outliers are not present. Last, in (d), Targets A, B, and F again outperform all of the other targets when outliers are present. Here the median value is similar to the median value in (c), but individual targets (A, B, and F) perform much better than they did using Pearson correlation.

In the case of normal data with no outliers, using the Spearman correlation coefficient raises MSE. Interestingly, Target E is incredibly affected while Targets C and D are fairly highly affected and A, B, and F are only moderately affected. Part of this discrepancy could come from the calculation of the shrinkage intensity for each target. If we refer back to Figure 2, we notice that the covariance between values in the sample covariance matrix plays a large role in the calculation of Target E's shrinkage intensity. If this value is large enough, the calculated intensity could become negative, leading the used shrinkage intensity to be zero. This is problematic, because $\hat{\Sigma}$ is no longer guaranteed to be positive definite.

Now in the case of data with outliers, using the Spearman correlation coefficient in calculations actually lowers the MSE. We see a similar pattern to (b) in Figure 3, with the slight difference being that Target D now has a lower MSE than Target C.

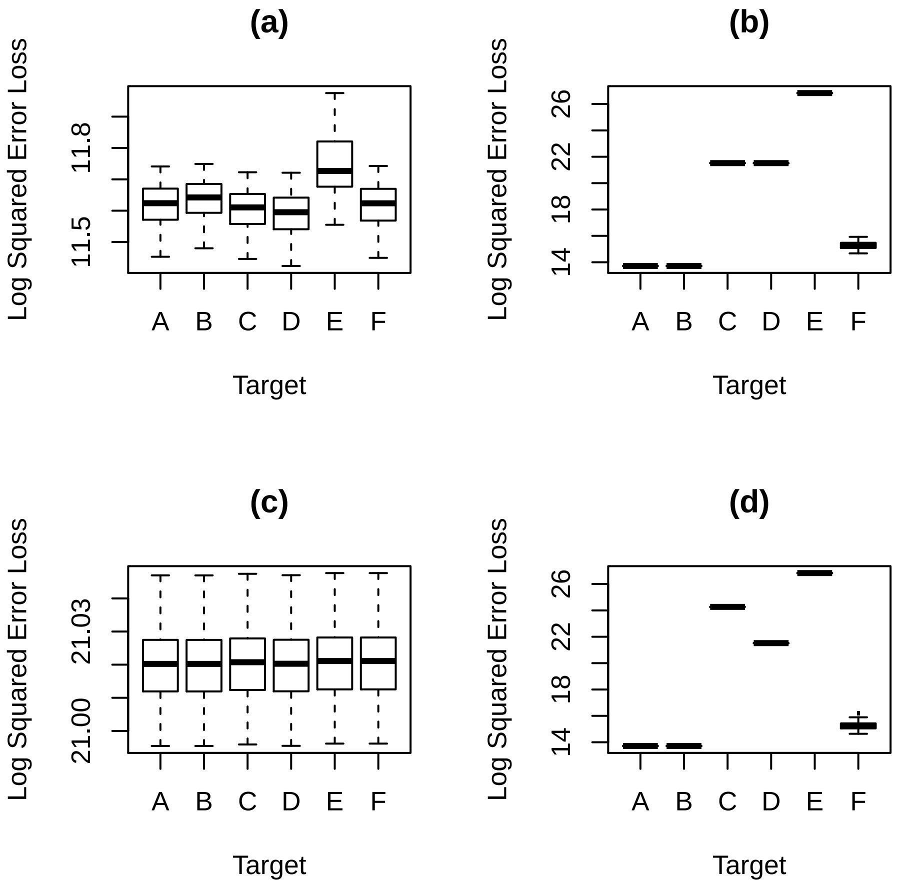
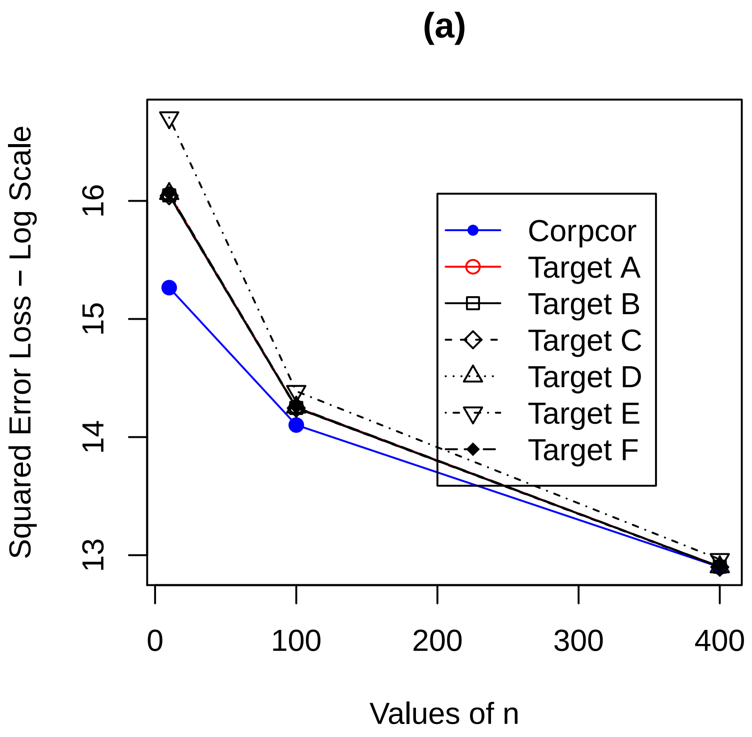


FIGURE 3. Log $MSE(\hat{\Sigma}-\Sigma)$ calculated for each of the six target matrices laid out in Figure 2. (a) Multivariate data with Pearson correlation, (b) Multivariate data with Spearman correlation, (c) Data with outliers and Pearson correlation, (d) Data with outliers and Spearman correlation.



4.2. **Shrinkage Intensity and MSE for Selected Values of n.** In Figure 4, as n increases, the MSE for each approaches a common value. We expect this to happen, because the target is especially important when n is small and the sample covariance matrix is a particularly poor estimator of Σ. As $n \to p$, however, we expect $\lambda \to 0$ and $\hat{\Sigma} \to \mathbf{S}$. This is natural, because when $n \geq p$, $\mathbf{S}$ is the MLE and the best estimate of Σ, and is also invertible. The Corpcor estimate consistently has the lowest MSE, which is to be expected since not only is it more recent than the others (developed in 2007 rather than 2005) but it is also code developed and optimized by the authors, rather than the others which were developed for use in this paper and not fully optimized. All of the other targets beside Target E perform the same. Target E is interesting, because as we noted earlier the covariance plays a large role and can sometimes drive the shrinkage intensity below zero, thus making the matrix used the sample covariance matrix. This has large problems with small n and large p, as we see in Figure 4. However, as $n \to p$, the estimate approaches the others in terms of MSE. Thus for this data we should use the Corpcor estimate for $n << p$, but as $n \to p$ the estimator doesn't have a huge effect.

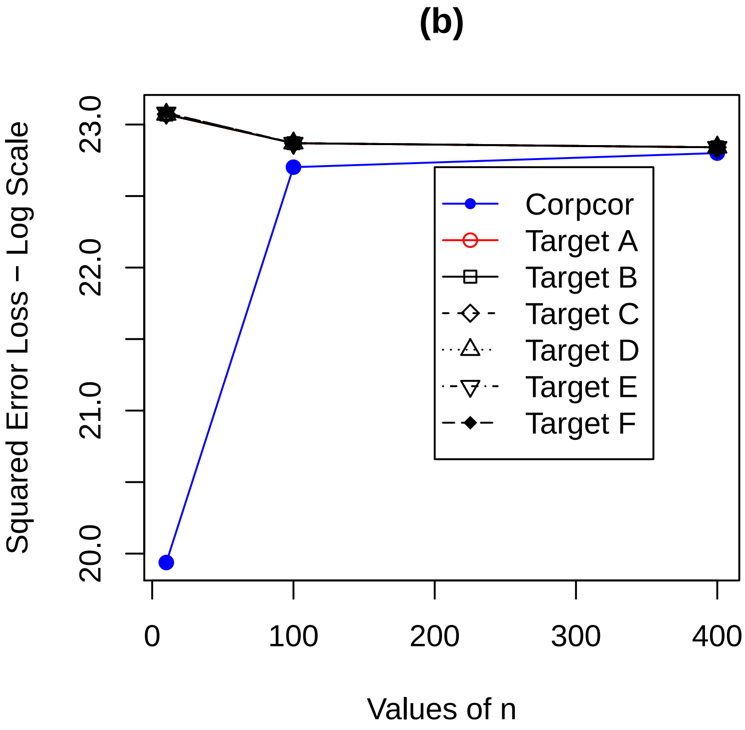


Figure 5. Log MSE calculated for each of the six target matrices laid out in Figure 2, plus the “Corpcor” estimator from the package “corpcor” [9]. The data is generated from the multivariate normal with outliers.

Figure 5 shows an interesting trend. This data has outliers, and here the Corpcor estimate performs significantly better with small n and actually does worse as $n \to p$. All of the other estimates perform the same. These values are significantly higher than their normal data counterparts. Similar to our prior conclusion, here we definitely use Corpcor for $n << p$, and only when $n \approx p$ can we justify using the others.

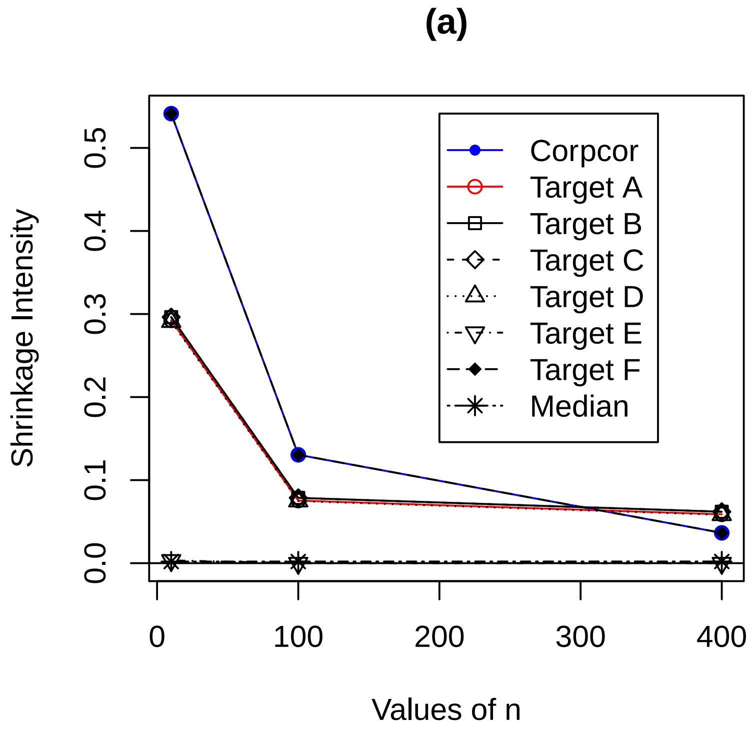


Figure 6. Shrinkage intensity calculated for each of the six targets laid out in Figure 2 plus "Corpcor" and the "median" value from [9]. We ran at three values of n, and the data is generated from the multivariate normal.

Now consider Figure 6, which provides the calculated shrinkage intensities for each of the estimates used in Figure 4. Notice that both Median and Target E have essentially zero as the optimal shrinkage intensity across the board. The slight increase we notice accounts for the decrease in MSE as $n \to p$, but the fact that the optimal intensity is so close to zero means that these are still the worst estimates. In the case of Target E, this means that the optimal intensity was calculated as less than zero and was brought up by our stipulation that the intensity must be in $[0, 1]$. However, Median doesn't suffer from this same issue. Interestingly, the Corpcor estimated shrinkage intensity is quite different from than calculated simply using Equation (58). Clearly there is something else going on in the corpcor package that produces this discrepancy. Also notice that Target F has the same calculated shrinkage intensity as Corpcor, but it performs worse than Corpcor in Figure 4. This must have to do with the target matrix chosen for Corpcor. All of the other targets have shrinkage intensity consistent with their performance in Figure 4.

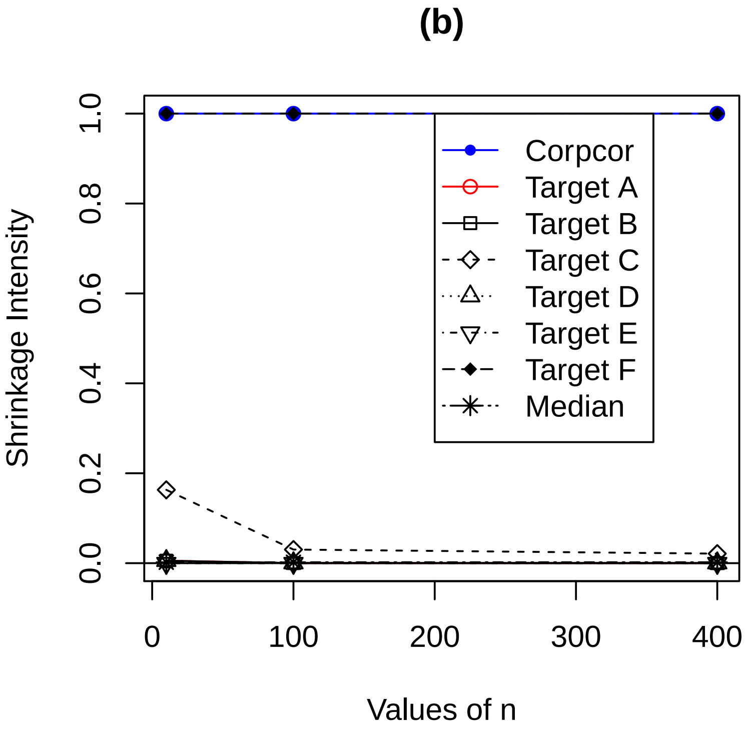


FIGURE 7. Shrinkage intensity calculated for each of the six targets laid out in Figure 2 plus "Corpcor" and the "median" value from [9]. We ran at three values of n, and the data is generated from the multivariate normal with outliers.

However, Figure 7 has the interesting information. Notice that the Corpcor estimate is close to one, corresponding to using almost entirely the target matrix, which is not known. The median, supposedly the way that "corpcor" calculates the optimal shrinkage intensity, is close to zero. This, as we know, corresponds to using almost entirely the sample covariance matrix. The incredible difference between these two values means that the package definitely performs some other calculations to determine the optimal shrinkage intensity. Now setting aside the Corpcor estimate, consider the others. They are all incredibly close to zero - except for F, which again is the same as Corpcor - which is unexpected given that the sample covariance matrix is a very bad estimate in the small n cases. Also, Target D is similar to all of the others, which is not the case in the normal data. Equally confusing is that Target C has a higher optimal shrinkage intensity, but there is no noticeable difference in its MSE. Here again we see the discrepancy between the MSE of Target F and Corpcor, while the optimal shrinkage intensity is calculated to be the same.

The fact that the MSE of the Corpcor estimate increases to approach the others as $n \to p$ indicates that as $n \to p$ the sample covariance matrix becomes a better estimate of the covariance matrix - which we expected - and that the target used to calculate the Corpcor estimate does a poorer job. It is unexpected that even when $n = 100$, which is $\frac{1}{5}p$, the Corpcor estimate is only slightly better than the other estimates. The huge difference in optimal shrinkage intensity leads us to believe that $\hat{\Sigma}$ will be different, but the estimates must not vary as much as we would expect. Perhaps if we ran trials with much higher dimension we would see more variation in the estimates, but this requires much more computation power than is available to us.

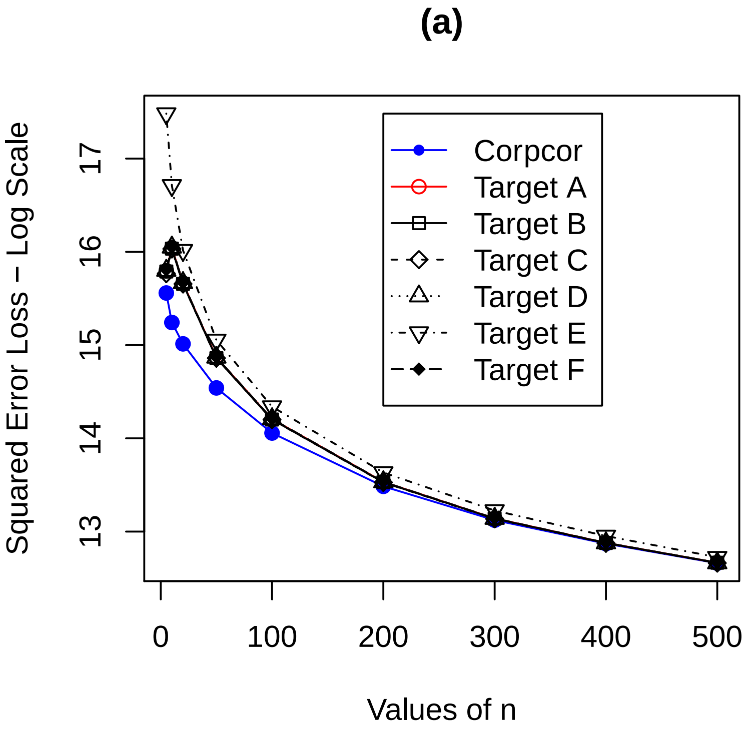


FIGURE 8. Log MSE calculated for each of the six target matrices laid out in Figure 2, plus the "Corpcor" estimator from the package "corpcor" [9]. The data is generated from the multivariate normal.

4.3. MSE **for a Range of Values of** n**.** Consider Figure 8, where we have run many different values of n to watch the approach as $n \to p$. Again, the Corpcor estimate has the lowest MSE and the estimates all approach the same value as $n \to p$. However, there are two interesting trends to notice. First is that Target E has a significantly higher MSE for the small values of n - 5, 10, and 20 - but is similar for the others. This echoes our results discussed above but is seen on a much larger scale here. Also, there is a strange spike up between $n = 5$ and $n = 10$ for the other Targets. In fact, the difference in log MSE between them and Corpcor is very small. It is very odd that for this incredibly small value of n the estimates should be so similar, and simply increasing n to 10 not only causes a jump up in the Target estimates but also sees a decrease in MSE in the Corpcor estimate. The $n = 5$ case must be special in that the data is so small that the estimate is very similar. The jump in MSE could possibly be explained by added noise from the added values outweighing the increase in sample size. This effect lowers with increased n after 10, as we see a decrease in MSE. The plot also adds power to the trends observed our previous figures with the extra iterations of the simulation.

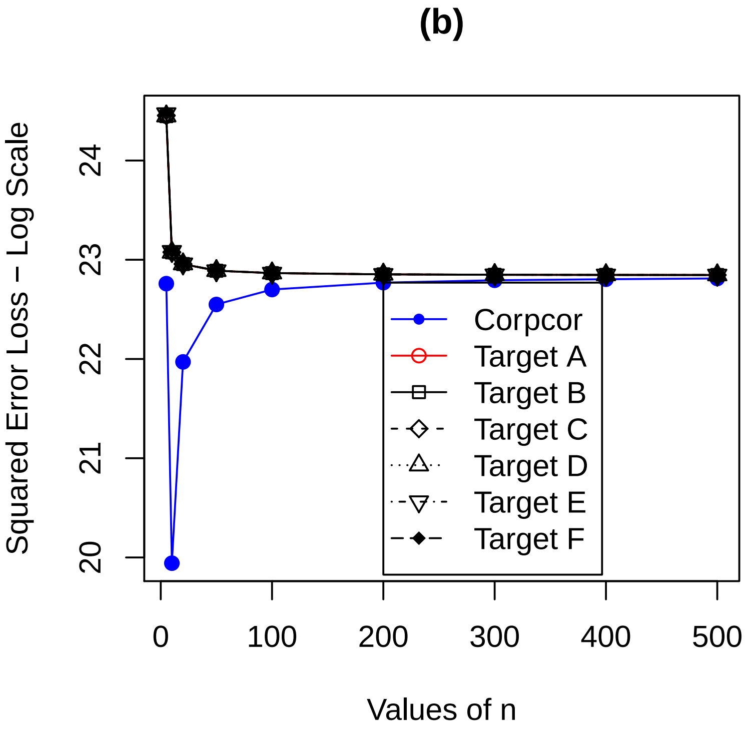


FIGURE 9. Log MSE calculated for each of the six target matrices laid out in Figure 2, plus the "Corpcor" estimator from the package "corpcor" [9]. The data is generated from the multivariate normal with outliers.

Finally, we arrive at Figure 9. Recall that in Figure 5 the corpcor estimate rose in MSE to reach the others. Now with the addition of more data points we see this trend again, with the exception of a spike downward in MSE for the Corpcor estimate between $n = 5$ and $n = 10$. Notice that this was precisely the spot where in Figure 8 we had a spike upwards in MSE for the Targets (minus Target E). This spike is very steep, and leads us to question what is going on under the hood in the "corpcor" package. It may be a similar phenomenon to what we observed in Figure 8, though in the opposite direction, but it is hard to tell given that we don't know the target matrix and the information from Figure 7 indicates that the shrinkage intensity would not have changed much from $n = 5$ to 10. Figure 9 also validates the conclusion that the Corpcor estimate is better to use in the case of outliers.

5. Conclusions

In this paper we have presented six different shrinkage estimators for the covariance matrix. We have discussed the main theorems leading to the properties that we want in an estimator of the covariance matrix - positive definiteness and good conditioning. The LW estimator assumes the least about the data, while the rest make a few different assumptions. Schäfer and Strimmer allow the user to make assumptions about the underlying structure of the data in order to choose a target matrix other than the identity. Chen et. al assume that the data is Gaussian for the RBLW and OAS estimators, and then assume that the data is distributed elliptically. Last, the STOA estimator combines the strenghts of shrinkage and tapering to handle the case where p goes to ∞ much faster than n. All of these estimators should behave far better than the sample covariance matrix in the small n large p case.

However, as we noticed in our simulations, the sample covariance matrix performs surprisingly well. Perhaps we only see this phenomenon at comparatively low ratios of n/p, which we were limited to by computation power. Ideally we would run dimensions of p higher than 500, and in the future it would be an interesting exercise to do so. We noticed that the "corpcor" package doesn't calculate the optimal shrinkage intensity the same way as the method in (58). Also, the targets described in Schäfer and Strimmer [11] all calculate a very similar estimate. It would be interesting to compare the other estimators with these, but unfortunately we were not able to acquire code for them and in the interest of time and analysis coding our own version was not feasible.

Of the estimators we were able to compare in the simulation study, Corpcor performs best for small n in both the case of normal data and data with outliers. Only when $n \approx p$ would it be more appropriate to use one of the other estimators. If we were to use one of these others, the best would be Targets A, B, or F. These have the best performance both in normal data and in outlier data besides Corpcor, regardless of sample size. Interestingly, Target A is the LW estimate and proves that it is a viable option. We hypothesize that the other estimators - RBLW estimator, the OAS estimator, the Chen estimator, and the Shrinkage-to-Tapering estimator - would perform better than all of the estimators considered in the simulation study, at least for the normal data. Since all of these estimators rely on assumptions about the distribution of the data, they might blow up in the case of outliers.

In the future, it would be interesting to implement the R code for the other estimators. These other estimators have been shown to perform better than the estimators that we ran simulations on, but we were not able to reproduce the results. Also, looking under the hood of the "corpcor" estimator to understand why it was so different from the "median" value in computation would be a good exercise. Along this vein optimizing our code to run more efficiently would significantly reduce computation time and allow for many more simulations to be run, especially in higher dimensions. To investigate the odd results from our simulations, I would want to try many more values of n and also many different combinations with different p, since a much larger difference between the two might contain a lot of information. I also want to try different methods of creating outliers, and using data distributed differently - perhaps with the multivariate t distribution, or something completely outside of the elliptical family to see how the estimators which rely on those assumptions handle the new data. I would also like to explore different algorithms for finding the optimal λ, which then would find a better $\hat{\Sigma}$. Also, there were plenty of techniques similar to shrinkage, like adaptive banding, which might be a better estimate in certain situations.

References

[1] Xiaohui Chen, Z. Jane Wang, and Martin McKeown. Shrinkage-to-Tapering Estimation of Large Covariance Matrices. *IEEE Transactions on Signal Processing*, 60:5640–5656, 2012.

[2] Yilun Chen, Ami Weisel, Yonina C. Eldar, and Alfred O. Hero. Shrinkage Algorithms for MMSE Covariance Estimation. *IEEE Transactions on Signal Processing*, 58:5016–5029, 2010.

[3] Yilun Chen, Ami Weisel, and Alfred O. Hero III. Robust Shrinkage Estimation of High-Dimensional Covariance Matrices. *IEEE Transactions on Signal Processing*, 59:4097–4107, 2011.

[4] Gabriel Frahm. Generalized Elliptical Distributions: Theory and Applications. *Ph.D. Dissertaion, Economic and Social Statistics Department*, University of Cologne, Germany, 2004.

[5] Johanna Hardin, Stephan Ramon Garcia, and David Golan. A method for generating realistic correlation matrices. *Annals of Applied Statistics*, 7:1249–1835, 2013.

[6] Olivier Ledoit and Michael Wolf. Honey, I Shrunk the Sample Covariance Matrix. *UPF Economics and Business Working Paper*, 691, 2003.

[7] Olivier Ledoit and Michael Wolf. Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of Empirical Finance*, 10, 2003.

[8] Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88:365–411, 2004.

[9] Rainer Opgen-Rhein and Korbinian Strimmer. Accurate Ranking of Differentially Expressed Genes by a Distribution-Free Shrinkage Approach. *Statistical Applications in Genetics and Molecular Biology*, 6, 2007.

[10] Adam J. Rothman, Elizaveta Levina, and Ji Zhu. A new approach to Cholesky-based covariance regularization in high dimensions. *Biometrika*, 3:539–550, 2010.

[11] Juliane Schäfer and Korbinian Strimmer. A Shrinkage Approach to Large-Scale Covariance Estimation and Implications for Functional Genomics. *Statistical Applications in Genetics and Molecular Biology*, 4, 2005.

[12] Harrison H. Zhou T. Tony Cai, Cun-Hui Zhang. Optimal Rates of Convergence for Covariance Matrix Estimation. *The Annals of Statistics*, 38:2118–2144, 2010.

[13] D.E. Tyler. A distribution-free M-estimator of multivariate scatter. *Annals of Statistics*, 15:234–251, 1987.

[14] Richard Weber. Statistics Course Notes from the University of Cambridge. page 14, 2007.