Challenges of Principal Component Analysis in High-Dimensional Settings when n<p

Nuwan Weeraratne

ncweera.sab@gmail.com

University of Waikato

Lyn Hunt

University of Waikato

Jason Kurz

University of Waikato

Research Article

Keywords: Dimensionality Reduction, Principal Component Analysis (PCA), High-Dimensional Covariance Estimation, Ledoit-Wolf Estimation

Posted Date: March 12th, 2024

DOI: https://doi.org/10.21203/rs.3.rs-4033858/v1

Additional Declarations: No competing interests reported.

Challenges of Principal Component Analysis in High-Dimensional Settings when n < p

Nuwan Weeraratne[1*], Lyn Hunt[2†] and Jason Kurz[1†]

[1*]Department of Mathematics, University of Waikato, New Zealand.
[2]Department of Software Engineering, University of Waikato, New Zealand.

*Corresponding author(s). E-mail(s): ncweera.sab@gmail.com;
Contributing authors: lah@waikato.ac.nz; jason.kurz@waikato.ac.nz;
†These authors contributed equally to this work.

Abstract

Principal Component Analysis (PCA) aims to reduce the dimensions of datasets by transforming them into uncorrelated Principal Components (PCs), retaining most of the data's variation with fewer components. However, standard PCA struggles in high-dimensional settings where there are more variables than observations due to limitations in covariance estimation. PCA relies on covariance matrices to measure variable relationships, using eigenvectors to determine data distribution directions and assessing eigenvalues' significance. This article examines the pros and cons of estimating high-dimensional covariance matrices and emphasizes the importance of well-conditioned covariance estimation for accurate finite sample PCA. Various methods are available for estimating population covariance, among which Ledoit-Wolf estimation is deemed optimal in scenarios where the number of observations is smaller than the number of variables. However, it tends to excessively shrink the sample covariates matrix, resulting in an underestimation of the true eigen spectrum and a dearth of sparsity. Therefore, there's a need for sparse and well-conditioned covariance matrix estimation to enhance PC estimation accuracy.

Keywords: Dimensionality Reduction, Principal Component Analysis (PCA), High-Dimensional Covariance Estimation, Ledoit-Wolf Estimation

1 Introduction

In various domains such as genomics, biometrics, medicine, e-commerce, network security, computer vision, ecology, and industry, the indexing of high-dimensional data has become increasingly vital in recent years. Within this context, we frequently encounter scenarios where decisions or outcomes hinge upon a multitude of factors, thereby complicating the decision-making process. This phenomenon is commonly referred to as the "curse of dimensionality." Addressing this challenge necessitates the application of either regularization or dimensionality reduction techniques.

Dimensionality reduction encompasses methods aimed at transforming the high-dimensional space of original data into a lower-dimensional space to derive a set of principal variables. Essentially, dimensionality reduction serves as a technique to eliminate less important information from a dataset. Moreover, dimensionality reduction can be subdivided into two main categories: feature selection and feature extraction. The primary distinction between feature selection and feature extraction lies in their respective approaches—while feature selection involves the process of selecting a subset from the original dataset, feature extraction entails generating a new set of features from the existing dataset.

Principal Component Analysis (PCA) emerges as a preferred statistical technique for reducing dimensionality and eliminating redundancy, particularly when all features are deemed equally relevant for the study. The fundamental principle of PCA revolves around the reduction of dataset dimensionality containing numerous interrelated variables while preserving a significant portion of its total variance. This is accomplished by transforming the original features into a new set of features known as principal components (PCs), which are uncorrelated. The initial PCs effectively capture most of the variation inherent in all original variables, thereby allowing the overall dataset variation to be succinctly described by only a few PCs.

Alternatively, PCA can be understood as a linear projection that minimizes the average cost of projection, represented by the mean squared distance between the data points and their respective projections. In this context, matrix decomposition plays a pivotal role in the PCA methodology. Matrix decomposition entails the transformation of a matrix into a product of matrices, facilitating the breakdown of the matrix into its constituent parts to enable specific operations. Specifically, within the realm of matrix decomposition methods, PCA relies on eigendecomposition—a technique that evaluates the relationship between variables using covariance matrices, determines the direction of data distribution through eigenvectors, and assesses the relative importance of these directions using eigenvalues.

Consequently, variance-covariance estimation serves as the cornerstone of PCA, forming the foundation upon which the technique operates. Through the assessment of the relationship between variables using covariance matrices, PCA enables the identification of the principal components that encapsulate the most significant

sources of variation within the dataset. PCA reduces the dimensionality of the original data set by maximizing the retained variance and minimizes the least square reconstruction error (see Fig. 1).

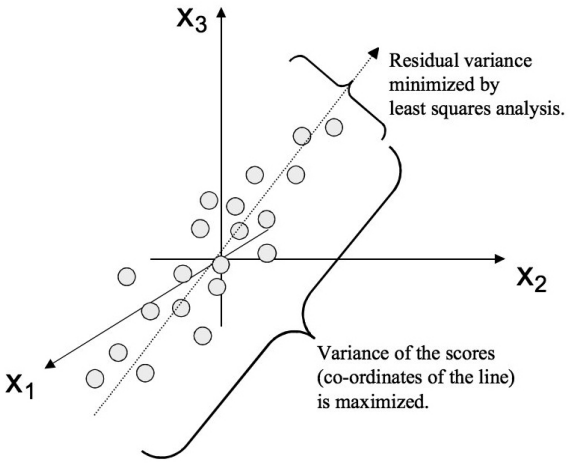


Fig. 1: The Principal Idea of PCA (taken from [1])

In other words, by leveraging eigendecomposition, PCA effectively transforms the original high-dimensional dataset into a lower-dimensional representation characterized by a reduced set of principal components that capture the essential features of the data. This attempt to reduce dimensionality can be described as "parsimonious summarization" of the data [2].

1.1 Notations

In this paper, we employ a set of notations and symbols to facilitate clarity and consistency in our presentation. These notations are introduced below for reference and ease of understanding.

- $\mathbf{X}$ is a $n \times p$ random data matrix with mean $\mathbf{0}$ and variance $\mathbf{\Sigma}$
- $\mathbf{\Sigma}$ is the true/population covariance matrix
- $\mathbf{\bar{S}}$ is the sample covariance matrix or the maximum likelihood sample covariance matrix and $\hat{\mathbf{\Sigma}}$ is the estimator of population covariance matrix
- $\mathbf{\bar{S}}_{\text{ii}}$ is the variance of i^{th} variable
- $\mathbf{\bar{S}}_{\text{ij}}$ is the covariance of i^{th} and j^{th} variables, $i \neq j$ and $\mathbf{\bar{S}}_{\text{ij}} = \mathbf{\bar{S}}_{\text{ji}}$

- p is the number of variables/dimensions
- n is the number of observations/ sample size
- λ_{i} is the i^{th} eigenvalue of the sample covariance matrix
- e_{i} is the i^{th} eigenvector of the sample covariance matrix
- $\mathbf{U}$ is an $n \times n$ orthogonal matrix containing the right singular vectors
- $\mathbf{V}$ is a $p \times p$ orthogonal matrix containing the left singular vectors
- $\mathbf{M}$ is an $n \times p$ diagonal matrix containing the singular values
- $\mathbf{0}$ is a $p \times 1$ zero vector
- $\mathbf{1}$ is a $p \times p$ unit matrix
- $\mathbf{I}$ is a $p \times p$ identity matrix
- r is the number of principal components
- $\hat{\mu}$ is the sample mean vector
- $\boldsymbol{\Theta}$ is the precision matrix ($\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1}$)
- $\mathbb{P}$ is a matrix of eigenvectors of the sample covariance matrix
- $|\boldsymbol{\Sigma}|$ is the determinant of covariance matrix
- $tr(\boldsymbol{\Sigma})$ is the sum of elements on the main diagonal of $\boldsymbol{\Sigma}$ (trace of $\boldsymbol{\Sigma}$)
- $||\boldsymbol{\Sigma}||_1$ is the maximum absolute column sum of $\boldsymbol{\Sigma}$
- $||\boldsymbol{X}||_2$ is the l_2 norm of N-dimensional vector $\mathbf{X}$ and it measures the shortest distance from the origin. It is defined as the root of the sum of the squares of the components of the vector
- $||\boldsymbol{X}||_{\mathrm{F}}$ is the frobenius norm of $n \times p$ matrix $\mathbf{X}$ and it measures the the square root of the sum of the squares of all the matrix entries

1.2 Sample Principal Components

Let the random matrix $\mathbf{X^T} = [\mathrm{X_1}, \mathrm{X_2}, ..., \mathrm{X_p}]$ have the covariance matrix $\mathbf{\Sigma}$ with eigenvalues $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_\mathrm{p} \geq 0$. Consider data on p variables for n observations. The first sample PC is then the linear combinations of the variables $\mathrm{X_1}, \mathrm{X_2}, ..., \mathrm{X_p}$,

$$PC_1 = Z_1 = \boldsymbol{e}_1{}^{\mathrm{T}}\boldsymbol{X} = e_{11}X_1 + e_{12}X_2 + ... + e_{1\mathrm{p}}X_\mathrm{p} \tag{1}$$

and it's variance is maximum for the individuals subject to $\mathbf{e}_1{}^{\mathrm{T}}\mathbf{e}_1 = 1$.

$$e_{11}{}^2 + e_{12}{}^2 + ... + e_{1\mathrm{p}}{}^2 = 1 \tag{2}$$

Thus the $\mathrm{Var(Z_1)}$ is as large as possible satisfying the above constraint. The constraints introduced here is to prevent increasing $\mathrm{Var(Z_1)}$ simply by increasing one of the coefficients. The second PC,

$$PC_2 = Z_2 = \boldsymbol{e}_2{}^{\mathrm{T}}\boldsymbol{X} = e_{21}X_1 + e_{22}X_2 + ... + e_{2\mathrm{p}}X_\mathrm{p} \tag{3}$$

is such that $\mathrm{Var(Z_2)}$ is as large as possible subject to the constraint, $\mathbf{e}_2{}^{\mathrm{T}}\mathbf{e}_2 = 1$.

$$e_{21}{}^2 + e_{22}{}^2 + ... + e_{2\mathrm{p}}{}^2 = 1 \tag{4}$$

with additional condition that $\mathrm{Z_2}$ is uncorrelated with $\mathrm{Z_1}$. Similarly the third PC,

$$PC_3 = Z_3 = \boldsymbol{e}_3{}^{\mathrm{T}}\boldsymbol{X} = e_{31}X_1 + e_{32}X_2 + ... + e_{3\mathrm{p}}X_\mathrm{p} \tag{5}$$

is such that $\mathrm{Var(Z_3)}$ is as large as possible subject to the constraint, $\mathbf{e}_3{}^{\mathrm{T}}\mathbf{e}_3 = 1$.

$$e_{31}{}^2 + e_{32}{}^2 + ... + e_{3\mathrm{p}}{}^2 = 1 \tag{6}$$

and also that $\mathrm{Z_3}$ is uncorrelated with both $\mathrm{Z_1}$ and $\mathrm{Z_2}$. Other PCs can be defined in the similar way and this can be continued until p PCs are defined.

Maximizing a function subject to constraints is usually achieved by using a mathematical method called the Lagrange Multiplier. Under the optimum condition, the variances of the PCs are the eigenvalues of the sample variance-covariance matrix [3], $\mathbf{S}$, of the p original variables and $\mathbf{S}$, which is of the form,

$$\boldsymbol{S} = \begin{bmatrix} S_{11}{}^2 & S_{12}{}^2 & ... & S_{1\mathrm{p}}{}^2 \\ S_{21}{}^2 & S_{22}{}^2 & ... & S_{2\mathrm{p}}{}^2 \\ . & . & ... & . \\ . & . & ... & . \\ . & . & ... & . \\ S_{\mathrm{p}1}{}^2 & S_{\mathrm{p}2}{}^2 & ... & S_{\mathrm{pp}}{}^2 \end{bmatrix}.$$

Since $\mathbf{S}$ is of size $p \times p$, there are p eigenvalue-eigenvector pairs $(\hat{\lambda}_1,\hat{\boldsymbol{e}}_1)$, $(\hat{\lambda}_2,\hat{\boldsymbol{e}}_2)$, ..., $(\hat{\lambda}_p,\hat{\boldsymbol{e}}_p)$ correspond to p PCs. If the eigenvalues are ordered as,

$$\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_\mathrm{p} \geq 0. \tag{7}$$

Then λ_1 correspond to the first PC ($\mathrm{Var}(\mathrm{Z}_1) = \lambda_1$), and λ_2 correspond to the second PC ($\mathrm{Var}(\mathrm{Z}_2) = \lambda_2$) and so on. Since the $\mathrm{Var}(\mathrm{Z}_1)$ is the largest followed by $\mathrm{Var}(\mathrm{Z}_2)$ and so on. The e_{11}, e_{12}, ..., $e_{1\mathrm{p}}$ are the elements of the corresponding eigenvector of λ_1 and thus the first PC is determined. Similarly other PCs can also be determined.

As a result of the relationship between eigenvalues and variance of principal components, the total sample variance can be represented by

$$Total\ Sample\ Variance = \sum_{i=1}^{p} \boldsymbol{S}_{\mathrm{i}}{}^2 = \sum_{i=1}^{p} \lambda_{\mathrm{i}} = \hat{\lambda}_1 + \hat{\lambda}_2 + ... + \hat{\lambda}_{\mathrm{p}}. \tag{8}$$

This shows that the principal components capture all of the variability in the original dataset. When data is collected under a large number of variables, it is often the case that different variables are measured in different units. Because of the unit of the measurement, sometimes the variance becomes large. In order to avoid certain variables having an undue influences on the PCs, often standardized variables are used for the analysis. When standardized variables are used,

$$Total\ Sample\ Variance = \sum_{i=1}^{p} \boldsymbol{S}_{\mathrm{i}}{}^2 = p \tag{9}$$

and thus proportion of variability explained by the i^{th} PC becomes λ_{i}/p. Also note that when variables are standardized, the covariance matrix becomes the correlation matrix.

There are various statistical tests to determine the optimum number of PCs [4] namely Kaiser-Guttman Criterion [5] and [6], Cattell's Scree Test [7] and Percent of Cumulative Variance.

1. According to the scree test, the components to the left of the "elbow" point of the graph should be retained as optimal number of PCs. (see Fig. 2)

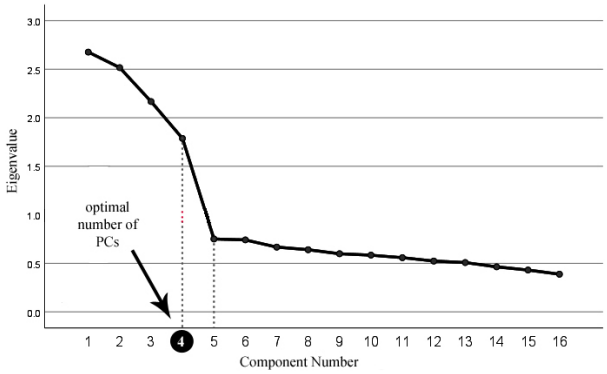


Fig. 2: Scree plot method

2. The Kayser-Guttman number of components to retain ($n_{\text{K-G}}$) is computed by $n_{\text{K-G}} = \text{Count}\ (\lambda_{\text{i}} > 1)$, where, λ_{i} is the i^{th} eigenvector.
3. Percent of cumulative variance method retain the number of components, that explain 70% or 80% (note: it's a thumb rule) of the total variance.

However, most of them commonly used non-statistical strategies to determine the optimum number of PCs [8].

1.3 Computations of PCA

The PCs are calculated by two main approaches: Using Singular Value Decomposition (SVD) or solving the covariance matrix.

1. **SVD Method**

   Let's denote the sample covariance matrix as $\mathbf{S}$, which is typically computed from a dataset $\mathbf{X}$ of size $n \times p$, where n is the number of samples and p is the number of features.The Singular Value Decomposition (SVD) of $\mathbf{S}$ can be expressed as:

$$\boldsymbol{S}_{p\times p} = \boldsymbol{U}_{n\times n}\boldsymbol{M}_{n\times p}{\boldsymbol{V}^{\mathrm{T}}}_{p\times p}$$

$$\begin{bmatrix} s_{11} & s_{12} & \dots & s_{1\text{p}} \\ s_{21} & s_{22} & \dots & s_{2\text{p}} \\ . & . & \dots & . \\ . & . & \dots & . \\ . & . & \dots & . \\ s_{\text{p1}} & s_{\text{p2}} & \dots & s_{\text{pp}} \end{bmatrix} = \begin{bmatrix} u_{11} & u_{21} & \dots & u_{\text{n1}} \\ u_{12} & u_{22} & \dots & u_{\text{n2}} \\ . & . & \dots & . \\ . & . & \dots & . \\ . & . & \dots & . \\ u_{1\text{n}} & u_{2\text{n}} & \dots & u_{\text{nn}} \end{bmatrix} \begin{bmatrix} m_1 & 0 & \dots & 0 \\ 0 & m_2 & \dots & 0 \\ . & . & \dots & . \\ . & . & \dots & . \\ . & . & \dots & . \\ 0 & 0 & \dots & 0 \end{bmatrix} \begin{bmatrix} v_{11} & v_{12} & \dots & v_{1\text{p}} \\ v_{21} & v_{22} & \dots & v_{2\text{p}} \\ . & . & \dots & . \\ . & . & \dots & . \\ . & . & \dots & . \\ v_{\text{p1}} & v_{\text{p2}} & \dots & v_{\text{pp}} \end{bmatrix}$$

   The singular values in $\mathbf{M}$ are typically arranged in descending order along the diagonal. The singular vectors in $\mathbf{U}$ and $\mathbf{V}$ are orthonormal, meaning that $\mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{I}$ and $\mathbf{V}^{\mathrm{T}}\mathbf{V} = \mathbf{I}$ where $\mathbf{I}$ is the identity matrix.

2. **Covariance Matrix Method**

   Perhaps the oldest and most widely used covariance estimation approach is Maximum Likelihood Estimation (MLE) of covariance. In 1922 R. A. Fisher introduced the method of maximum likelihood. Let x_1, x_2, x_3,..., x_n be a set of observations of random variables X_1, X_2, X_3,..., X_n with joint probability density function f$x(x_1, x_2, x_3, \dots, x_n, \theta)$, when considered as a function of θ, is called the likelihood function of θ for the set of observations x_1, x_2, x_3,..., x_n. It is denoted by L(θ; x_1,x_2,x_3,...,x_n). If $\hat{\theta}$ is the value of θ which maximizes L(θ; x_1, x_2, x_3,..., x_n), then it is the MLE of the sample. The steps of the maximum likelihood method may be outlined as follows.

For an example, assume that the variable X $\sim$ N(μ,$\sigma_x{}^2$).

Step I: Form the likelihood function.

$$L = f(X_1,X_2,X_3,...,X_n;\ \mu,\sigma_x{}^2)$$
$$L = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_x{}^2}}.exp-\tfrac{1}{2}(\tfrac{X_i-\mu}{\sigma_x})^2$$
$$L = \tfrac{1}{\sqrt{2\pi\sigma_x{}^2}}{}^n.exp-\tfrac{1}{2}\sum_{i=1}^{n}(\frac{X_i-\mu}{\sigma_x})^2$$

To simplify the partial differentiation we express the likelihood function in the logarithms of the variables.

$$log_e\ L = -nlog_e\sigma_x - nlog_e\sqrt{2\pi} - \tfrac{1}{2}\sigma_x{}^{-2}\Sigma(X_i-\mu)^2$$

Step II: Take the partial derivatives of the likelihood function with respect to the parameters which we want to estimate and equate to zero.

$$\tfrac{\partial logL}{\partial\mu} = -\tfrac{1}{2}\sigma_x{}^{-2}\Sigma 2(X_i-\mu) = 0$$

$$\tfrac{\partial logL}{\partial\sigma_x} = -n\tfrac{1}{\sigma_x} + \sigma_x{}^{-3}\Sigma(X_i-\mu)^2 = 0$$

Step III: Solve the equations of the partial derivatives for the unknown parameters, to obtain their maximum likelihood estimates.

a). ML estimator of population mean:

$$\Sigma X_i - n\mu = 0$$
$$\hat{\mu} = \tfrac{\Sigma X_i}{n} = sample\ arithmetic\ mean$$

b). ML estimator of population covariance:

$$-\tfrac{n}{\hat{\sigma_x}} + \tfrac{\Sigma(X_i-\mu)^2}{\hat{\sigma_x}{}^3} = 0$$
$$\tfrac{1}{\hat{\sigma_x}}[-n + \tfrac{\Sigma(X_i-\mu)^2}{\hat{\sigma_x}{}^2}] = 0$$
$$\tfrac{\Sigma(X_i-\mu)^2}{\hat{\sigma_x}{}^2} = n$$
$$\hat{\Sigma}_{ML} = \hat{\sigma_x}{}^2 = \tfrac{\Sigma(X_i-\mu)^2}{n} = sample\ variance$$

In this method, PCs are given by the eigenvectors of the covariance matrix of the data matrix **X**. (see Fig. 3)

In summary, PCA reduces dimensionality in high-dimensional datasets by transforming features into uncorrelated principal components, preserving essential variation and providing insights into data structure.

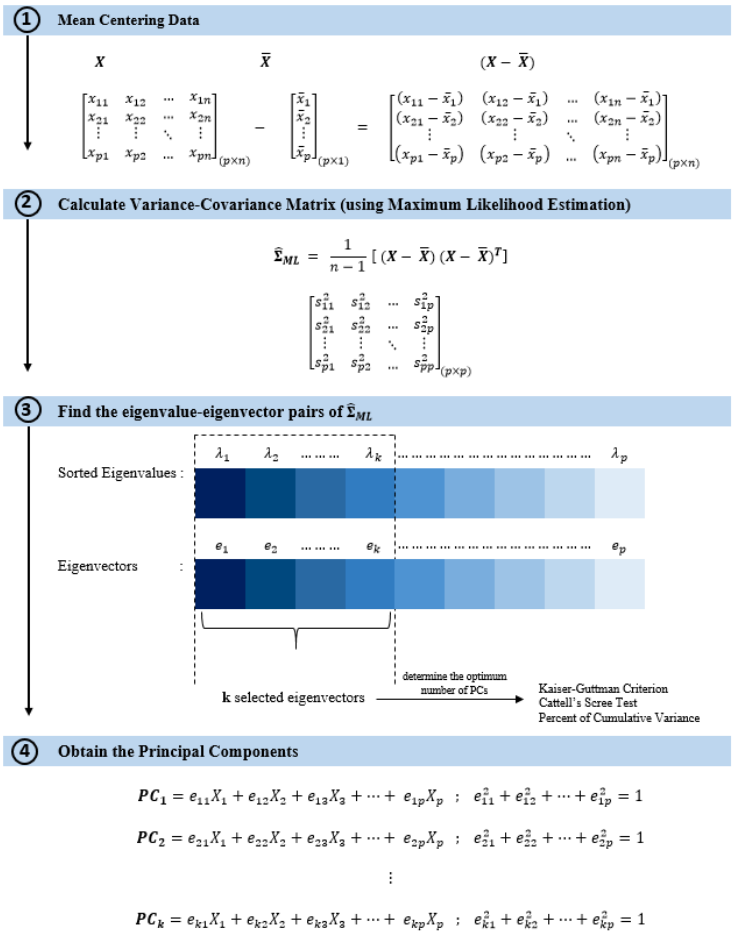


Fig. 3: Geometric interpretation of diagonalizing the covariance matrix in 2D

2 Principal Component Analysis when n < p

The Maximum Likelihood Estimation (MLE) method, as introduced by Anderson in 1970, stands out as a widely utilized and prominent technique for covariance estimation in practical applications. While the usual maximum likelihood estimator of covariance **S** is asymptotically unbiased, it tends to provide inadequate estimates of PCs and poorly conditioned estimates of the covariance matrix $\mathbf{\Sigma}$ in high-dimensional settings, particularly when n (the number of observations) is less than p (the number of variables).

When PCA is applied to high-dimensional data, several challenging issues arise, with three major concerns standing out prominently.

1. **Consistency of MSE (S) breaks down as p/n $\to \infty$**

The depicted Fig.4 demonstrates that as the ratio of variables to observations increases, the Mean Squared Error (MSE) of **S** also experiences exponential growth. This observation indicates that as the number of variables in the dataset surpasses the number of observations, the MSE of **S** exhibits exponential escalation. Consequently, it becomes evident that the sample covariance matrix is not a reliable estimator of the population covariance matrix in high-dimensional settings characterized by "large p small n" scenarios, particularly when utilized as a dimensionality reduction technique.

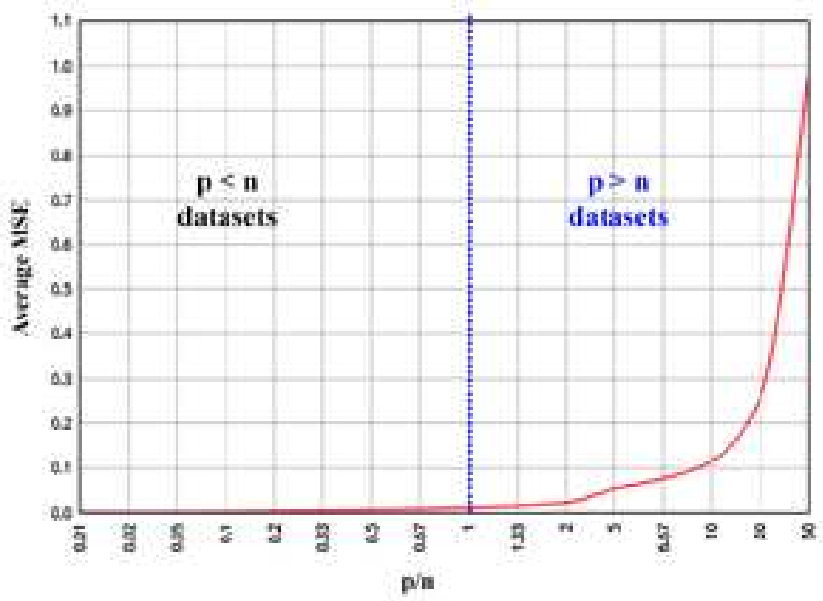


Fig. 4: GEffect of p/n with respect to MSE of **S**

Let $\mathbf{X} \in \mathbb{R}^{n \times p}$ represent n observations with p features where we primarily focus on the situation such that $p/n > \epsilon$ for some $\epsilon > 0$ (or sufficiently large). Moreover, Stuart Gemen [9] showed that, if $\mathbf{X} \sim \mathrm{N}(\mathbf{0}, \sigma^2)$ and $p/n \to \theta$ as $n \to \infty$ for some $0 < \theta < \infty$, then the maximum sample eigenvalue λ_1 satisfies $(1+\theta^{1/2})^2\sigma^2$. That means if $\mathbf{X} \sim \mathrm{N}(\mathbf{0}, \mathbf{I})$, then the largest eigenvalue of the sample tends to $(1+\theta^{1/2})^2$. This tells us, if p/n, does not converge to zero, that the first sample eigenvalue is not a consistent estimator of the maximum eigenvalue ${\sigma_{11}}^2$ of $\mathbf{\Sigma}$. Moreover, if $n=2p$, $n=p$, $n=0.75p$,$n=0.5p$ and $n=0.25p$, then the largest sample eigenvalue λ_1 tends to 2.91, 4, 4.64, 9 or 25 respectively whereas ${\sigma_{11}}^2$ is 1. This implies, when $n < p$, the problem of consistency of covariance estimation is worse.

2. **Eigenvalues of $\mathbf{S}$ are over-dispersed because *(p-n+1)* eigenvalues are exactly equal to zero.**

Numerous empirical studies, including those [10], [11], [12], [13], [14], [15], and [16], have highlighted the inadequacy of the sample covariance matrix $\mathbf{S}$ as an estimator of the true population covariance when the dataset's dimensions exceed the sample's observations. Furthermore, these studies emphasized that the eigenvalues of the sample covariance matrix $\mathbf{S}$ exhibit over-dispersion, particularly due to *(p-n+1)* eigenvalues being exactly equal to zero, as depicted in Fig. 5. Thus, it is crucial to consider alternative methods for covariance estimation in high-dimensional datasets where the number of dimensions surpasses the number of observations to ensure accurate and reliable results in statistical analyses.

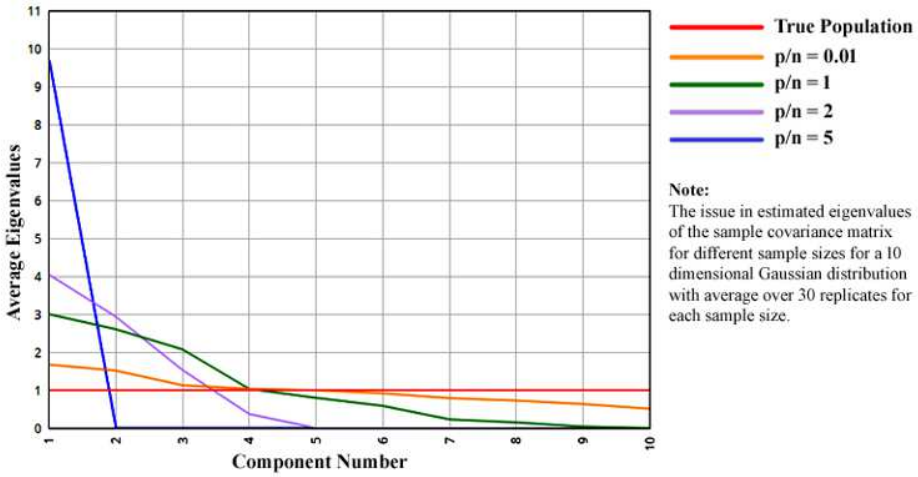
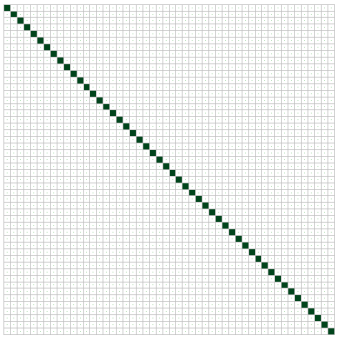


Fig. 5: Over-dispersion of sample eigenvalues $\mathbf{S}$

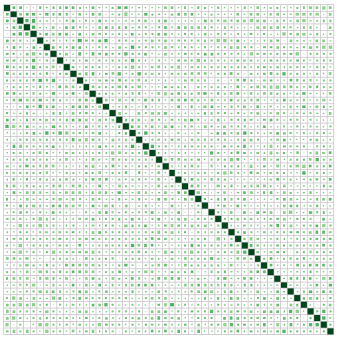
In Fig.5, it is evident that when $n \leq p$, the eigenvalues are often significantly misestimated. Specifically, the largest eigenvalue tends to be excessively large or over-dispersed, while the smallest eigenvalue is typically too small or under-dispersed. Consequently, if the covariance matrix of $\mathbf{X}$ possesses a rank (r) less than p, then the total variation of $\mathbf{X}$ can be entirely explained by the first r principal components. This stems from the fact that if the covariance matrix has a rank r, then the last $(p\text{-}r)$ eigenvalues are uniformly zero, underscoring the importance of understanding the relationship between eigenvalues and rank in accurately characterizing the variation within the dataset.

3. $\mathbf{S}$ is very noisy/ non-sparse matrix (matrix with mostly non-zero values) and therefore biased for true/population covariance matrix $\mathbf{\Sigma}$.

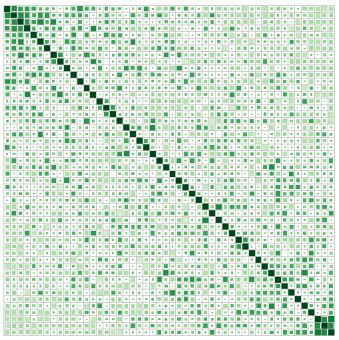
A covariance matrix is deemed sparse when the majority of its entries are zero, serving as a means to mitigate the curse of dimensionality. However, as depicted in Fig. 6, the maximum likelihood estimation of the covariance matrix often yields a noisy and biased approximation of the true population covariance. Moreover, the eigenvectors of $\mathbf{S}$, a crucial output of PCA, can significantly deviate from those of $\mathbf{\Sigma}$, as identified by Johnstone and Lu [17] and Ledoit and Pèchè [18]. Consequently, the limitations of the sample covariance matrix $\mathbf{S}$ underscore the necessity for a well-conditioned estimator of the covariance matrix. Such an estimator can ease estimation errors (projection cost) by enhancing the accuracy of eigenvalue estimation, while also confirming the positive definiteness and invertibility of the covariance matrix.



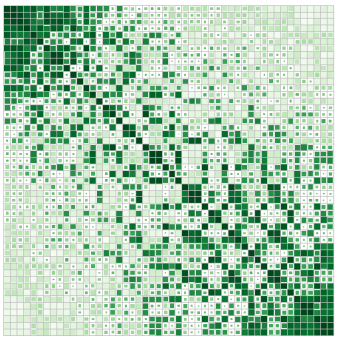
((a)) $\mathbf{\Sigma}$ (True Covariance)



((b)) $\mathbf{S}$ (where; $p = 50$ and $n = 250$)



((c)) $\mathbf{S}$ (where; $p = 50$ and $n = 25$)



((d)) $\mathbf{S}$ (where; $p = 50$ and $n = 5$)

Fig. 6: Graph of True & Sample Covariance Matrices

3 High Dimensional Covariance Estimation

Due to the limitations of maximum likelihood (empirical) sample covariance estimation when $n < p$, researchers have been motivated to explore various regularized covariance estimators in diverse contexts. Here, we provide a selective review of high-dimensional covariance and precision matrix estimations, focusing on their effectiveness in addressing the challenges associated with sample covariance in high-dimensional datasets. For that purpose, we illustrate the sample data, where $\mathbf{X}_{\mathrm{i}} \sim N_{10}(\mathbf{0}, \mathbf{I})$; where i = 1,2,3,...,500 with average over 500 replicates for sample size $n = 3$. (For instance, let $\mathbf{X}$ denote a multivariate normal distribution with p variables ($p = 10$), where the mean vector is zero and the variance-covariance matrix is the identity matrix.)

3.1 Stein's Estimation - 1975

Since 1950s, Charles Stein observed that the the eigenvalues of the sample variance-covariance matrix $\mathbf{S}$ are much more spread out than the eigenvalues of the population covariance matrix $\mathbf{\Sigma}$ [19]. In 1975, Charles Stein proposed a new covariance estimation method called "Stein Estimator" [20] under Stein's loss function. Let $\mathbf{S} = RLR^{\mathrm{T}}$ be the eigen-decomposition of the maximum likelihood covariance estimator $\mathbf{S}$, where L is diagonal matrix of eigenvalues, and R is matrix of eigen vectors. The Stein's covariance matrix estimator is,

$$\hat{\mathbf{\Sigma}}_{\mathrm{ST}} = R\varphi(L)R^{\mathrm{T}}$$

where
$L = diag(l_1,l_2,...,l_p)$, $l_1 \geq l_2 \geq ... \geq l_{\mathrm{p}}$ are the eigenvalues of $\mathbf{S}$
R is an orthogonal matrix such that $\mathbf{S} = RLR^{\mathrm{T}}$,
and $\varphi(L) = diag(\varphi_1(L), \varphi_2(L), ..., \varphi_{\mathrm{p}}(L))$ with non-negative elements.

The eigenvalues are shrinking towards a central value, while the eigenvectors are kept as they are in this approach. This estimator is known as a rotation equivariant covariance estimator because the eigenvectors are not changed or regularized. For the purpose of selecting the value of φ, we can use the entropy loss function or the Frobeneous loss function shown below.

Entropy loss function:

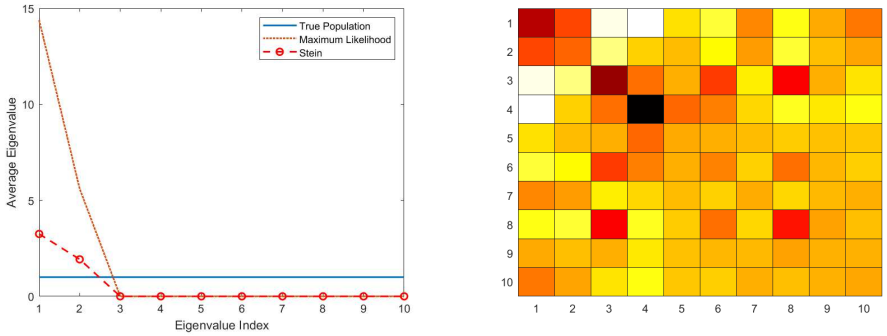
$$tr(\hat{\mathbf{\Sigma}}\mathbf{\Sigma}^{-1}) - log(\hat{\mathbf{\Sigma}}\mathbf{\Sigma}^{-1}) - p$$

Frobeneous loss function:

$$tr(\hat{\mathbf{\Sigma}}\mathbf{\Sigma}^{-1} - \boldsymbol{I})^2$$

Stein's Covariance Estimation method helps estimate covariance matrices, offering advantages like risk reduction and minimax properties (the minimax property refers to the method's ability to minimize the maximum possible risk or error across all possible covariance matrices). However, it has drawbacks worth noting:

1. It only sometimes ensures positive definiteness, which is crucial for meaningful interpretation. Positive definiteness breaks down in Charles Stein's covariance estimation when the sample size is small relative to the number of variables, when the data exhibit high variability or outliers, or when there are deviations from the normality assumption in the underlying data distribution.
2. It relies on the assumption of normality in the data distribution, which may not hold in real-world datasets, potentially leading to biased estimates.
3. Its performance is limited by sample size conditions, being most effective when the sample size is greater than or equal to the number of variables.



((a)) Eigenvalue Analysis ((b)) Sample Covariance

Fig. 7: Eigenvalue distribution and Noise of $\hat{\Sigma}_{\text{Stein}}$ (where $\mathbf{X}_{\text{i}} \sim N_{10}(\mathbf{0}, \mathbf{I})$ with average over 500 replicates for sample size $n = 3$.)

According to Figure 7, the primary limitations of Stein's estimator include its lack of sparsity, requirement for positive definiteness, and unsuitability for high-dimensional data (typically effective when $n \geq p$ and p is not excessively large). While Stein's estimator addresses over-dispersion in eigenvalues, it does not yield a sparse covariance matrix, making it unsuitable for high-dimensional settings.

Additionally, it may need to handle extreme differences in population eigenvalues better, leading to biased estimates. Furthermore, it lacks robustness to outliers or extreme values in the data, which can distort covariance estimates, limiting its applicability in datasets with high variability or outliers. In conclusion, while Stein's method offers benefits, its limitations include a lack of positive definiteness, reliance on normality, sample size constraints, bias with extreme eigenvalues, and reduced robustness to outliers, necessitating careful evaluation for suitability in analysis.

3.2 Ledoit-Wolf Estimation - 2004

Sample covariance matrix ($\mathbf{S}$) eigenvalues can also be truncated using the empirical Bayesian estimator, which is the linear combination of the estimator $\mathbf{S}$ and the estimator $\mathbf{I}$. This is an appropriate approach as the estimator $\mathbf{S}$ is unbiased but is highly unstable in high dimensional data, while the estimator $\alpha\mathbf{I}$ has very low variability with potentially high bias.

$$\hat{\mathbf{\Sigma}} = \frac{np - 2n - 2}{n^2 p}\alpha \boldsymbol{I} + \frac{n}{n+1}\boldsymbol{S}$$

In 2004, Ledoit and Wolf proposed a more general form of estimator would be a well-known stenian-type shrinkage (minimizing a Mean Squared Error (MSE) criterion) estimator for the $\hat{\mathbf{\Sigma}}$ [21]. That linear shrinkage estimator as given by,

$$\hat{\mathbf{\Sigma}}_{\mathrm{LW}} = \hat{\rho}[\frac{1}{n}\sum_{i=1}^{n}(X_{\mathrm{i}} - \bar{X})(X_{\mathrm{i}} - \bar{X})^{'}] + (1 - \hat{\rho})((tr(\boldsymbol{S})/p) \times \boldsymbol{I})$$

where;

$$\hat{\rho} = \frac{(1 - \frac{2}{p})tr(\boldsymbol{S}^2) + tr(\boldsymbol{S})^2}{(n + 1 - \frac{2}{p})tr(\boldsymbol{S}^2) + (1 - \frac{n}{p})tr(\boldsymbol{S})^2}; \hat{\rho} \in [0, 1)$$

$$\boldsymbol{S} = \frac{1}{n\ \text{-}\ 1}\sum_{i=1}^{n}(x_{\mathrm{i}} - \bar{x})(x_{\mathrm{i}} - \bar{x})^{\mathrm{T}}$$

According to the results of this Ledoit-Wolf estimator, the modified eigenvalues $\hat{\lambda}_{\mathrm{i}}$ is denoted as,

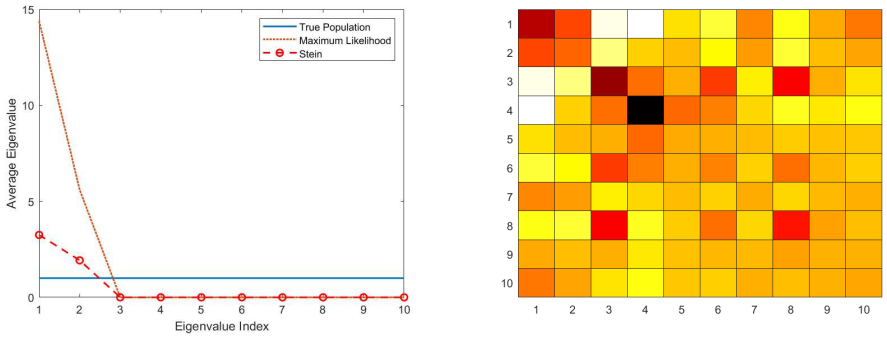
$$\hat{\lambda}_{\mathrm{i}} = (1 - \rho)l_{\mathrm{i}} + \rho c$$

with $c = tr(\mathbf{S})\ /\ p$ and l_{i} i^{th} eigenvalue of $\mathbf{S}$.

The ledoit-Wolf covariance estimation method presents several drawbacks:

1. Lack of Sparsity: The method produces dense covariance matrices even in scenarios where sparsity is expected, increasing computational complexity and potentially obscuring relationships between variables.
2. Uncertainty in Shrinkage Intensity Parameter: Ledoit-Wolf relies on an estimator for the shrinkage intensity parameter, which introduces uncertainty and can result in suboptimal performance due to inaccuracies in its estimation.
3. Lack of Adaptability to Data Structure Changes: Ledoit-Wolf preserves eigenvectors of the sample covariance matrix even when they do not accurately represent the data structure. This leads to inconsistent estimates, especially when the underlying covariance structure undergoes significant changes.

4. Biased Linear Shrinkage: The method applies linear shrinkage to eigenvalues, which may lead to bias, particularly for eigenvalues far from the center, resulting in inaccuracies in high-dimensional settings.
5. Suboptimal Performance in High-Dimensional Settings: Ledoit-Wolf may yield biased or unstable estimates in scenarios where the number of variables is comparable to or greater than the number of observations.
6. Inadequate Handling of Outliers: The method may need to handle outliers or anomalies in the data robustly, leading to biased or unreliable results, particularly in datasets with extreme observations.
7. Computational Inefficiency: Ledoit-Wolf estimation can be computationally demanding, limiting its scalability in practical applications with large-scale datasets or constrained computational resources.
8. Sensitivity to Model Assumptions: The method's performance may vary depending on the underlying distribution of the data and the validity of assumptions regarding the covariance structure, potentially leading to biased or unreliable covariance estimates.



((a)) Eigenvalue Analysis ((b)) Sample Covariance

Fig. 8: Eigenvalue distribution and Noise of $\hat{\Sigma}_{\text{LW}}$ (where $\mathbf{X}_{\text{i}} \sim N_{10}(\mathbf{0}, \mathbf{I})$ with average over 500 replicates for sample size $n = 3$.)

However, Figure 8 shows that when $n < p$, reduces the over-dispersion in eigenvalues with non-zero values (but $\lambda_3, ..., \lambda_{10}$ give the same eigenvalue). In other words, the Ledoit-Wolf Estimator is not sparse, $\hat{\rho}$ depends on the unknown S, and uniform shrinkage. ([22] and [23])

3.3 Graphical Lasso (GLasso) Estimation - 2008

The GLasso estimator is a sparse penalized maximum likelihood estimation for a precision matrix (the inverse of the covariance matrix), of any given member of a family of generalised multivariate normal distributions. In the case of small-samples, sparse inverse covariances are generally more effective than shrunk covariances. Therefore, In

2008, Friedman [24] proposed a covariance estimation method to other estimator and distribution types as an extension for the Dempster covariance selection problem for a multivariate Gaussian distribution when observations are constrained. That covariance estimator as given by,

$$\hat{\boldsymbol{\theta}} = argmin_{\boldsymbol{\theta} \geq 0}[tr(\boldsymbol{S\theta}) - log\det(\boldsymbol{\theta}) + \lambda\Sigma_{j \neq k} \mid \boldsymbol{\theta}_{jk} \mid]$$

Where $\boldsymbol{\theta} = \boldsymbol{\Sigma}^{-1}$ and $\hat{\boldsymbol{\theta}}$ is the precision matrix to be estimated, $\mathbf{S}$ is the sample covariance matrix, and λ is the penalizing parameter.

GLasso estimation, leveraging an l_1 penalty for promoting sparsity in the precision matrix, confronts significant challenges in high-dimensional covariance estimation. Despite its benefits, including ease of interpretation and sparsity emphasis, GLasso faces computational inefficiencies in datasets with many variables relative to the sample size. The optimization problem becomes computationally demanding, consuming substantial resources and time to converge to a solution, limiting its scalability and practicality for real-time or large-scale datasets where computational efficiency is paramount. Furthermore, GLasso's sensitivity to the choice of regularization parameter complicates its application, as selecting an appropriate parameter to control sparsity becomes non-trivial. Suboptimal choices may yield biased or unreliable estimates, particularly in datasets with complex covariance structures where GLasso struggles to accurately model the relationships between variables.

GLasso's shortcomings extend to handling outliers and adherence to model assumptions and data distributional properties. Outliers can unduly influence the estimated covariance matrix, leading to biased or unreliable results and constraining GLasso's applicability in datasets prone to data quality issues or extreme observations. The method's performance also hinges on strict adherence to underlying data distributional properties and model assumptions, which may only sometimes hold in practice. Deviations from these assumptions can further compromise the robustness and generalizability of GLasso estimation, necessitating careful evaluation of its suitability for high-dimensional or complex data settings. In summary, while GLasso estimation offers certain advantages, its drawbacks underscore the importance of thoughtful consideration and evaluation regarding its applicability to specific datasets and analytical objectives in high-dimensional covariance estimation tasks.

3.4 Covariance Estimation via Hard Thresholding - 2008

In 2008, Bickel and Levina developed a method of sparse estimation for the application of thresholding to off-diagonal components of a sample covariance matrix of p variables estimated from n observations [25]. They demonstrate the robustness of the thresholded estimate in the operator norm, provided that the actual covariance matrix is sufficiently sparse, the parameters are Gaussian (or sub-Gaussian), and log p/n is set to zero, and explicit rates are obtained.

The entry of sample covariance matrix $\mathbf{S}_{\mathrm{i,j}} = 0$ if $\mid \boldsymbol{S}_{\mathrm{i,j}} \mid \leq \tau$ where τ is a thresholding value. This is an user-defined threshold value. If it is a vector of regularization values, it automatically selects one that minimizes cross validation risk. Thresholding factor is given by $\tau = \sqrt{log(\frac{p}{n})}$, where $\mathbf{X}$ is the $n \times p$ data matrix.

Bickel and Levina's Covariance Estimation via the Hard Thresholding method presents several drawbacks in high-dimensional covariance estimation. One significant limitation lies in its sensitivity to the selection of the threshold parameter. The method's performance heavily relies on this parameter, making it challenging to determine an appropriate value, especially in scenarios with high-dimensional data exhibiting sparsity or noise. As the dimensionality of the data increases, accurately determining the threshold becomes more complex, potentially leading to suboptimal covariance estimates.

Moreover, the method may need help to handle datasets with complex covariance structures effectively. When variables display intricate correlations or dependencies beyond the simplistic thresholding approach, Bickel and Levina's method may fail to estimate the underlying covariance matrix accurately. Additionally, the method lacks robustness in the presence of outliers or extreme values within the data, as it does not incorporate robustness measures into the estimation process. This susceptibility to outliers can skew the estimated covariance matrix, compromising the reliability of results, particularly in datasets where outliers are prevalent or influential. Furthermore, the computational complexity associated with determining an optimal threshold parameter challenges the method's scalability, particularly in large-scale or high-dimensional datasets with limited computational resources. These drawbacks emphasize the importance of cautious evaluation and consideration of alternative methods for practical applications in high-dimensional covariance estimation.

3.5 Oracle Approximating Shrinkage (OAS) Estimation - 2010

In 2010 Chen [22] developed a formula that, assuming the data were Gaussian-distributed, could be used to select a decrease in the shrinkage coefficient, resulting in a MSE lower than the one reported in the 2014 Ledoit-Wolf formula. They developed an improved version of the Ledoit-Wolf method by conditioning it with the Rao-Blackwell Ledoit-Wolf (RBLW) Covariance Estimation.

Let $\mathbf{X}$ be independent p-dimensional Gaussian vectors with covariance $\mathbf{\Sigma}$, and let $\mathbf{S}$ be the sample covariance of $\mathbf{X}$. The conditioned expectation of the LW covariance estimator is

$$\hat{\mathbf{\Sigma}}_{\mathrm{RBLW}} = (1 - \hat{\rho}_{\mathrm{RBLW}})\boldsymbol{S} + \hat{\rho}_{\mathrm{RBLW}}\hat{\boldsymbol{F}}$$

Where

$$\hat{\rho}_{\mathrm{RBLW}} = \frac{(n-2)/n.tr(\boldsymbol{S}^2) + (tr(\boldsymbol{S}))^2}{(n+2)[tr(\boldsymbol{S}^2) - (tr(\boldsymbol{S}))^2/p]}$$

The drawbacks of RBLW Covariance Estimation primarily revolve around its computational complexity and potential instability in high-dimensional settings. While RBLW aims to improve upon the Ledoit-Wolf estimator by leveraging Rao-Blackwellization to reduce bias, this approach may introduce additional computational overhead, particularly for large datasets with high dimensionality. Furthermore, RBLW may need to adequately address the challenges associated with sparsity and non-sparse covariance structures, potentially leading to suboptimal covariance estimates in datasets characterized by complex covariance patterns. Additionally, the performance of RBLW may be sensitive to the choice of tuning parameters, requiring careful calibration to achieve optimal results.

Additionally, it proposes an iterative approach that is comparable to the predictive shrinkage estimator in order to further minimize the estimation error. The resultant estimator, referred to as the Oracle Approximating Shrinkage [22], is an iterative approach to the OAS of a covariance matrix.

$$\hat{\mathbf{\Sigma}} = \rho\hat{F} + (1-\rho)\boldsymbol{S}$$

Where $\rho \in (0,1)$ is a control parameter/weight, $\mathbf{S}$ is an empirical covariance matrix, and $\hat{F}$ is a target matrix. It is proposed to use a structured estimate $\hat{F} = tr(\boldsymbol{S}/p) \cdot \mathbf{I}_{p\times p}$ where $\mathbf{I}_{p\times p}$ is an identity matrix of dimension p.

Numerical simulations demonstrate that the OAS approach can perform even better than RBLW, especially when n is much less than p [22].

OAS Estimation faces limitations beyond its struggles with low sphericity conditions, as its reliance on such assumptions restricts its utility to datasets with spherical covariance structures, potentially yielding inaccurate estimates. Additionally, OAS may falter when confronted with complex covariance patterns or high-dimensional datasets, lacking robustness against outliers that can skew results. These drawbacks underscore the necessity of carefully evaluating alternative covariance estimation methods for practical applications, especially in scenarios where precise covariance estimation is crucial for downstream analyses or model fitting. Considering OAS's limitations, exploring alternative approaches offering improved robustness, scalability, and accuracy in covariance estimation becomes essential, particularly in datasets with complex structures or high dimensionality, ensuring more reliable and valid statistical analyses and model predictions.

3.6 The Adaptive Thresholding Estimation - 2011

Cai and Liu introduced an adaptive variant of hard thresholding covariance estimation [26] in 2011, which was originally suggested by Bickel and Levina in 2008. The application of the thresholding technique to a correlation matrix is what adaptive thresholding is all about. Relationship, in which it adapts to each variable. That covariance estimator as given by,

$$\hat{\mathbf{\Sigma}}(\delta) = (\hat{\sigma}^*_{ij})_{p*p}$$

Where
$\hat{\sigma}^*_{ij} = S_{\lambda_{ij}}(\hat{\sigma}_{ij})$ and $\lambda_{ij} = \lambda_{ij}(\delta) = \delta\sqrt{\frac{\hat{\theta}_{ij}logp}{n}}$
$\hat{\theta}_{ij} = \frac{1}{n}\sum_{k=1}^{n}[(X_{ki} - \bar{X}^i)(X_{kj} - \bar{X}^j) - \hat{\sigma}_{ij}]^2$
$\bar{X}^i = \frac{1}{n}\sum_{k=1}^{n} X_{ki}$ and $\bar{X}^j = \frac{1}{n}\sum_{k=1}^{n} X_{kj}$
$\delta > 0$ is a regularization parameter

The demonstrated estimators adaptively converge over numerous sparse correlated matrices under standard deviation variations.

Adaptive Thresholding Estimation faces limitations beyond its narrow scope, primarily tailored for sparse normal mean vectors in specific contexts such as wavelet function estimation. This targeted approach restricts its effectiveness across various statistical scenarios, potentially hindering its utility in diverse applications. Furthermore, Adaptive Thresholding Estimation may need help to adequately capture the intricate covariance structure in datasets characterized by complex or highly variable relationships between variables. Its design, focusing on sparse mean vectors, may lead to suboptimal performance in scenarios diverging from its intended scope. These constraints underscore the need for alternative methods that offer greater versatility and robustness across a broader spectrum of datasets and statistical applications.

3.7 Positive Definite Sparse Covariance Estimation (PDSCE) - 2012

Using Convex optimization, Rothman proposed a covariance estimator called Positive Definite Sparse Covariance Estimator (PDSCE) in 2012 [27] to generate a sparse estimate of a covariance matrix, which is positive definite, and is suitable for high-dimensional environments.

To promote sparsity, a penalty of the lasso type is applied, and the logarithm barrier function is employed to enforce positive definiteness. It's denoted as,

$$\hat{\mathbf{\Sigma}} = \underset{\Sigma=\Sigma^{\mathrm{T}}}{\operatorname{argmin}}[\| \mathbf{\Sigma} - \boldsymbol{S} \|_F^2 + \lambda \| \mathbf{\Sigma}^{-1} \|_1 - \gamma log(|\mathbf{\Sigma}|)], \lambda, \gamma > 0$$

PDSCE has some limitations, especially when dealing with small sample sizes compared to the number of variables. It might not work well in datasets with few observations compared to the data's complexity. Additionally, PDSCE may need help accurately representing the relationships between variables in datasets with complex structures. This is because it is designed for sparse covariance matrices and might work less effectively with different covariance structures. These limitations show that we need other methods for estimating covariance that can adapt better to different statistical situations and be more reliable overall.

3.8 Non-Linear Shrinkage Estimation of Large Dimensional Covariance Estimation - 2015

In 2015, Ledoit and Wolf proposed a nonlinear shrinkage eigenvalue estimator for population covariance matrices that satisfies a mean-squared criterion for large-scale asymptotic functions [12].

That covariance estimator as given by,

$$\hat{\mathbf{\Sigma}} = Pdiag(P^T \mathbf{\Sigma} P)P^T$$

Where;
P = $(p_1, p_2, ..., p_p)$ is a matrix of eigenvectors for sample covariance matrix $\boldsymbol{S}$.

The eigenfunction is a numerical inverse of a multivariate nonrandom function, which they refer to as a Quantized Eigenvalues Sampling Transform (QuEST) function. Non-Linear Shrinkage Estimation of Large Dimensional Covariance Estimation is asymptotically optimal under the framework $\frac{p}{n} \to k > 0$ with respect to the class of rotation-equivariant estimators.

The limitations of Non-Linear Shrinkage Estimation extend beyond its applicability to non-sparse population parameters. This method may need help in datasets where the underlying covariance matrix lacks sparsity or contains many non-zero elements, potentially compromising its performance. Moreover, Non-Linear Shrinkage Estimation may need help to accurately characterize the covariance structure in datasets exhibiting intricate or highly variable relationships among variables. Its primary design for scenarios with sparse covariance matrices may hinder its ability to generalize effectively to other covariance structures prevalent in diverse datasets. These drawbacks emphasize the importance of considering alternative covariance estimation approaches that offer greater adaptability and robustness across various statistical contexts.

3.9 Joint PENalty Estimation of Covariance (JPEN) - 2016

In 2016, Maurya proposed a method for the estimation of well-defined and sparse covariance and inverse covariance from a high-dimensional sample of vectors derived from a sub-Gaussian distribution. The estimators proposed are derived from the minimization of the quadratic loss function, the joint penalty of the l_1 norm, and the variance of the eigenvalues of that norm.

The corresponding JPEN covariance matrix estimator [16] is;

$$\hat{\mathbf{\Sigma}}_R = D\widehat{\Gamma}D^T$$

Where;

$$\widehat{\Gamma} = \underset{\Gamma=\Gamma^T|(\lambda,\gamma)\epsilon\widehat{S}_1^R, tr(\Gamma)=tr(R)}{\operatorname{argmin}} [\| \Gamma - R \|_F^2 + \lambda \| \Gamma^- \|_1 + \gamma \textstyle\sum_{i=1}^p \sigma_i(\Gamma) - \bar{\sigma}(\Gamma)^2]$$

$$\widehat{S}_1^R = (\lambda, \gamma) : \lambda, \gamma > 0, \lambda \asymp \gamma \asymp \sqrt{\tfrac{log(p)}{n}} : \lambda < \tfrac{\sigma_{min}(R+\gamma I)}{C\sigma_{max}}(sign(R))$$

$C \geq 0.5$, and sign(R) is matrix of signs of elements of R and D is diagonal matrix of sample standard deviations.

In addition, the JPEN estimator reconstructs the sparsity pattern for the real covariance matrix and offers an approximate representation of the fundamental eigenrange.

Beyond its dependence on sparsely distributed covariance matrices, JPEN Estimation encounters limitations that affect its performance in scenarios where sparsity is absent in the covariance matrix of datasets. This means JPEN might not give accurate estimations when sparsity is missing, limiting its usefulness. Additionally, JPEN might struggle to understand the underlying covariance structure in datasets with complex or highly variable relationships between variables, since it's mainly designed for sparse covariance matrices. This limitation affects its performance in scenarios with different covariance structures. These drawbacks highlight the importance of exploring alternative covariance estimation methods that are more versatile and adaptable across various statistical situations.

3.10 Other Contributions

In 2011, Bien and Tibshirani suggested a Penalized Maximum Likelihood (PML) covariance estimation with a weighted lasso type penalty method on the basis of a sample of vectors drawn from a multivariate gaussian distribution [28].

$$\hat{\mathbf{\Sigma}} = \underset{\Sigma \succ 0}{\operatorname{argmin}}[log(det(\mathbf{\Sigma})) + tr(\boldsymbol{S}\mathbf{\Sigma}^{-1}) + \lambda \parallel \mathbf{\Sigma} \parallel_1], \lambda > 0$$

The prime objective of this process is to approximately reduce the size (over Sigma) of the following optimization problem that is not conformal.

$$minimize; \log\det(\mathbf{\Sigma}) + \operatorname{tr}(\mathbf{S}\ \mathbf{\Sigma}^{-1}) + \parallel \lambda * \mathbf{\Sigma} \parallel_1 \text{ subject to } \mathbf{\Sigma} \text{ positive definite.}$$

The empirical covariance matrix of the optimization problem must be positive definite in order for the problem to have a bounded objective. If this is not the case, a small constant must be added to the diagonal of the diagonal of $\mathbf{S}$. As the problem is non-convex, the matrix returned is not necessarily a global minimum for the problem.

Several popular covariance estimation methods, including PML covariance estimation and Augmented Sparse Principal Component Analysis (ASPCA), exhibit noteworthy drawbacks in specific contexts. PML covariance estimation is sensitive to the choice of penalty parameter, a critical factor influencing its performance, particularly in high-dimensional settings where data sparsity or noise is prevalent. Additionally, introducing a penalty term can lead to biased covariance estimates, and inaccuracies may arise when the penalty function needs to appropriately reflect the true sparsity or structure of the covariance matrix. Moreover, PML may need more robustness in the presence of outliers or extreme values due to the absence of robustness measures in its estimation process. On the other hand, while ASPCA addresses the challenge of estimating eigenvectors by decomposing the covariance matrix into

low-rank and sparse components, it encounters limitations in terms of computational complexity, especially for large datasets. Furthermore, ASPCA's performance relies heavily on selecting regularization and tuning parameters, posing challenges in scenarios with limited prior knowledge about the data structure. Additionally, ASPCA may not perform optimally in datasets with complex or highly variable relationships between variables, as it is primarily designed for scenarios with sparse and low-rank covariance matrices. These drawbacks underscore the importance of carefully considering the limitations and suitability of covariance estimation methods in various statistical contexts.

Xue [29] proposed a Positive Definite l_1 Penalized Estimation of Large Covariance Matrices. Positive Definite l_1 Penalized Estimation of Large Covariance (PDGLasso) is a method aimed at estimating high-dimensional covariance matrices that are both positive definite and sparse, which is particularly relevant when the number of variables (p) exceeds the number of observations (n). By applying an l_1 penalty to the off-diagonal elements of the covariance matrix, PDGLasso encourages many of these elements to be precisely zero, resulting in a sparse covariance matrix. This approach ensures that the estimated covariance matrix maintains positive definiteness, a crucial requirement for various statistical analyses. However, PDGLasso has limitations, including its computational complexity, especially for large datasets with numerous variables, requiring substantial computational resources. Additionally, the performance of PDGLasso can be sensitive to the choice of penalty parameter and tuning parameters, posing challenges in scenarios with limited prior knowledge about the data structure. Furthermore, PDGLasso may not perform optimally in datasets with intricate or highly variable relationships between variables, as it is primarily tailored for scenarios where the covariance matrix displays sparsity. Despite these drawbacks, PDGLasso offers advantages in estimating sparse and positive definite covariance matrices in high-dimensional settings, highlighting the importance of carefully considering its computational demands and parameter choices in practical applications of covariance estimation.

Bickel, P.J. proposed a hierarchical selection of variables in high-dimensional regression with a sparse representation of the regression function [30]. While effective for regression tasks, this approach's drawback as a covariance estimation method lies in its limited applicability to covariance matrices, particularly in scenarios where sparse representation may not accurately capture complex relationships between variables or adequately address computational challenges in covariance estimation for large datasets.

Friedman [24] introduced a simple and fast algorithm for estimation of a sparse inverse covariance matrix using an l_1 penalty. While efficient, its drawback in high-dimensional covariance estimation lies in its potential sensitivity to the choice of penalty parameter and the limited scope of the l_1 penalty in capturing complex covariance structures. Additionally, its performance may be compromised in datasets with intricate relationships between variables or when dealing with computational

constraints in large datasets.

In another interesting Levina's paper [31] proposed a sparse estimation of large covariance matrices via a nested lasso penalty. This approach aims to promote sparsity in the covariance matrix, facilitating easier interpretation and computational efficiency. However, its drawbacks include sensitivity to the choice of penalty parameters and potential challenges in accurately capturing complex covariance structures, particularly in high-dimensional settings with limited sample sizes.

Vu and Lei [32] analyzed the problem of estimating the subspace spanned by the principal eigenvectors of the population covariance matrix and they introduced two complementary notions of l_{q} subspace sparsity: row sparsity and column sparsity. This approach aims to capture the underlying structure of high-dimensional data more effectively. However, drawbacks may include computational complexity in implementing these concepts and challenges in generalizing to datasets with diverse covariance structures or limited sample sizes.

In 2013, Won [33] was proposed that the estimator should be maximized in relation to the normality of the data, but the condition number constraint should be applied to the estimator. This would result in the eigenvalues of the estimator being winsorized, while preserving the sample Eigenvectors of $\mathbf{\Sigma}$. Furthermore, it is demonstrated that the estimator has a lower entropy loss in comparison to $\mathbf{S}$, but it is not demonstrated that optimal nonlinearly shrinkability can be achieved with such a loss. The sample eigenvalue of the middle part of the sample remains unchanged, however, the higher eigenvalues are winsorized at specific constants.

It's denoted as,

$$\hat{\mathbf{\Sigma}} = \underset{\Sigma}{\operatorname{argmax}} L(\boldsymbol{S}, \mathbf{\Sigma})$$

subject to $\frac{\sigma_{max}(\mathbf{\Sigma})}{\sigma_{max}(\mathbf{\Sigma})} \leq K_{max}$. $\hat{\mathbf{\Sigma}}$ invertible if K_{max} finite and well-conditioned if K_{max} is moderate.

One disadvantage of Condition-Number-Regularized Covidance estimation is that it relies on the condition number regularization method. This method may sometimes only be appropriate for datasets with complicated or highly heterogeneous covariance structures. Additionally, this method may not accurately capture the underlying covariance relationships in datasets where the condition number regularization does not adequately address the covariance matrix's properties. Moreover, Won et al.'s approach may be sensitive to parameter tuning and could require careful calibration to achieve optimal performance, potentially limiting its applicability in practice. Overall, while this method may offer advantages in certain scenarios, its drawbacks emphasize the importance of considering alternative covariance estimation approaches that address a broader range of data characteristics and modeling requirements.

The sparsity assumption directly based on $\mathbf{\Sigma}$ is not suitable in many cases due to the presence of homogeneous factors. Instead, in 2013, Fan et al. proposed a non-parametric estimator for $\mathbf{\Sigma}$ is based on principal component analysis called Principal Orthogonal ComplEment Thresholding (POET) Estimation [34]. Suppose the eigenvalues of $\hat{\mathbf{\Sigma}}$ and $\hat{\xi}_{i\,i=1}^{\,p}$, i.e. $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq ... \geq \hat{\lambda}_p$, are the ordered eigenvectors of $\hat{\mathbf{\Sigma}}$. The resulting sample covariance has the spectral decomposition as follows.

$$\hat{\mathbf{\Sigma}} = \sum_{i=1}^{k} \hat{\lambda}_i \hat{\xi}_i \hat{\xi}_i^{'} + \hat{R}_k^{\tau}$$

Where;

$$\hat{R}_k^{\tau} = (\hat{r}_{ij}^{\tau})_{p*p}, \hat{r}_{ij}^{\tau} = \begin{cases} \hat{r}_{ii}, & i = j \\ \boldsymbol{S}_{ij}(\hat{r}_{ij})\boldsymbol{I}(\mid \hat{r}_{ij} \mid \geq \tau_{ij}) & i \neq j \end{cases}$$

$\tau_{ij} = \tau(\hat{r}_{ii}\hat{r}_{jj})^{1/2}$, for a given $\tau > 0$ where $\hat{r}_{ii}$ is the ith diagonal of $\hat{R}_K$. This corresponds to applying the thresholding with parameter τ to the correlation matrix of $\hat{R}_K$. This estimator is equivalent to the thresholding estimator (with proper chioce of τ) of Bickel and Levina (2008), Rothman (2009) and Cai and Liu (2011) (with a more generalized thresholding function) when $K = 0$.

The drawbacks of POET Estimation include its limited suitability for scenarios where the underlying covariance matrix deviates significantly from a diagonal structure. This means that POET Estimation may not perform optimally in datasets with complex or highly variable covariance structures, as it is designed primarily for scenarios where the covariance matrix is approximately diagonal or nearly sparse. Additionally, POET Estimation may not accurately capture the covariance relationships in datasets with strong off-diagonal dependencies or non-diagonal covariance patterns, potentially leading to suboptimal estimation performance. Overall, while POET Estimation may offer advantages in certain contexts, its drawbacks highlight the importance of considering alternative covariance estimation methods that can better accommodate diverse data structures and relationships.

In 2014, Abadir et al. proposed to split the data into two parts, $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$ where $\mathbf{X}_1$ has size $p * n_1$ and $\mathbf{X}_2$ is $p * n_2$ with $n = n_1 + n_2$ and provided a theoretical-supported data partitioning scheme for asymmetric efficiency. This method of covariance estimation called Grand Average Covariance Estimation [35].

That covariance estimator as given by,

$$\hat{\mathbf{\Sigma}} = P(M^{-1} \sum_{j=1}^{M} diag(P_{1j}^T \widetilde{\Sigma}_2^{(j)} P_{1j}))P^T$$

Where;

$\widetilde{\Sigma}^{(j)} = n_i^{-1} X_i^{(j)} X_i^{(j)T} = P_{ij} D_{ij} P_{ij}^T, i = 1, 2.$
P_i is the matrix of eigenvectors for $\widetilde{\Sigma}_i$
$\widetilde{\Sigma}_i = n_i^{-1} X_i X_i^T, i = 1, 2.$

The eigenvectors are estimated from a limited subset of the data, which are then transformed into roughly orthogonal series, resulting in a well-conditioned estimation despite the fact that the data contains fewer observations than dimensions. Furthermore, the estimator is designed without any assumptions regarding the distribution of the random sample or any parametric structure associated with the variance matrix.

The drawbacks of Grand Average Covariance Estimation include its limited suitability for datasets with high variability or non-stationarity in covariance structures. This method relies on averaging individual covariance estimates across multiple datasets, assuming a consistent covariance structure across all datasets. However, in scenarios where the covariance structure varies significantly between datasets or changes over time, the grand average may not accurately represent the true underlying covariance. Additionally, this method may be sensitive to outliers or extreme values in individual datasets, leading to biases in the estimated grand average covariance. Overall, while Grand Average Covariance Estimation may offer simplicity and computational efficiency, its reliance on the assumption of consistent covariance structure across datasets limits its applicability in diverse data settings.

The Nonparametric Eigenvalue Regularized Covariance Matrix Estimation (NERCOME) was proposed by Lam in 2016 using the same data partitioning concept in the Non-Linear Shrinkage Estimation of Large Dimensional Covariance Estimation [36].

That covariance estimator as given by,

$$\hat{\mathbf{\Sigma}} = P_1 diag(P1_1^T \widetilde{\Sigma}_2 P_1) P1_1^T$$

Importance about Nonparametric Eigenvalue Regularized Covariance Matrix Estimation is that the convergence to Nonlinear shrinkage eigenvalue estimator $\hat{\mathbf{\Sigma}}, 1$ means that NERCOME estimator is also a nonlinear shrinkage estimator like $\hat{\mathbf{\Sigma}}$, with only P replaced by P_1.

The drawbacks of NERCOME include its sensitivity to the choice of regularization parameters and kernel functions. NERCOME relies on nonparametric methods to estimate the covariance matrix, which involves selecting appropriate smoothing parameters and kernel functions. However, the performance of NERCOME can be heavily influenced by these choices, and selecting optimal parameters can be challenging, especially in high-dimensional datasets or when the underlying covariance structure is complex. Additionally, NERCOME may not perform well when the dataset contains outliers or non-Gaussian data, as it assumes a certain smoothness in

the covariance structure that may not hold in such scenarios. Overall, while NERCOME offers flexibility in estimating covariance matrices without strict parametric assumptions, its performance is sensitive to the selection of regularization parameters and kernel functions.

In 2018, NOVEL Integration of the Sample and Thresholded (NOVELIST) covariance estimators was introduced by Huang and Fryzlwicz [37], which is a combination of linear shrinkage with sparse estimators, as follows.

$$\hat{\mathbf{\Sigma}}_N = (1-\delta)\boldsymbol{S} + \delta T_\lambda(\boldsymbol{S})$$

$T_\lambda(\boldsymbol{S})$ is a thresholded estimator of $\mathbf{S}$ with parameter λ $(T_\lambda(\boldsymbol{S}) = (s_{ij}\mathbf{1}(\mid s_{ij} \mid \geq \lambda)))$

The sample variance is reduced by the NOVELIST estimator to a variant that is thresholded. High dimensionality can lead to a low-rank component for the sample variance component, which is not sparse. The addition of this component ensures the stability of NOVELIST's invertibility because the thresholded sample variance component is sparse. The NOVELIST estimator is advantageous due to its simplicity, straightforward implementation, high computational efficiency, and the absence of eigenanalysis in its application.

The drawbacks of NOVELIST covariance estimation include its sensitivity to the choice of threshold parameter and the potential for bias in estimating the covariance matrix. NOVELIST combines information from the sample covariance matrix with thresholding techniques to improve estimation accuracy. However, the performance of NOVELIST heavily depends on selecting an appropriate threshold value, which can be challenging, especially in datasets with complex covariance structures or high levels of noise. Additionally, the thresholding process may introduce bias into the estimation, particularly if the chosen threshold is not optimal for the dataset. Overall, while NOVELIST offers a novel approach to covariance estimation, its performance is contingent on the selection of threshold parameters and may be susceptible to bias.

In addition to frequentist statistics, Bayesian perspectives [38], [24], [39], [40], [41], [42]) proposed to regularize $\hat{\mathbf{\Sigma}}$ by shrinking sample eigenvalues.

4 Discussion and Future Work

Estimating covariance matrices in high dimensions poses a significant challenge due to the poor performance of the sample covariance matrix. As such, explicit regularization techniques are necessary to improve its accuracy. This paper explores various approaches to covariance regularization, including imposing specific structures during estimation and shrinking extreme eigenvalues of the sample covariance matrix. The choice of method depends on the application context and the availability of prior knowledge about the population covariance matrix's structures, with shrinkage emerging as a viable option in the absence of such knowledge.

In 1976, British statistician George Box famously remarked, "All statistical models and tools are wrong, but some are useful." This sentiment rings true in the realm of high-dimensional covariance estimations, where numerous challenges persist despite the availability of various methods.

Among the options considered for high-dimensional matrix covariance estimation, Ledoit-Wolf Estimation emerges as particularly suitable for datasets with $n < p$ settings. However, it is not without limitations. One drawback is its lack of sparsity, which can impact its performance in scenarios where the covariance matrix is sparse. Additionally, the dependency of the estimated shrinkage intensity $\hat{\rho}$ on the unknown true covariance matrix $\mathbf{\Sigma}$ poses a challenge, as it requires accurate estimation of $\mathbf{\Sigma}$ for effective shrinkage. Furthermore, Ledoit-Wolf Estimation tends to overestimate non-zero elements in the covariance matrix, and its inconsistency due to the preservation of eigenvectors may limit its applicability in certain scenarios. Despite these drawbacks, Ledoit-Wolf Estimation remains valuable for its effectiveness in specific contexts, especially when dealing with high-dimensional datasets where traditional methods may fall short.

Looking ahead to future research in principal component analysis (PCA) and covariance matrix estimation, there is an opportunity to address the limitations of existing techniques, particularly in scenarios where the sample size is smaller than the dimensionality of the data $n < p$. This could involve the development of novel approaches that improve estimation accuracy, enhance robustness to outliers, and adapt to varying data structures. Exploring the impact of non-normal distributions and complex covariance structures on covariance estimation could also provide valuable insights for developing more robust and versatile techniques. By addressing these challenges, future research endeavors aim to enhance the applicability and effectiveness of covariance estimation methods in high-dimensional settings.

References

[1] Sartorius: What Is Principal Component Analysis (PCA) and How It Is Used? (2020). https://www.sartorius.com/en/knowledge/science-snippets/what-is-principal-component-analysis-pca-and-how-it-is-used-507186

[2] Mardia, K.: Kent. jt; bibby, jm 1979: Multivariate analysis. Probability and Mathematical Statistics., Academic Press, London

[3] Anderson, T.W.: Estimating linear statistical relationships. The Annals of Statistics, 1–45 (1984)

[4] Gorsuch, R.L.: Common factor analysis versus component analysis: Some well and little known facts. Multivariate Behavioral Research **25**(1), 33–39 (1990)

[5] Guttman, L.: Some necessary conditions for common-factor analysis. Psychometrika **19**(2), 149–161 (1954)

[6] Kaiser, H.F.: The application of electronic computers to factor analysis. Educational and psychological measurement **20**(1), 141–151 (1960)

[7] Cattell, R.B.: The scree test for the number of factors. Multivariate behavioral research **1**(2), 245–276 (1966)

[8] Zwick, W.R., Velicer, W.F.: Comparison of five rules for determining the number of components to retain. Psychological bulletin **99**(3), 432 (1986)

[9] Geman, S.: A limit theorem for the norm of random matrices. The Annals of Probability, 252–261 (1980)

[10] Bickel, P.J., Levina, E., *et al.*: Some theory for fisher's linear discriminant function,naive bayes', and some alternatives when there are many more variables than observations. Bernoulli **10**(6), 989–1010 (2004)

[11] Ledoit, O., Wolf, M., *et al.*: Some hypothesis tests for the covariance matrix when the dimension is large compared to the sample size. The Annals of Statistics **30**(4), 1081–1102 (2002)

[12] Ledoit, O., Wolf, M.: Spectrum estimation: A unified framework for covariance matrix estimation and pca in large dimensions. Journal of Multivariate Analysis **139**, 360–384 (2015)

[13] Ledoit, O., Wolf, M., *et al.*: Optimal estimation of a large-dimensional covariance matrix under stein's loss. Bernoulli **24**(4B), 3791–3832 (2018)

[14] Durrant, R.J., Kabán, A.: Random projections as regularizers: learning a linear discriminant from fewer observations than dimensions. Machine Learning **99**(2), 257–286 (2015)

[15] Yata, K., Aoshima, M.: Effective pca for high-dimension, low-sample-size data with singular value decomposition of cross data matrix. Journal of multivariate analysis **101**(9), 2060–2077 (2010)

[16] Maurya, A.: A well-conditioned and sparse estimation of covariance and inverse covariance matrices using a joint penalty. The Journal of Machine Learning Research **17**(1), 4457–4484 (2016)

[17] Johnstone, I.M., Lu, A.Y.: On consistency and sparsity for principal components analysis in high dimensions. Journal of the American Statistical Association **104**(486), 682–693 (2009)

[18] Ledoit, O., Péché, S.: Eigenvectors of some large sample covariance matrix ensembles. Probability Theory and Related Fields **151**(1-2), 233–264 (2011)

[19] Rajaratnam, B., Vincenzi, D.: A theoretical study of stein's covariance estimator. Biometrika **103**(3), 653–666 (2016)

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

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

[22] Chen, Y., Wiesel, A., Eldar, Y.C., Hero, A.O.: Shrinkage algorithms for mmse covariance estimation. IEEE Transactions on Signal Processing **58**(10), 5016–5029 (2010)

[23] Won, J.H., Kim, S.-J.: Maximum likelihood covariance estimation with a condition number constraint. In: 2006 Fortieth Asilomar Conference on Signals, Systems and Computers, pp. 1445–1449 (2006). IEEE

[24] Friedman, J., Hastie, T., Tibshirani, R.: Sparse inverse covariance estimation with the graphical lasso. Biostatistics **9**(3), 432–441 (2008)

[25] Bickel, P.J., Levina, E.: Covariance regularization by thresholding (2008)

[26] Cai, T., Liu, W.: Adaptive thresholding for sparse covariance matrix estimation. Journal of the American Statistical Association **106**(494), 672–684 (2011)

[27] Rothman, A.J.: Positive definite estimators of large covariance matrices. Biometrika **99**(3), 733–740 (2012)

[28] Bien, J., Tibshirani, R.J.: Sparse estimation of a covariance matrix. Biometrika **98**(4), 807–820 (2011)

[29] Xue, L., Ma, S., Zou, H.: Positive-definite $\ell 1$-penalized estimation of large covariance matrices. Journal of the American Statistical Association **107**(500), 1480–1491 (2012)

[30] Bickel, P.J., Ritov, Y., Tsybakov, A.B.: Hierarchical selection of variables in sparse high-dimensional regression. In: Borrowing Strength: Theory Powering Applications–a Festschrift for Lawrence D. Brown, vol. 6, pp. 56–70 (2010). Institute of Mathematical Statistics

[31] Levina, E., Rothman, A., Zhu, J., *et al.*: Sparse estimation of large covariance matrices via a nested lasso penalty. The Annals of Applied Statistics **2**(1), 245–263 (2008)

[32] Vu, V.Q., Lei, J., *et al.*: Minimax sparse principal subspace estimation in high dimensions. The Annals of Statistics **41**(6), 2905–2947 (2013)

[33] Won, J.-H., Lim, J., Kim, S.-J., Rajaratnam, B.: Condition-number-regularized covariance estimation. Journal of the Royal Statistical Society Series B: Statistical Methodology **75**(3), 427–450 (2013)

[34] Fan, J., Liao, Y., Mincheva, M.: Large covariance estimation by thresholding principal orthogonal complements. Journal of the Royal Statistical Society Series B: Statistical Methodology **75**(4), 603–680 (2013)

[35] Abadir, K.M., Distaso, W., Žikeš, F.: Design-free estimation of variance matrices. Journal of Econometrics **181**(2), 165–180 (2014)

[36] Lam, C.: Nonparametric eigenvalue-regularized precision or covariance matrix estimator (2016)

[37] Huang, N., Fryzlewicz, P.: Novelist estimator of large correlation and covariance matrices and their inverses. Test **28**, 694–727 (2019)

[38] Rajaratnam, B., Massam, H., Carvalho, C.M., *et al.*: Flexible covariance estimation in graphical gaussian models. The Annals of Statistics **36**(6), 2818–2849 (2008)

[39] Qian, W., Brown, P.: Bayes sequential decision theory in clinical trials. Bayesian Statistics **6**, 829–838 (1999)

[40] Banerjee, O., Ghaoui, L.E., d'Aspremont, A.: Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. Journal of Machine learning research **9**(Mar), 485–516 (2008)

[41] Daniels, M.J., Kass, R.E.: Shrinkage estimators for covariance matrices. Biometrics **57**(4), 1173–1184 (2001)

[42] Peng, J., Wang, P., Zhou, N., Zhu, J.: Partial correlation estimation by joint sparse regression models. Journal of the American Statistical Association **104**(486), 735–746 (2009)