Improving Finite Sample Estimates of Principal Components

A thesis submitted in
partial fulfillment of the requirements for the Degree of
Doctor of Philosophy in Statistics
at the Department of Mathematics
The University of Waikato

By
Nuwan Weeraratne

THE UNIVERSITY OF
WAIKATO
Te Whare Wānanga o Waikato

The University of Waikato
August 2024

Abstract

Principal Component Analysis (PCA) is a method of compressing high-dimensional data into a lower-dimensional format that captures the essence of the original structure. PCA is a matrix decomposition technique based on eigendecomposition. It quantifies relationships between variables using covariance matrices, captures the shape of the data distribution, and evaluates the importance of directions using eigenvalues. Therefore, the accuracy of the variance-covariance estimation is crucial for reliable PCA. In high-dimensional settings where the number of observations (n) is much smaller than the number of variables (p) (i.e., $n \ll p$), the conventional maximum likelihood estimator (MLE) of covariance becomes poorly conditioned and yields unreliable principal components. To address these limitations, we propose a novel estimation framework called Pairwise Differences Covariance (PDC), along with four regularized extensions: Standardized PDC (SPDC), Local Scaled PDC (LSPDC), Maximum Absolute Scaled PDC (MAXPDC), and Range Scaled PDC (RPDC). These estimators increase the effective sample size by utilizing all pairwise differences within the data, thereby enhancing estimation stability without requiring additional data collection. Extensive experiments on synthetic and real datasets demonstrate that the proposed estimators, particularly SPDC, significantly reduce the overdispersion of the first principal component and improve directional accuracy. On average, SPDC reduced cosine similarity error by approximately 10–30% and narrowed eigenvalue spread by 10–20% compared to MLE and Ledoit-Wolf estimators in $n \ll p$ HDLSS scenarios. Real-world applications confirm the practical utility and robustness of these methods for high-dimensional data analysis.

Keywords: Dimensionality Reduction, Principal Component Analysis, Covariance Estimation, High-Dimensional Data.

Note on Publications

This Ph.D. is a **"by Thesis Only"** Ph.D. The details of the publications are as follows.

Journal Papers

1. Weeraratne, N., Hunt, L. and Kurz, J. (2024) Challenges of Principal Component Analysis in High-Dimensional Settings when $n < p$. Preprint, 1-32. https://www.researchsquare.com/article/rs-4033858/v1 https://doi.org/10.21203/rs.3.rs-4033858/v1
   Note: This article has **six citations.**

2. Weeraratne, N., Hunt, L., & Kurz, J. (2025). Principal Component Analysis When n < p: Challenges and Solutions. arXiv preprint arXiv:2503.17560. https://doi.org/10.48550/arXiv.2503.17560
   Note: (Accepted for Publication) Computational Statistics - Springer Journal.

R Packages

Ongoing: Development of an R package for Pairwise Differences Covariance (PDC) Estimation, including implementing four regularized versions. This project focuses on creating a comprehensive and user-friendly tool to facilitate the application of these advanced statistical techniques, contributing to the broader field of covariance estimation in statistical analysis.

Conference Proceedings

1. On U-estimation of Principal Components when n < p, 16th International Conference of the ERCIM WG on Computational and Methodological Statistics (CMStatistics 2023) — HTW Berlin, University of Applied Sciences, Berlin, Germany, 16-18 December 2023.

2. Pairwise Differences Covariance (PDC) Estimation in Principal Component Analysis when n < p, 73rd Annual New Zealand Statistical Association (NZSA) Conference, University of Canterbury, Christchurch, New Zealand — 04-06 December 2023 — High-Dimensional Statistics Track.

3. Improving the precision of finite sample estimates of the principal components using the Pairwise Differences Covariance Estimation, 8th AUT Mathematical Modelling and Analytics Symposium, Department of Mathematical Sciences at Auckland University of Technology and Mathematical Modelling and Analytics Research Centre (MMARC), New Zealand — 29-30 November 2023.

4. Issues with the Principal Component Analysis under High Dimensionality when n < p, 46th Annual Research Students' Conference on Probability & Statistics at the University of Sheffield, United Kingdom, sponsored by The Royal Statistical Society.

5. Improving the Principal Component Analysis using the Pairwise Differences Covariance Estimation, 18th New Zealand Mathematics and Statistics Postgraduate Conference, Victoria University of Wellington, New Zealand — 29 November – 01 December — Statistics Track.

6. Optimising PCA for Health and Care Research: A Reliable Approach to Component Selection, NIHR Statistics Group Conference 2025, Sheffield, UK — 19 June - 20 June — Poster Presentation.

Public Seminar

1. A seminar entitled "Is PCA telling statistical lies?" at the Department of Statistics - University of Auckland, New Zealand — Statistics Seminar Series – Semester II, 2023 (Talk #3) — October 25, 2023 @ 11.00 am – 12.00 pm.

Acknowledgements

I am deeply grateful to the following individuals and institutions, whose support has been instrumental in completing my Ph.D. thesis.

Firstly, I extend my heartfelt appreciation to my wife, whose unwavering encouragement, understanding, and calm listening have provided immense comfort and motivation. She has been our family hero, multitasking between earning, managing household responsibilities, and caring for our daughter. Her dedication has sometimes meant sacrificing her time, as she has seamlessly taken on roles of both dad and mom and everything in between. Her support has allowed me to focus on my research and has alleviated stress during challenging times.

I offer my daughter my deepest thanks for her unconditional love and warm hugs, which have been a constant source of motivation and stress relief. Despite her age, her belief in me has been unwavering. She often tells me, "Dad, you are my hero," and reassures me with encouraging words such as, "Don't worry, Dad; you can do anything. Trust me." Her support has been a beacon of strength throughout my academic journey.

I am profoundly grateful to the AHEAD World Bank Scholarship for providing the funding opportunity that enabled me to undertake my studies at the University of Waikato. This support has been crucial in advancing my research and academic journey.

I would also like to express my deep appreciation to Dr. Lyn Hunt, my chief supervisor, and Dr. Jason Kurz, my co-supervisor, for their invaluable advice and continuous support throughout my Ph.D. studies. Their guidance has been instrumental in shaping my research proposal, and their immense knowledge and experience have been a source of encouragement in my academic and daily life.

I wish to thank Dr. Bob Durrant especially for accepting my Ph.D. research proposal and initiating my research problem as my first chief supervisor despite his unforeseen departure from the university.

I would also like to acknowledge Dr. Niroshan Withanage and Dr. Nimalakith Samarakoon for initially pushing me to apply for a Ph.D. placement. Their encouragement and motivation were pivotal in inspiring me to embark on this academic track.

Additionally, I sincerely appreciate the continuous care and support from my wife's and my moms, dads, and brothers. Their presence and encouragement have helped me stay strong and focused throughout this journey.

I am also grateful to Dr. Hiruni Rupasinghe, Head of the Department, and Dr. Sampath Fernando, Dean of the Faculty, for assisting with leave approval and other administrative tasks at my home country university. Their support has been invaluable in smoothing the process of my Ph.D. journey.

Lastly, I would like to express my gratitude to Mrs. Amanda Sajiwanie and Mr. Achintha Ekanayake, who have graciously agreed to act as bailors to accept my university bond in my home country. Their assistance has been a significant help in facilitating this critical aspect of my academic journey.

These individuals and organizations have played a crucial role in my journey, and I am truly grateful for their support and contributions.

Contents

List of Tables

List of Figures

Nomenclature

n	Number of observations
p	Number of dimensions
$\boldsymbol{\mu}$	Population mean vector
$\hat{\boldsymbol{\mu}}$	Sample mean vector
$\boldsymbol{\Sigma}$	Population covariance matrix
$\boldsymbol{S}$	Sample covariance matrix or $\hat{\boldsymbol{\Sigma}}$
$\mathbf{S}_{\mathrm{ii}}$	Variance of i^{th} dimension
$\mathbf{S}_{\mathrm{ij}}$	Covariance of i^{th} and j^{th} dimensions
$\boldsymbol{\Theta}$	Precision matrix ($\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1}$)
$\mathbf{X}$	A $n \times p$ random data matrix with mean $\mathbf{0}$ and variance $\boldsymbol{\Sigma}$
r	Number of principal components
k	Condition number of the sample covariance matrix
$\mathbf{P}$	Matrix of eigenvectors of the sample covariance matrix
$\| \boldsymbol{X} \|_2$	The l_2 norm of N-dimensional vector $\boldsymbol{X}$
$\| \boldsymbol{X} \|_{\mathrm{F}}$	Frobenius norm of $n \times p$ matrix $\mathbf{X}$
$\mathbb{R}$	Real numbers
λ_{i}	i^{th} eigenvalue of the sample covariance matrix
e_{i}	i^{th} eigenvector of the sample covariance matrix

$\mathbf{U}$	An $n \times n$ orthogonal matrix containing the right singular vectors
$\mathbf{V}$	A $p \times p$ orthogonal matrix containing the left singular vectors
$\mathbf{M}$	An $n \times p$ diagonal matrix containing the singular values
$\mathbf{0}$	A $p \times 1$ zero vector
$\mathbf{1}$	A $p \times p$ unit matrix
$\mathbf{I}$	A $p \times p$ identity matrix
$\mid \mathbf{\Sigma} \mid$	The determinant of the covariance matrix
$tr(\mathbf{\Sigma})$	Sum of elements on the main diagonal of $\mathbf{\Sigma}$ (trace of $\mathbf{\Sigma}$)
$\parallel \mathbf{\Sigma} \parallel_1$	The maximum absolute column sum of $\mathbf{\Sigma}$

Abbreviations

AI Artificial Intelligence

CSE Cosine Similarity Error

DR Dimensionality Reduction

HD High Dimensional

LDA Linear Discriminant Analysis

LSPDC Local Scaled Pairwise Differences Covariance

LW Ledoit Wolf

MAXPDC MAXimum absolute value Pairwise Differences Covariance

ML Machine Learning

MLE Maximum Likelihood Estimation

PC Principal Component

PCA Principal Component Analysis

PDC Pairwise Differences Covariance

PDSCE Positive Definite Sparse Covariance Estimation

RPDC Range Pairwise Differences Covariance

SPDC Standardized Pairwise Differences Covariance

SVD Singular Value Decomposition

Chapter 1

Introduction

1.1 Research Background

Over the last two decades, technology has developed rapidly, minimizing the processing time of most statistical applications and algorithms. Because technology advances quickly, researchers constantly introduce new technological concepts and methodologies to the research world. In that context, Machine Learning (ML) is a trending topic in the Research and Development (R&D) industry. As Arthur Samuel said in 1959, ML is a vital part of Artificial Intelligence (AI) that focuses on developing a statistical algorithm that can learn from data and generalize to new unsighted data without detailed human guidance or instruction [1]. Supervised, unsupervised, semi-supervised, and reinforcement learning are the four main ML methods used in practice. A wide range of fields use ML algorithms. However, the choice of an ML algorithm depends on the characteristics of the data, expected usage, desired performance, domain of the statistical application, and other factors.

However, ML algorithms are now often used to operate High-Dimensional (HD) data in various applications across diverse fields, such as bioinformatics, forensic science, cyber security, agriculture, healthcare, finance, and computer science. However, fitting models in HD spaces demands substantial resources and time. The challenge intensifies due to the "Curse of Dimensionality," [2] where increasing dimensions lead to sparsity in data, complicating model training and potentially causing overfitting. Regularization or Dimensionality Reduction (DR) techniques are essential to mitigate this challenge. DR aids in compressing data while retaining essential information, facilitating more efficient and effective model-building processes across domains. We can elaborate on the general idea of the phenomena called "the curse of dimensionality" as shown in Figure 1.1 (taken from [3]).

DR is a fundamental technique used in intermediate data analysis and Unsupervised ML algorithms, transforming the original data into a lower-dimensional space while preserving its essential characteristics. These algorithms find hidden patterns in given data without any detailed human guidance or instruction. To gain better insights into HD data, it enhances the effectiveness of ML models by simplifying the structure of data, eliminating redundancy, reducing computation time, and improving visualization capability.

In general, DR techniques can be divided into two main types: feature selection and feature extraction. The feature selection method comprises selecting the most critical subset of the original dimensions from the existing dataset [4]. It does not create any new dimensions based on the existing dataset. Meanwhile, feature extraction methods transform the original HD data into a lower-dimensional space, and in the case of linear methods such as Principal Component Analysis (PCA), the new dimensions are constructed as linear combinations of the original variables. [5]. Researchers use feature extraction methods more widely than feature selection methods. Several feature extraction methods exist, each with its unique approach and underlying principles.

Some of the most widely used techniques are PCA, Linear Discriminant Analysis (LDA), t-distributed Stochastic Neighbor Embedding (t-SNE), Autoencoders, Independent Component Analysis (ICA), Multidimensional Scaling (MDS), Non-negative Matrix Factorization (NMF), and others.

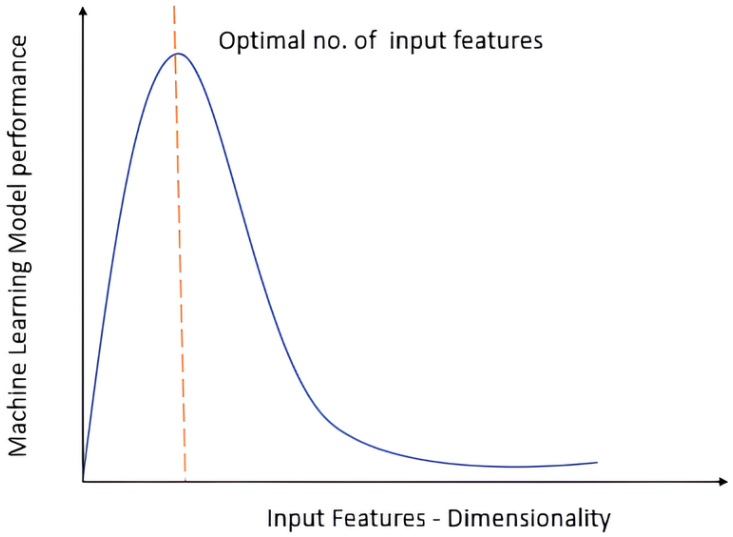


Figure 1.1: The Curse of Dimensionality

PCA is the widely used method for DR among these feature extraction methods. PCA ensures reasonable data representation by transforming original HD data into a lower-dimensional space while keeping most of the total variance. When dealing with HD data, PCA facilitates more efficient ML algorithms by shaping the data meaningfully. Otherwise, the data's redundancy minimizes the effectiveness of most HD statistical approaches. However, in recent years, there has been significant growth in utilizing HD data in situations where fewer observations than the number of dimensions. However, existing most commonly used DR techniques, including PCA, have not been effective in such kind of situations because most of the DR methods rely on the mean estimation, and with limited observations, the mean is more likely to be influenced by random noise and outliers, leading to inaccurate centering. This challenge has verved a significant research interest to the researchers who focused mostly on issues of HD data analysis. Therefore, this research aims to rigorously examine the effectiveness of PCA under high-dimensional settings where $n \ll p$, with particular focus on the asymptotic regime in which both n and p grow large while $p/n \to c > 1$. These scenarios are known to challenge the reliability of classical PCA due to sample insufficiency and rank deficiency in the sample covariance matrix. By doing so, the goal is to enhance our ability to extract meaningful insights from HD data despite the limitations imposed by the insufficient sample size and the number of dimensions.

1.2 Problem Statement

As we discussed earlier, PCA transforms the original dataset into a differently structured dataset by using linear combinations of original dimensions while keeping the original dataset's variance-covariance structure as much as possible. That means PCA depends on the variance-covariance structure of the dataset. As primarily used covariance estimation, PCA also uses the Maximum Likelihood Estimation (MLE) of covariance estimation to obtain the variance-covariance structure of the original dataset and it was initially introduced by Anderson in 1970. While the traditional MLE of covariance, denoted as $\mathbf{S}$, is recognized for its asymptotic unbiasedness, its efficacy becomes notably challenged in HD scenarios where $n < p$. This problem arises because MLE needs more observations than the number of dimensions. When $n < p$, the sample mean becomes more sensitive to random noise and outliers, which increases variability and reduces the reliability of the estimated central tendency for each dimension. As a result, the covariance estimates might not accurately reflect the data structure, and its Principal Component (PC) estimates will be less effective. Fixing this issue is essential for reliable covariance estimates, especially with HD data. Using PCA on such data presents four main challenges.

1. **Impact of $n < p$ on Covariance Estimation Accuracy**

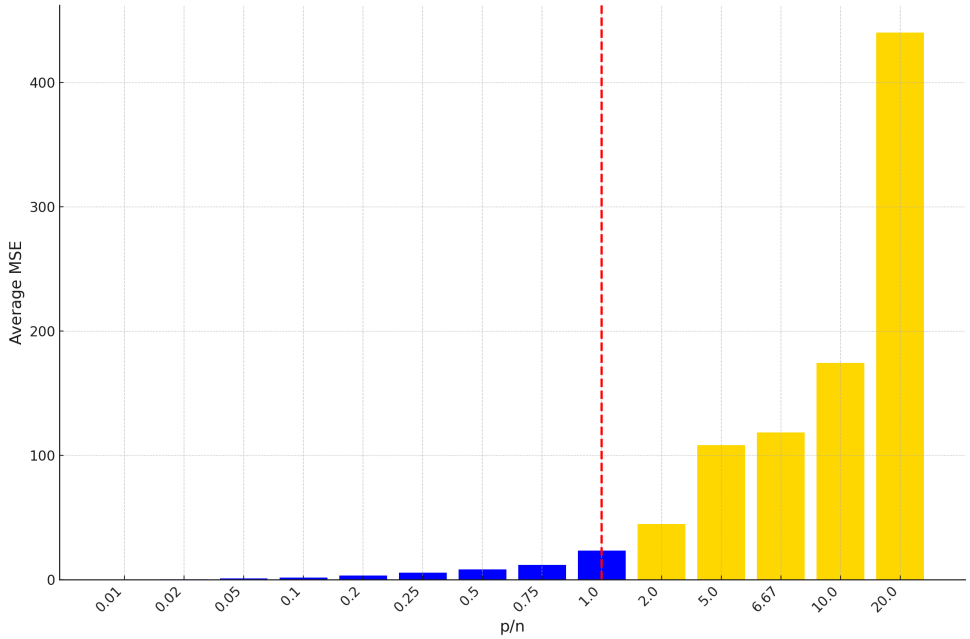


Figure 1.2: Effect of p/n with respect to MSE of $\mathbf{S}$

Figure 1.2 illustrates how the *Mean Squared Error (MSE)* of the sample covariance matrix S changes as the sample size n decreases, with the number of variables fixed at $p = 100$. The data are simulated from a multivariate normal distribution $X \sim \mathcal{N}_p(0, \Sigma)$, where the true covariance matrix Σ has a general, fixed structure and is scaled by a constant factor σ^2.

The MSE of S is defined as:

$$\text{MSE}(S) = \mathbb{E}\left[\|S - \Sigma\|_F^2\right],$$

where $\|\cdot\|_F$ denotes the Frobenius norm, and the expectation is taken over repeated samples of X.

Each point in the figure corresponds to the average MSE computed over 1000 simulated datasets for a given value of n, with n ranging from 10,000 down to 5. The MSE is calculated element-wise over all entries of the $p \times p$ covariance matrix and then averaged.

The plot clearly shows that when n is large relative to p, the sample covariance matrix S closely approximates Σ, resulting in a low MSE. However, as n decreases—especially when $n < p$—the estimation accuracy deteriorates rapidly, and the MSE increases sharply. This increase is due to the high variability and overfitting risk associated with estimating a large number of covariance parameters from a limited number of observations.

Let $\mathbf{X} \in \mathbb{R}^{n \times p}$ represent a data matrix consisting of n observations in p dimensions. We primarily focus on the high-dimensional (HD) setting where $p/n > \epsilon$ for some constant $\epsilon > 0$, particularly the challenging case where $n \ll p$, which is common in practical applications.

Stuart Geman [6] showed that if the rows of $\mathbf{X}$ are drawn independently from a multivariate normal distribution, i.e., $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$, and if $p/n \to \theta$ as $n \to \infty$ for some $0 < \theta < \infty$, then the largest sample eigenvalue $\hat{\lambda}_1$ of the sample covariance matrix satisfies

$$\hat{\lambda}_1 \to \sigma^2(1 + \sqrt{\theta})^2.$$

In particular, when $\sigma^2 = 1$, the limiting value of the largest sample eigenvalue becomes $(1+\sqrt{\theta})^2$. Although our simulation study does not follow this exact asymptotic regime—since it focuses on finite-sample, high-dimensional scenarios where $n \ll p$—Geman's result is cited to provide theoretical context. It highlights the tendency for the leading sample eigenvalue to become increasingly biased when p/n does not approach zero, a phenomenon also reflected in our empirical results.

For example, when $n = 2p$, p, $0.75p$, $0.5p$, and $0.25p$, the corresponding limiting values of $\hat{\lambda}_1$ are approximately 2.91, 4, 4.64, 9, and 25, respectively, while the true largest eigenvalue remains $\sigma^2 = 1$. This illustrates how the inconsistency of the sample covariance estimator becomes more severe as n decreases relative to p, supporting the motivation for improved estimation methods in such settings.

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

   Numerous empirical studies, including [7], [8], [9], [10], [11], [12], and [13], 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.

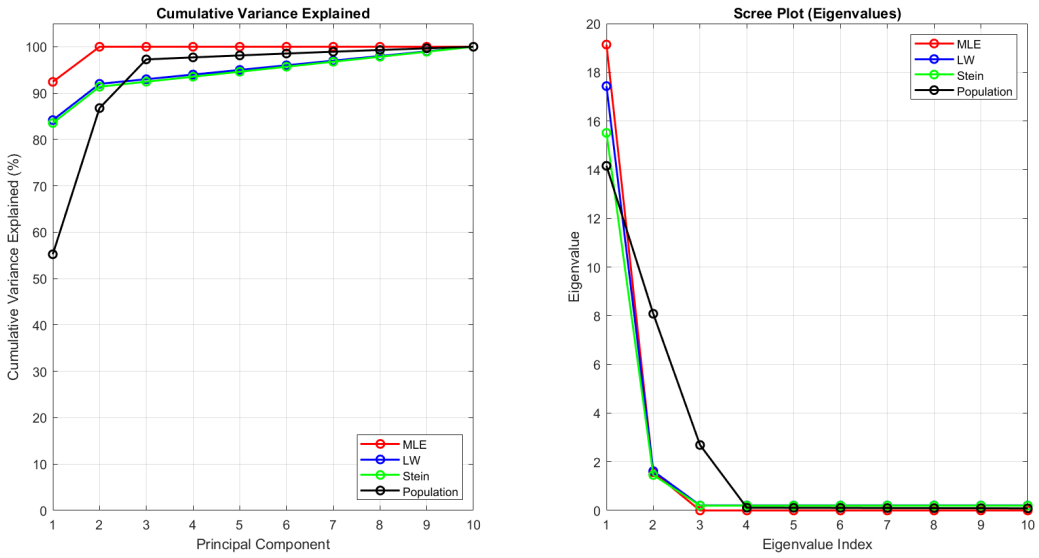


Figure 1.3: Over-dispersion of sample eigenvalues $\mathbf{S}$

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 Figure 1.3. Thus, it is crucial to consider alternative methods for covariance estimation in HD datasets where the number of dimensions surpasses the number of observations to ensure accurate and reliable results in statistical analyses.

Figure 1.3 illustrates the overdispersion of sample eigenvalues when $n < p$, based on simulated data. We set $p = 10$ and generated a population covariance matrix with strong low-rank structure using $\Sigma = AA^\top + 0.1I$, where $A \in \mathbb{R}^{10\times 3}$ is a random matrix. A small sample of size $n = 3$ was drawn from $\mathcal{N}(\mathbf{0}, \Sigma)$ to represent a HD $(n < p)$ scenario.

The left panel shows the *cumulative variance explained* by the PCs. It is evident that the MLE overestimates the contribution of the leading components, while shrinkage estimators such as LW and Stein partially correct this bias.

The right panel scree plot reveals how the MLE inflates the largest eigenvalues and underestimates the smaller ones, resulting in an overly steep drop in the spectrum. This pattern—where the leading eigenvalues are much larger than those of the population, and the trailing ones are disproportionately small—demonstrates classical overdispersion. It falsely suggests that

most variation is explained by only a few components. Shrinkage estimators reduce this inflation–deflation effect by pulling extreme eigenvalues toward the center, yielding a more balanced and realistic eigenvalue distribution. This simulation highlights the importance of accounting for estimation bias in HD settings, particularly when using PCA for interpreting variation or reducing dimensionality.

3. **As n approaches zero relative to p, the Cosine Similarity Error (CSE) of sample PCs increases progressively**

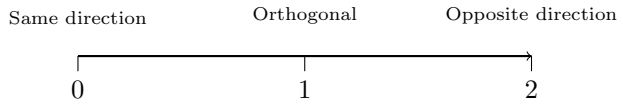
   When n becomes small and p increases, the overall average CSE increases, indicating that the sample PCs are increasingly misaligned with the population PCs. This lack of directional agreement is especially pronounced for the last $p - n + 1$ components, whose CSEs rise disproportionately as n decreases relative to p.

   The CSE is computed using the following formula:

$$\text{CSE}(\text{PC}_i) = 1 - \frac{\hat{\mathbf{PC}}_i \cdot \mathbf{PC}_i}{\|\hat{\mathbf{PC}}_i\| \cdot \|\mathbf{PC}_i\|},$$

   which reflects the angular deviation between the estimated and true eigenvectors. Under this definition (without taking absolute values), CSE values lie in the range $[0, 2]$, where:

   - CSE $= 0$ means the sample and population PCs point in the exact same direction,
   - CSE $= 2$ means they point in exactly opposite directions (180 degrees apart),
   - CSE $= 1$ corresponds to orthogonal directions (90 degrees apart).



   In simpler terms, as the number of observations becomes small relative to the number of variables, the sample PCs increasingly diverge from the true population PCs, especially for the later components. This highlights the instability of principal directions in HD, low-sample-size regimes.

   Table 1.1 shows the average CSEs from simulations where the number of observations n varies from 3 to 25, while the number of dimensions p is fixed at 50. The data were generated from a multivariate normal distribution with mean vector $\mu = \mathbf{0}$ and a randomly generated covariance matrix. Results are averaged over 100 simulation runs.

Table 1.1: CSE of selected cases of $n < p$ when $p = 50$

No: of Dimensions (p)	**No: of Observations (n)**	**Average CSEs**		
		Overall	**n-1**	**p-n+1**
50	3	0.98502	0.59969	1.00107
50	5	0.98007	0.74709	1.00033
50	10	0.97872	0.87527	1.00143
50	20	0.97340	0.93464	0.99715
50	25	0.97644	0.95135	0.99959

As n decreases, the overall CSE increases, indicating weaker alignment between sample and population PCs. For example, the overall CSE is 0.98502 when $n = 3$, and improves slightly to 0.97644 when $n = 25$. Notably, the CSEs for the last $p - n + 1$ PCs are much higher than the overall CSE, especially when $n \ll p$. This gap is most visible for small sample sizes like $n = 3$ and $n = 5$, suggesting that the later PCs are poorly estimated in such settings.

Such findings highlight the critical impact of sample size relative to data dimensionality on representation accuracy. In HD spaces, where p is large, ensuring adequate observations becomes paramount to mitigate disproportionate errors, as observed for the latte PCs. Therefore, this analysis emphasizes balancing sample size and dimensionality to maintain data representation fidelity, particularly in scenarios with limited observations relative to the dimension space.

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

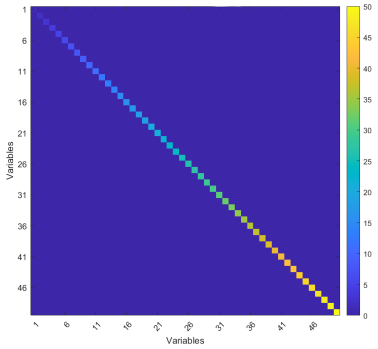
   A covariance matrix is considered sparse when most of its off-diagonal entries are zero. Sparsity helps mitigate the curse of dimensionality by simplifying the structure and reducing estimation error. In our simulation setting, the true population covariance matrix Σ is a diagonal matrix with increasing variance values, specifically $\Sigma = \text{diag}(1, 2, \ldots, 50)$, where the dimensionality is fixed at $p = 50$. This structure ensures sparsity and clear separation among the variable variances.

   Figure 1.4 illustrates how well the sample covariance matrix approximates this true structure under varying sample sizes. As shown, when the number of observations n is relatively large (e.g., $n = 250$) compared to $p = 50$, the sample covariance matrix closely resembles the true diagonal form. However, as n decreases (e.g., $n = 25$ and especially $n = 5$), the sample covariance matrix becomes increasingly noisy and deviates from the sparse structure of Σ.

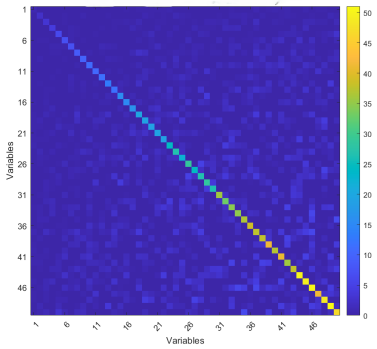
This leads to inflated off-diagonal entries and a distorted eigenstructure, making the sample covariance matrix $\mathbf{S}$ unreliable in high-dimensional, low-sample settings.

These limitations of $\mathbf{S}$ are consistent with findings by Johnstone and Lu [14] and Ledoit and Pèchè [15], who observed that sample eigenvectors can significantly deviate from those of the true population covariance matrix in such regimes. This reinforces the need for a well-conditioned covariance estimator that ensures positive definiteness, sparsity, and improved eigenvalue estimation. Such estimators reduce projection error and enhance the reliability of PCA, especially when $n \ll p$.

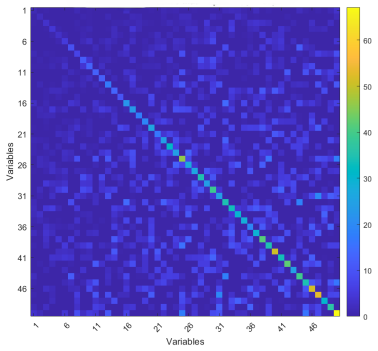
(a) Σ (True Covariance)



(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 $n = 5$

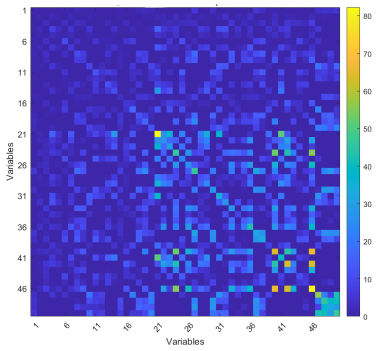


Figure 1.4: Graph of True & Sample Covariance Matrices

1.3 Research Objective(s) and Significance

There are lots of theories already established for sparse and well-conditioned covariance estimation to mitigate the HD covariance estimation problem. However, among them, most of the research papers mainly concern with the broader interest in data science apart from PCA. Therefore, this research aims to address and achieve the following objectives.

1. Propose a novel sample covariance estimation to minimize the overdispersion first PC when $n < p$ settings.
2. Improve the finite sample estimates of PCs when $n < p$ settings in terms of the first PC's low CSE and low overdispersion.
3. Highlight the challenges and issues in selecting the optimal number of PCs when $n < p$ settings.

1.4 Research Questions

- **Effectiveness of Existing Covariance Estimations for PCA in HD Data when $n < p$:** What is the effectiveness of using current well-known covariance estimation methods (i.e., MLE of covariance $\mathbf{S}$ and Ledoit-Wolf estimation of covariance $\hat{\mathbf{\Sigma}}_{\text{LW}}$) for PCA when $n < p$ datasets? How could the direction and magnitude of those PCs be improved?
- **Factors Affecting Classical PCA Inconsistency when $n < p$ Data:** When $n < p$ data sets, what are the factors affecting the inconsistency of classical PCA? How does the following affect the higher CSE and overdispersion of PCs in "$n < p$" HD data sets?
  - number of observations, number of dimensions and p/n
  - type of covariance matrix
  - global outliers
  - level of correlation
- **Proposed Methods for Improved PCA when $n < p$ Data:** How do our proposed covariance estimation methods improve the PCA in terms of both the direction and magnitude when $n < p$ data sets?
- **Proposed Methods for Enhanced Covariance Applications when $n < p$ Data:** How do our proposed covariance estimation methods improve the other covariance applications (such as MANOVA) when $n < p$ data sets?

- **Optimum Number of PCs in HD Data when $n < p$:** When $n < p$ data sets, what are the factors affecting the contradiction of determining the optimum number of PCs retained? How could it be improved in terms of consistency?

Figure 1.5 outlines the various research questions and the corresponding chapters that address them in the thesis. The effectiveness of existing covariance estimations for PCA in HD datasets when $n < p$ is discussed in Chapters 2, 3, 5, and 6. Factors affecting classical PCA inconsistency in such scenarios are analyzed in Chapters 5 and 6. Proposed methods for improved PCA are evaluated in Chapters 4, 5, and 6. The benefits of new covariance estimation methods for applications like MANOVA are assessed in Chapters 4 and 7. Finally, the optimal number of PCs in HD datasets is explored in Chapter 7.

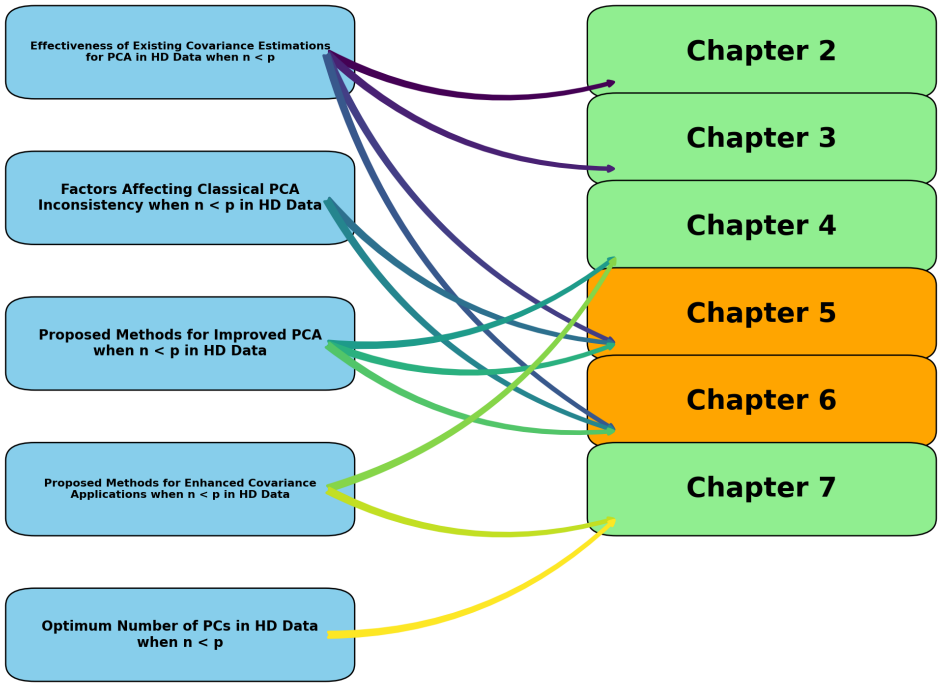


Figure 1.5: Map of Research Questions and Chapters

1.5 Contribution of this Thesis

This research introduces a simple and novel approach to address the challenges experienced in PCA when $n < p$ data. The main issue in this $n < p$ data is inaccuracy in estimating the sample covariance due to the insufficient sample size compared to the number of dimensions. This will lead to ill-conditioned covariance estimation. Based on such ill-conditioned sample covariance, it is unlikely to estimate the more representative finite sample PCs to the population PCs. Therefore, we suggest a different method for estimating the covariance matrix to minimize this issue. Instead of using the usual MLE method, we suggest a different one that works better for this case. To overcome the limitations $n < p$ in the PCA context and enhance the accuracy of PCA outcomes, we propose a simple and novel sample covariance estimation method without increasing the original sample size by collecting new sample observations. Therefore, we use pairwise differences from the original dataset to increase the sample size instead of taking new samples. This research aims to show the effectiveness and applicability of our proposed approach across a diverse range of domains by conducting extensive experiments.

1.6 Organization of this Thesis

This thesis has eight detailed chapters, with each chapter focusing on different aspects of PCA when $n < p$. Chapter 2 explains the fundamental concepts behind PCA and how it deals with challenges in HD settings. Chapter 3 briefly examines both practical and theoretical methods for estimating covariance in similar scenarios. Chapter 4 introduces a new method called Pairwise Differences Covariance (PDC) estimation, which aims to improve the accuracy of PCA estimates in this context. Chapter 5 gives details about how factors like the number of observations, dimensions, types of covariance matrices, correlations, and outliers affect PCA results. Then, Chapter 6 takes the innovative PDC estimation method and applies it to various situations beyond just PCA. Chapter 7 evaluates different statistical tests to figure out the best number of PCs. Finally, Chapter 8 summarizes our findings, pointing out unresolved issues and suggesting areas for future research to advance statistical methods in analyzing HD data.

Chapter 2

Principal Component Analysis (PCA)

2.1 Introduction

PCA explains the variance-covariance structure of a set of dimensions through linear combinations. PCA is the prime linear DR technique used in the higher dimensional space. PCA reduces the dimensionality of the original data set by maximizing the retained variance and minimizes the least square reconstruction error. (see Figure 2.1 [taken from [16]]). In other words, PCA is a technique to reduce the dimensions by taking uncorrelated (orthogonal) linear combinations of the original dimensions that explain the most variance in the data it can (see Figure 2.2 [taken from [17]]). This attempt to reduce dimensionality can be described as a "parsimonious summarization" of the data [18].

The PCA transforms the original p dimensions X_1, X_2, ..., X_p of the observations into its linear combinations called Z_1, Z_2, ..., Z_p that are uncorrelated with each other. These transformed dimensions Z_1, Z_2, ..., Z_p are called PCs. PCs' orthogonality is useful as it indicates that they assess the clearly different "dimensions" or "aspects" of the information. This helps us to make clear interpretations.

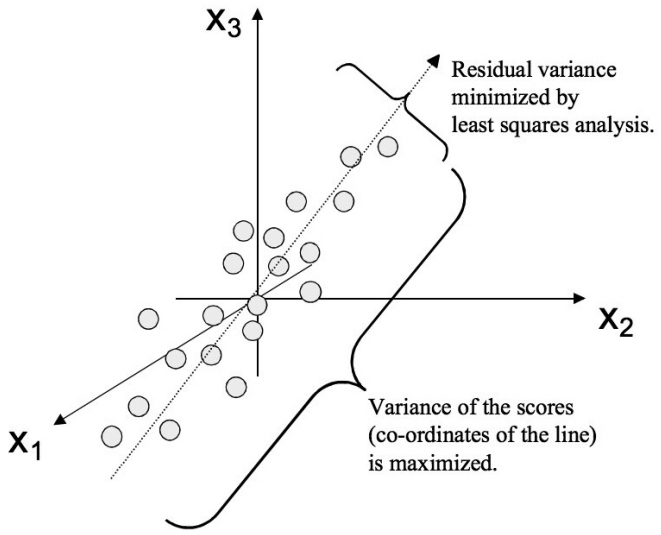


Figure 2.1: The Principal Idea of PCA

The indices are derived in such a way that Z_1 is the direction that explains the largest fraction of the sample variance (of the p orthogonal direction Z_1, ..., Z_p), Z_2 is the direction that explains the largest fraction of the sample variance (of the p orthogonal direction Z_2, ..., Z_p), and so on (see Eqn. 2.1). That is,

$$Var(Z_1) \geq Var(Z_2) \geq ... \geq Var(Z_p) \tag{2.1}$$

In other words, the first PC is associated with the largest variance in the data, and the second PC explains the maximum amount of the remaining variance in the data, subject to being uncorrelated with the first PC, and so on. When performing PCA, we can neglect the PCs, which contribute a much smaller amount to the total variance of the data set. Thus, a few PCs can describe the total variation of the data set.

The first PC is considered the most important among these few PCs since it describes the largest fraction of the sample variability, followed by the second PC, and so on. Ultimately, variation of p original dimensions is accounted for by a few Z dimensions (PCs) (see Figure 2.2). Thus, the achievement here is that if we have a multivariate normal data set with p dimensions, we can project data down to lower-dimensional transformed multivariate normal dimensions (PCs) that are orthogonal to each other. However, it is important to note that, the best results from PCA can be obtained when the original dimensions are highly correlated, positively or negatively [19],[18]. In such a case, 20 or 30 original dimensions can be adequately represented by 2 or 3 PCs [20]. This is

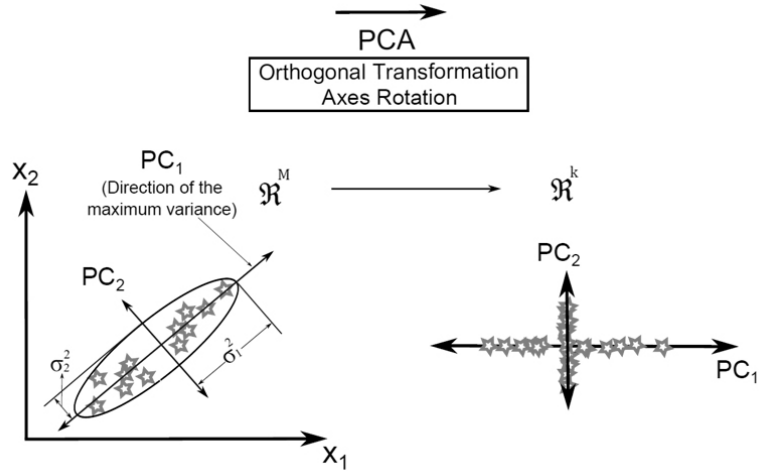


Figure 2.2: The direction of the PCA

usual since when data is collected under many dimensions, there will be a great deal of redundancy with several dimensions measuring more or less the same thing.

2.1.1 Sample PCs

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 dimensions for n observations. The first sample PC is then the linear combinations of the dimensions $\mathrm{X_1}$, $\mathrm{X_2}$, ..., $\mathrm{X_p}$,

$$PC_1 = Z_1 = \boldsymbol{e}_1{}^{\mathrm{T}}\boldsymbol{X} = e_{11}X_1 + e_{12}X_2 + ... + e_{1\mathrm{p}}X_\mathrm{p} \tag{2.2}$$

and its variance is maximum for the individuals subject to $\mathbf{e}_1{}^{\mathrm{T}}\mathbf{e}_1 = 1$.

$$e_{11}{}^2 + e_{12}{}^2 + ... + e_{1\mathrm{p}}{}^2 = 1 \tag{2.3}$$

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

$$PC_2 = Z_2 = \boldsymbol{e}_2{}^{\mathrm{T}}\boldsymbol{X} = e_{21}X_1 + e_{22}X_2 + ... + e_{2\mathrm{p}}X_\mathrm{p} \tag{2.4}$$

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

$$e_{21}{}^2 + e_{22}{}^2 + ... + e_{2\mathrm{p}}{}^2 = 1 \tag{2.5}$$

with additional condition that Z_2 is uncorrelated with Z_1. Similarly the third PC,

$$PC_3 = Z_3 = \boldsymbol{e}_3{}^{\mathrm{T}}\boldsymbol{X} = e_{31}X_1 + e_{32}X_2 + ... + e_{3\mathrm{p}}X_{\mathrm{p}} \tag{2.6}$$

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

$$e_{31}{}^2 + e_{32}{}^2 + ... + e_{3\mathrm{p}}{}^2 = 1 \tag{2.7}$$

and also that Z_3 is uncorrelated with both Z_1 and Z_2. Other PCs can be defined in a similar way, and this can be continued until p PCs are defined.

Maximizing a function subject to constraints is usually achieved using a mathematical method called Lagrange Multiplier. Under the optimum condition, the variances of the PCs are the eigenvalues of the sample variance-covariance matrix [21], $\hat{\mathbf{\Sigma}}$, of p the original dimensions and $\hat{\mathbf{\Sigma}}$, which is of the form,

$$\hat{\mathbf{\Sigma}} = \begin{bmatrix} S_{11} & S_{12} & \cdots & S_{1p} \\ S_{21} & S_{22} & \cdots & S_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ S_{p1} & S_{p2} & \cdots & S_{pp} \end{bmatrix}$$

Here, S_{ij} represents the sample covariance between the ith and jth variables. The diagonal elements S_{ii} are the sample variances of each variable.

Since $\hat{\mathbf{\Sigma}}$ is a symmetric $p \times p$ matrix, it has p real eigenvalue-eigenvector pairs:

$$(\hat{\lambda}_1, \hat{\boldsymbol{e}}_1), (\hat{\lambda}_2, \hat{\boldsymbol{e}}_2), \ldots, (\hat{\lambda}_p, \hat{\boldsymbol{e}}_p)$$

If the eigenvalues are ordered as:

$$\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_p \geq 0 \tag{2.8}$$

then $\hat{\lambda}_1$ corresponds to the first PC (i.e., $\mathrm{Var}(Z_1) = \hat{\lambda}_1$), $\hat{\lambda}_2$ to the second PC, and so on. The first eigenvector $\hat{\boldsymbol{e}}_1 = (e_{11}, e_{12}, \ldots, e_{1p})$ defines the direction of the first PC. Each subsequent PC is defined similarly by its corresponding eigenvector.

An important property of the eigenvalues is that their sum equals the trace of $\hat{\mathbf{\Sigma}}$, i.e., the sum of the variances of the individual variables:

$$\text{Total Sample Variance} = \sum_{i=1}^{p} S_{ii} = \sum_{i=1}^{p} \hat{\lambda}_i \tag{2.9}$$

This proves the PCs account for all the original data variation. When data is collected under many dimensions, different dimensions are often measured in different units. Because of the unit of measurement, sometimes variance becomes large. Standardized dimensions are often used for the analysis to avoid certain dimensions having undue influence on the PCs. When standardized dimensions are used,

$$\text{Total Sample Variance} = \sum_{i=1}^{p} S_{ii} = \sum_{i=1}^{p} \hat{\lambda}_i = p \tag{2.10}$$

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

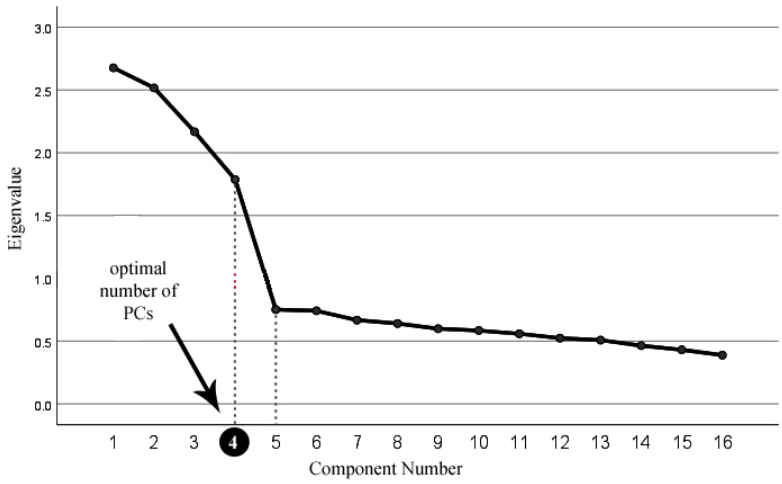


Figure 2.3: Scree Plot Method

There are various statistical tests to determine the optimum number of PCs, such as the Kaiser-Guttman criterion [22, 23], Cattell's Scree Test (1966) [24] and Percent of Cumulative Variance.

1. The Kaiser-Guttman number of components to retain ($n_{\text{K-G}}$) is computed by $n_{\text{K-G}}$ = Count ($\lambda_{\text{i}} > 1$), where, λ_{i} is the i^{th} eigenvector.

2. According to the scree plot, the components to the left of the "elbow" point in the graph are typically retained as the optimal number of PCs (see Figure 2.3 [taken from [25]]).

3. Percent of cumulative variance method retains the number of components that explain 70% or 80% (note: it's a thumb rule) of the total variance.

However, most of the researchers commonly used non-statistical strategies to determine the optimum number of PCs [26].

2.1.2 Computations of PCA

We use two primary methodologies to calculate the PCs, such as the Singular Value Decomposition (SVD) of the sample covariance matrix, and produce PCs by solving the covariance matrix. These methodologies represent two distinct techniques for extracting PCs.

1. **SVD Method**

   Let $\mathbf{X}$ be a centered data matrix of size $n \times p$, where n is the number of observations and p is the number of variables. The sample covariance matrix is defined as:

   $$\boldsymbol{S} = \frac{1}{n-1}\boldsymbol{X}^\top \boldsymbol{X}, \quad \boldsymbol{S} \in \mathbb{R}^{p\times p}.$$

   To study the structure of $\mathbf{S}$, we perform SVD on $\mathbf{X}$:

   $$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top,$$

   where:

   - $\boldsymbol{U} \in \mathbb{R}^{n\times n}$ contains the left singular vectors, with $\boldsymbol{U}^\top\boldsymbol{U} = \boldsymbol{I}_n$,
   - $\boldsymbol{V} \in \mathbb{R}^{p\times p}$ contains the right singular vectors, with $\boldsymbol{V}^\top\boldsymbol{V} = \boldsymbol{I}_p$,
   - $\boldsymbol{\Sigma} \in \mathbb{R}^{n\times p}$ is a diagonal matrix with singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$ on the diagonal, where $r = \text{rank}(\boldsymbol{X})$.

   Derivation of $\mathbf{V}$:

   From the SVD, we have:

   $$\boldsymbol{X}^\top\boldsymbol{X} = (\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top)^\top(\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top) = \boldsymbol{V}\boldsymbol{\Sigma}^\top\boldsymbol{U}^\top\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top.$$

   Since $\boldsymbol{U}^\top\boldsymbol{U} = \boldsymbol{I}_n$, it simplifies to:

   $$\boldsymbol{X}^\top\boldsymbol{X} = \boldsymbol{V}(\boldsymbol{\Sigma}^\top\boldsymbol{\Sigma})\boldsymbol{V}^\top.$$

Then:

$$\boldsymbol{S} = \frac{1}{n-1}\boldsymbol{X}^\top \boldsymbol{X} = \boldsymbol{V}\left(\frac{1}{n-1}\boldsymbol{\Sigma}^\top\boldsymbol{\Sigma}\right)\boldsymbol{V}^\top.$$

This shows that:

- Columns of $\mathbf{V}$ are eigenvectors of $\mathbf{S}$,
- The eigenvalues of $\mathbf{S}$ are given by $\lambda_i = \sigma_i^2/(n-1)$.

Derivation of $\mathbf{U}$:

Similarly, we can derive:

$$\boldsymbol{X}\boldsymbol{X}^\top = (\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top)(\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top)^\top = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top\boldsymbol{V}\boldsymbol{\Sigma}^\top\boldsymbol{U}^\top = \boldsymbol{U}(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)\boldsymbol{U}^\top,$$

so:

- Columns of $\mathbf{U}$ are eigenvectors of $\boldsymbol{X}\boldsymbol{X}^\top$,
- The eigenvalues are also σ_i^2.

Matrix Form Representation of $\mathbf{S}$:

$$\boldsymbol{S} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^\top, \quad \text{where } \boldsymbol{\Lambda} = \frac{1}{n-1}\boldsymbol{\Sigma}^\top\boldsymbol{\Sigma} = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_r).$$

$$\underbrace{\begin{bmatrix} s_{11} & s_{12} & \ldots & s_{1p} \\ s_{21} & s_{22} & \ldots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \ldots & s_{pp} \end{bmatrix}}_{\boldsymbol{S}\in\mathbb{R}^{p\times p}} = \underbrace{\begin{bmatrix} v_{11} & v_{12} & \ldots & v_{1p} \\ v_{21} & v_{22} & \ldots & v_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ v_{p1} & v_{p2} & \ldots & v_{pp} \end{bmatrix}}_{\boldsymbol{V}\in\mathbb{R}^{p\times p}} \underbrace{\begin{bmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \lambda_p \end{bmatrix}}_{\boldsymbol{\Lambda}\in\mathbb{R}^{p\times p}} \underbrace{\begin{bmatrix} v_{11} & v_{21} & \ldots & v_{p1} \\ v_{12} & v_{22} & \ldots & v_{p2} \\ \vdots & \vdots & \ddots & \vdots \\ v_{1p} & v_{2p} & \ldots & v_{pp} \end{bmatrix}}_{\boldsymbol{V}^\top\in\mathbb{R}^{p\times p}}.$$

In practical computations, it is often more accurate and stable to perform SVD on the original data matrix $\boldsymbol{X}$, rather than directly computing the eigenvalues and eigenvectors of the sample covariance matrix $\boldsymbol{S} = \frac{1}{n-1}\boldsymbol{X}^\top\boldsymbol{X}$. Instead of directly computing the eigen-decomposition of $\boldsymbol{S}$, we use the SVD of $\boldsymbol{X}$, computed via the Golub–Kahan–Lanczos bidiagonalization. This method avoids explicitly forming $\boldsymbol{S}$ and provides stable and efficient computation.

Step 1: Start with the matrix $\boldsymbol{X}$

We aim to compute the SVD:

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top,$$

where:

- $\boldsymbol{U} \in \mathbb{R}^{n\times r}$: left singular vectors (orthonormal),
- $\boldsymbol{V} \in \mathbb{R}^{p\times r}$: right singular vectors (orthonormal),
- $\boldsymbol{\Sigma} \in \mathbb{R}^{r\times r}$: diagonal matrix of singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq 0$.

Step 2: Initialize the bidiagonalization process

- Choose an initial vector $\boldsymbol{v}_1 \in \mathbb{R}^p$ such that $\|\boldsymbol{v}_1\| = 1$.
- Set $\beta_0 = 0$ and $\boldsymbol{u}_0 = \boldsymbol{0} \in \mathbb{R}^n$.

Step 3: Iterate to build orthonormal vectors

For $j = 1, 2, \ldots, k$, repeat:

$$\begin{aligned}\boldsymbol{u}_j &= \boldsymbol{X}\boldsymbol{v}_j - \beta_{j-1}\boldsymbol{u}_{j-1}, \quad \alpha_j = \|\boldsymbol{u}_j\|, \quad \boldsymbol{u}_j \leftarrow \boldsymbol{u}_j/\alpha_j,\\ \boldsymbol{v}_{j+1} &= \boldsymbol{X}^\top\boldsymbol{u}_j - \alpha_j\boldsymbol{v}_j, \quad \beta_j = \|\boldsymbol{v}_{j+1}\|, \quad \boldsymbol{v}_{j+1} \leftarrow \boldsymbol{v}_{j+1}/\beta_j.\end{aligned}$$

This produces:

- $\boldsymbol{U}_k = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k] \in \mathbb{R}^{n\times k}$,
- $\boldsymbol{V}_k = [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k] \in \mathbb{R}^{p\times k}$.

Step 4: Form the bidiagonal matrix $\boldsymbol{B}_k \in \mathbb{R}^{k\times k}$

$$\boldsymbol{B}_k = \begin{bmatrix} \alpha_1 & \beta_1 & & \\ & \alpha_2 & \beta_2 & \\ & & \ddots & \ddots \\ & & & \alpha_k \end{bmatrix}$$

Step 5: Compute the SVD of $\boldsymbol{B}_k$

$$\boldsymbol{B}_k = \boldsymbol{P}_k\boldsymbol{\Sigma}_k\boldsymbol{Q}_k^\top,$$

where:

- $\boldsymbol{\Sigma}_k$: diagonal matrix of approximate singular values,
- $\boldsymbol{P}_k \in \mathbb{R}^{k\times k}$: orthogonal matrix of left singular vectors of $\boldsymbol{B}_k$,
- $\boldsymbol{Q}_k \in \mathbb{R}^{k\times k}$: orthogonal matrix of right singular vectors of $\boldsymbol{B}_k$.

Step 6: Construct the approximate SVD of $\boldsymbol{X}$

Define:

$$\boldsymbol{U} \approx \boldsymbol{U}_k\boldsymbol{P}_k, \quad \boldsymbol{V} \approx \boldsymbol{V}_k\boldsymbol{Q}_k, \quad \boldsymbol{\Sigma} \approx \boldsymbol{\Sigma}_k.$$

Then:

$$\boldsymbol{X} \approx \boldsymbol{U}_k\boldsymbol{P}_k\boldsymbol{\Sigma}_k\boldsymbol{Q}_k^\top\boldsymbol{V}_k^\top = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top.$$

Step 7: Compute the sample covariance matrix from SVD

$$\boldsymbol{S} = \frac{1}{n-1}\boldsymbol{X}^\top\boldsymbol{X} = \boldsymbol{V}\left(\frac{1}{n-1}\boldsymbol{\Sigma}^\top\boldsymbol{\Sigma}\right)\boldsymbol{V}^\top.$$

Thus:

- The eigenvalues of $\boldsymbol{S}$ are $\lambda_i = \sigma_i^2/(n-1)$,
- The eigenvectors of $\boldsymbol{S}$ are the columns of $\boldsymbol{V}$.

Convergence Criteria:

In iterative algorithms for computing the SVD (such as the Golub–Kahan method), the procedure generates a sequence of approximations to the true decomposition. The algorithm is considered to have converged when the difference between two successive approximations becomes very small.

Mathematically, convergence is achieved when:

$$\|\boldsymbol{X}_k - \boldsymbol{X}_{k-1}\|_F < \varepsilon,$$

where:

- $\boldsymbol{X}_k = \boldsymbol{U}_k \boldsymbol{\Sigma}_k \boldsymbol{V}_k^\top$ is the current approximation of the matrix $\boldsymbol{X}$,
- $\boldsymbol{X}_{k-1}$ is the previous approximation,
- $\|\cdot\|_F$ denotes the Frobenius norm, which measures the total squared difference between matrices,
- $\varepsilon > 0$ is a small tolerance value (e.g., 10^{-6}) that determines how close the result must be to stop iterating.

In addition to this numerical tolerance, two other properties are often checked:

- **Orthogonality:** Ensure that the computed singular vectors are nearly orthonormal (e.g., $\boldsymbol{U}^\top \boldsymbol{U} \approx \boldsymbol{I}$),
- **Reconstruction accuracy:** Verify that the product $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$ closely approximates the original matrix $\boldsymbol{X}$.

These checks ensure both numerical accuracy and stability of the computed SVD.

2. **Covariance Matrix Method**

MLE of covariance is one of the foundational estimation methods in statistics. In 1922, R. A. Fisher introduced the MLE framework [27].

Let $x_1, x_2, \ldots, x_n$ be a set of observations from a random variable X, with joint probability density function $f(x_1, x_2, \ldots, x_n; \theta)$. When considered as a function of θ, this is called the likelihood function:

$$L(\theta; x_1, x_2, \ldots, x_n)$$

If $\hat{\theta}$ is the value that maximizes the likelihood, it is called the MLE of θ. To illustrate, consider the case where $X \sim \mathcal{N}(\mu, \sigma_x^2)$.

Step I: Form the likelihood function.

$$L = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_x^2}} \exp\left(-\frac{1}{2}\frac{(X_i - \mu)^2}{\sigma_x^2}\right) = \left(\frac{1}{\sqrt{2\pi\sigma_x^2}}\right)^n \exp\left(-\frac{1}{2\sigma_x^2}\sum_{i=1}^{n}(X_i - \mu)^2\right)$$

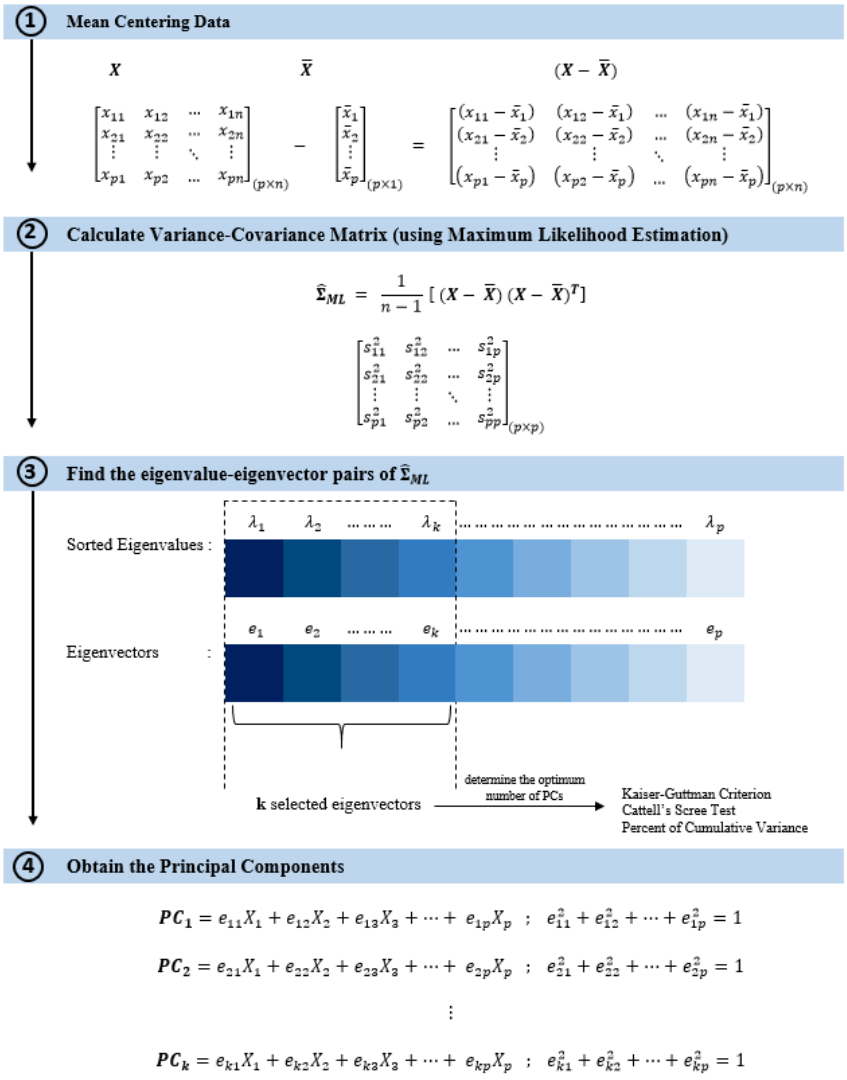
Taking the natural log of the likelihood:

$$\log L = -n \log \sigma_x - n \log \sqrt{2\pi} - \frac{1}{2\sigma_x^2}\sum_{i=1}^{n}(X_i - \mu)^2$$

Step II: Take partial derivatives with respect to μ and σ_x:

$$\frac{\partial \log L}{\partial \mu} = \frac{1}{\sigma_x^2}\sum(X_i - \mu) = 0 \quad \Rightarrow \quad \hat{\mu} = \frac{1}{n}\sum X_i$$

$$\frac{\partial \log L}{\partial \sigma_x} = -\frac{n}{\sigma_x} + \frac{1}{\sigma_x^3}\sum(X_i - \mu)^2 = 0 \quad \Rightarrow \quad \hat{\sigma}_x^2 = \frac{1}{n}\sum(X_i - \mu)^2$$



$$PC_1 = e_{11}X_1 + e_{12}X_2 + e_{13}X_3 + \cdots + e_{1p}X_p \;;\; e_{11}^2 + e_{12}^2 + \cdots + e_{1p}^2 = 1$$

$$PC_2 = e_{21}X_1 + e_{22}X_2 + e_{23}X_3 + \cdots + e_{2p}X_p \;;\; e_{21}^2 + e_{22}^2 + \cdots + e_{2p}^2 = 1$$

$$\vdots$$

$$PC_k = e_{k1}X_1 + e_{k2}X_2 + e_{k3}X_3 + \cdots + e_{kp}X_p \;;\; e_{k1}^2 + e_{k2}^2 + \cdots + e_{kp}^2 = 1$$

Figure 2.4: Geometric interpretation of diagonalizing the covariance matrix in 2D

Step III: Final MLEs:

- **MLE of the mean:** $\hat{\mu} = \frac{1}{n}\sum X_i$
- **MLE of the variance (known μ):** $\hat{\sigma}_x^2 = \frac{1}{n}\sum(X_i - \mu)^2$

Note that this MLE variance uses a denominator of n, and is therefore *biased* for finite n. In contrast, the *unbiased sample variance* uses a denominator of $n-1$:

$$\hat{\sigma}^2_{\text{sample}} = \frac{1}{n-1}\sum(X_i - \bar{X})^2$$

Thus, in this context:

$$\hat{\sigma}^2_{\text{MLE}} = \frac{1}{n}\sum(X_i - \bar{X})^2 \neq \hat{\sigma}^2_{\text{sample}} = \frac{1}{n-1}\sum(X_i - \bar{X})^2$$

In this method, PCs are obtained from the eigenvectors of the covariance matrix of the data matrix $\mathbf{X}$ (see Figure 2.4). Figure 2.4 illustrates the geometric interpretation of PCA via diagonalization of the covariance matrix in 2D.

When the sample covariance matrix is well-conditioned, both methods yield equivalent results for PCA. However, they differ in computational strategy and stability. SVD is often preferred in high-dimensional settings where $n < p$ because it avoids directly computing the potentially singular sample covariance matrix. Instead, it operates on the centered data matrix and provides improved numerical stability, particularly when the number of observations is small relative to the number of variables.

In this research, we primarily use the covariance-based method (with scaling factor $1/(n-1)$) to extract PCs. This is consistent with the formulation and application of the proposed Pairwise Differences Covariance (PDC) estimator, which is constructed to estimate the covariance structure directly. Since the PDC estimator does not involve matrix factorization through SVD, all PCA analyses in this thesis are based on eigendecomposition of the estimated covariance matrix. This choice ensures consistency across comparisons and aligns with standard practice when evaluating the quality of covariance estimators.

In summary, PCA reduces the dimensionality of high-dimensional data by transforming the input variables into uncorrelated PCs, preserving as much variation as possible, and revealing the underlying data structure.

2.2 PCA when $n < p$

The general PCA methods have significant challenges in $n < p$ data settings. Traditional MLE covariance estimation methods, upon which PCA relies, need help to produce reliable outcomes when faced with limited observations to estimate covariance accurately. This imbalance, often termed the "small n, large p" problem, introduces instability into the computations of PCs. When the ratio of p/n is high, the sample covariance estimation may suffer from singularity or ill-conditioning. This issue badly affects the reliability of finite sample estimates of eigenvalues and eigenvectors of the sample covariance matrix. Because of this issue, the finite sample estimates of PCA do not adequately represent the population parameters.

In this section, we present several results and figures based on controlled synthetic data simulations. These examples are specifically designed to illustrate how classical PCA behaves under various high-dimensional settings where the number of observations n is smaller than the number of dimensions p ($n < p$), as well as selected cases where $n > p$.

We simulate data from a multivariate normal distribution $\boldsymbol{X} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$, where the mean vector is zero and the covariance matrix $\boldsymbol{\Sigma}$ is generated using either random entries or low-rank structures with predefined eigenvalue patterns. In most simulations, we fix $p = 10$, except in Table 2.5, where we also consider $p = 20$ and $p = 50$. The sample size n is varied from small values (e.g., $n = 3, 5, 10$) up to $n = 500$, covering both high-dimensional ($n < p$) and well-conditioned ($n > p$) scenarios. Each setting is repeated over 100 independent runs to ensure reliable estimation of performance measures.

These simulated settings allow us to demonstrate typical issues encountered in practice, such as eigenvalue overdispersion, ill-conditioning, and instability in estimating the PCs from sample covariance matrices. Although these are not general theoretical results, they effectively illustrate the challenges that arise in high-dimensional data analysis.

As we underscored before, the weaknesses of typical sample covariance estimation in $n < p$ data settings create numerous concerns in a wide range of statistical applications, including PCA. Therefore, identifying the factors affecting the instability of typical sample covariance estimation in $n < p$ data settings and how it affects the essential. Sample PCs are derived from the eigenvectors of the data's sample covariance matrix $\mathbf{S}$, and the corresponding eigenvalues give the amount of variance each component captures. In $n < p$ data settings, the sample covariance estimates produce poorly estimated eigenvalues and eigenvectors. Because of this poor estimation of the sample covariance matrix, the first n-1 estimated eigenvalues are too large, and the last (p-n+1) eigenvalues are too small in the $n < p$ data settings.

When n is much lower than p, the ratio of the largest eigenvalue to the smallest eigenvalue becomes higher. As per the theories of linear algebra, the ratio of the largest and smallest eigenvalue is identified as the condition number of a matrix. In addition to that, the value of k ($k = \lambda_{\max}$ / $\lambda_{\min}$) of the sample covariance matrix tells us whether the sample covariance matrix is well-conditioned or ill-conditioned. If k of the typical sample covariance matrix is higher, the particular sample covariance matrix is ill-conditioned [28], [29]. Otherwise, it is well-conditioned.

Table 2.1: Instability of Covariance Estimates

No: of Dimensions (p)	**No: of Observations (n)**	**$k = \lambda_{\max}$ / $\lambda_{\min}$**
10	500	1925.9
10	250	1969.4
10	100	2128.0
10	50	2439.9
10	25	3444.6
10	10	∞[1]
10	5	∞[1]
10	3	∞[1]

[1] Theoretical condition number is infinite when $n < p$ due to zero eigenvalues.

An ill-conditioned covariance matrix describes a situation where the sample covariance matrix is nearly singular or almost non-invertible. This error may happen in $n < p$ data settings because the sample covariance has an extremely large ratio of its largest and smallest eigenvalues. It can be seen in Table 2.1. The increasing values of k when n goes down compared to p highlight the severity of the problem. If $\mathbf{X}$ is an $n \times p$ random data matrix and the covariance matrix is full rank, it can produce p-dimensional linearly independent combinations (PCs) from the original data. The maximum number of linearly independent rows or columns in the sample covariance matrix is called the rank of the sample covariance matrix. If $n > p$, rank equals p and if $n < p$ rank equals n-1. That means $n < p$ data sets cannot produce more than n-1 linearly independent combinations (PCs) from the original data. This is called rank deficiency of the sample covariance matrix [30], [5]. Due to this drawback of the sample covariance matrix, we have to interpret the sample variance-covariance structure using n-1 PCs instead of p PCs. However, n-1 PCs cannot accurately explain the whole variance-covariance structure of the original data. Therefore, applying the DR technique to such situations makes unstable and unreliable PC estimates. When n goes much lower compared to p, this issue worsens finite sample estimates for the PCs.

Table 2.2: Rank Deficiency or Singularity of Covariance Estimates

	No: of Dimensions (p) = 10							
Singular Values ID	**No: of Observations (n)**							
	2	**3**	**4**	**5**	**6**	**7**	**8**	**9**
1	**97.3**	**93.1**	**19.7**	**32.4**	**59.1**	**15.6**	**22.6**	**47.6**
2	0	**2.3**	**4.1**	**5.4**	**2.6**	**3.2**	**2.6**	**3.7**
3	0	0	**1.7**	**2.6**	**1.1**	**1.1**	**1.1**	**2.5**
4	0	0	0	**0.4**	**0.4**	**0.9**	**0.8**	**0.9**
5	0	0	0	0	**0.1**	**0.2**	**0.3**	**0.7**
6	0	0	0	0	0	**0.07**	**0.1**	**0.3**
7	0	0	0	0	0	0	**0.008**	**0.08**
8	0	0	0	0	0	0	0	**0.006**
9	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0
Rank	1	2	3	4	5	6	7	8

Table 2.2 clearly shows the effect of $n < p$ ($p = 10$ and $n < p$) in sample covariance estimation using eigenvalue distribution and the rank of the sample covariance matrix. Table 2.2 shows that when n goes lower compared to p, the number of non-zero eigenvalues also gets lower. In $n < p$ data settings, we will observe n-1 non-zero eigenvalues and p-n+1 zero eigenvalues. For example, when $n = 9$ has eight non-zero eigenvalues, and $n = 2$ has only one non-zero eigenvalue. That means that when $n \ll p$, the rank of the sample covariance matrix is away from the full rank situation. For example, in the $n = 2$ scenarios, one PC may not capture a substantial proportion of the total variance. It does not need any DR techniques to reduce the dimensionality. That means it overestimates the sample PCs, leading to unreliable finite sample estimates for PCs. The difference between the average values of the eigenvalues derived from the sample covariance matrix and the true eigenvalues of the population covariance matrix is called the mean biases of eigenvalues of the sample covariance matrix [31]. As shown in Figure 2.5, in $n < p$ data settings, when n goes down while maintaining a fixed number of p, the mean biases of eigenvalues of the sample covariance matrix tend to increase. Due to this bias, the estimated sample eigenvalues may significantly deviate from the true eigenvalues, leading to misinterpreted finite sample PCs.

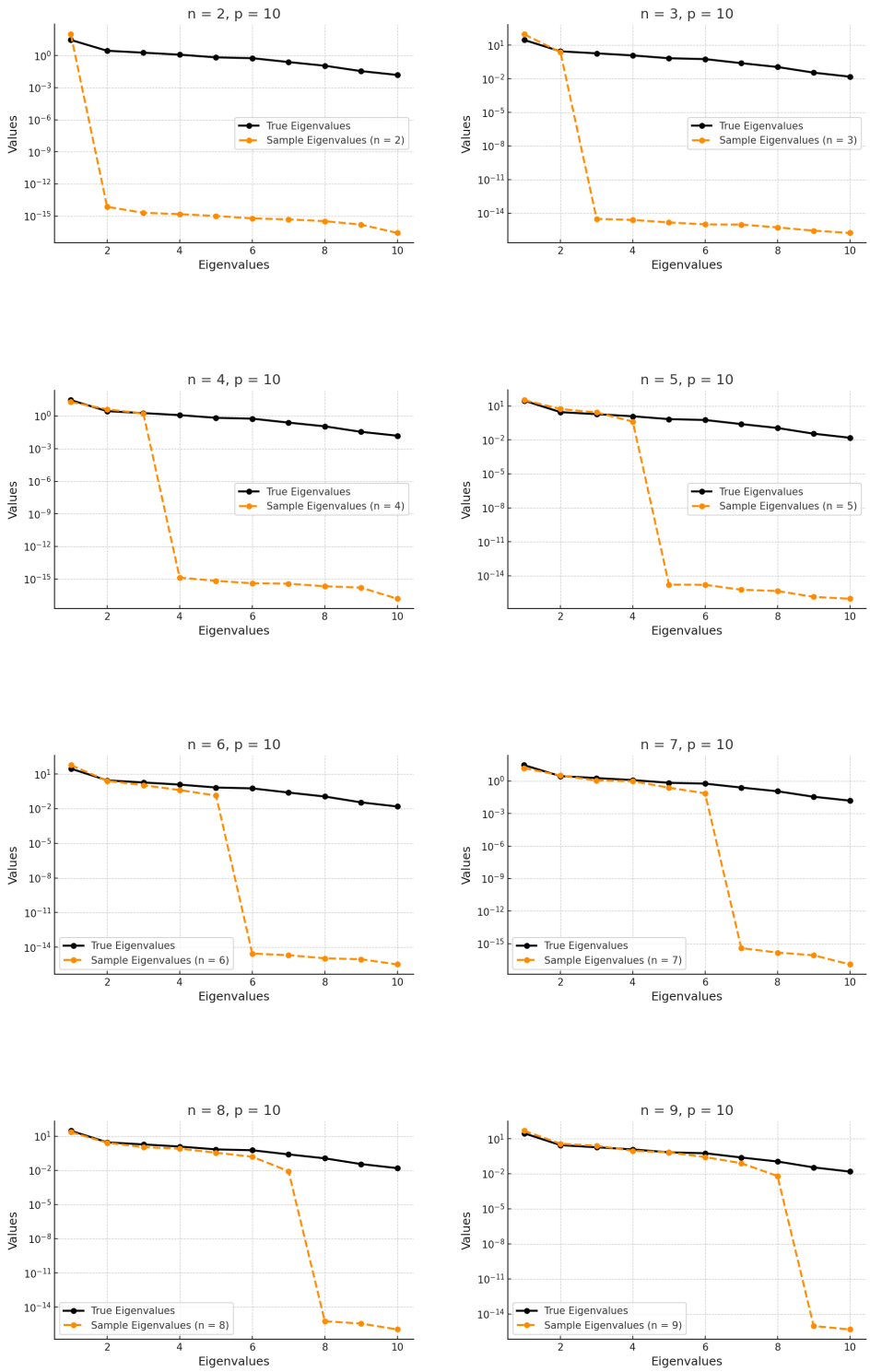


Figure 2.5: Bias in Eigenvalues of Covariance Estimates

Table 2.3: Mean and SD of Eigenvalue Estimates (Simulated Data)

No: of Dimensions (p)	**No: of Observations (n)**	**mean**	**SD**
Population		**3.6611**	**8.9657**
10	2	3.6013	11.3882
10	3	3.6076	10.2768
10	4	3.4492	9.3512
10	5	3.6271	9.6274
10	6	3.6322	9.5302
10	7	3.6154	9.3258
10	8	3.6798	9.5084
10	9	3.5301	8.9357

In PCA, the eigenvalues tell us the variance each PC captured. The average of these eigenvalues gives us a sense of the overall variance in the data. A higher mean indicates that the dataset represents much variance captured from the original data; however, if we experienced a lower number of non-zero eigenvalues with a comparatively high average of eigenvalues, that indicates the overdispersion of the sample PCs [32]. A high standard deviation of the average eigenvalues indicates that the first few PCs capture much more variance than others. However, in $n < p$, data settings give us overdispersed eigenvalues for the first few PCs. Therefore, in $n < p$ data settings, we do not have enough evidence to describe the variance-covariance structure of the original data using fake sample PCs.

Table 2.3 shows that as n goes down compared to p, the mean and standard deviation of the average sample eigenvalues are comparatively higher. This implies the effect of overdispersion of the first few PCs and the loss of information from original data.

If there is a large number of too-small eigenvalues in the sample covariance matrix and a sharp drop-off in the scree plot, it implies the existence of overfitting. However, $n < p$ data settings always produce the last p-n+1 eigenvalues with nearly zero (too small) magnitude. Therefore, in $n < p$ data settings, the first n-1 PCs are overdispersed due to overestimated eigenvalues. We generally use the eigenvalue distribution to compute the percentage of explained variance for each PC. However, due to the first n-1 eigenvalues being overestimated, the calculated percentage of explained variance of the first n-1 PCs may overdispersed.

Table 2.4 shows a clear pattern of each component's percentage of explained variance under each $n < p$ setting for $n = 10$. According to that, we can see that when n goes down compared to p, the overestimation of the first n-1 eigenvalues worsens and badly affects the interpretations of PCs. It also captures noise or irrelevant information [33]. This suggests that a few PCs should be retained because of overdispersed PCs with an artificial percentage of explained variance.

Table 2.4: Variance Explained by Each Principal Component

	No: of Dimensions (p) = 10								
ID	**No: of Observations (n)**								**Population Eigenvalues**
	2	**3**	**4**	**5**	**6**	**7**	**8**	**9**	
1	100.00	90.64	86.60	84.98	84.16	82.85	83.04	81.50	**79.34**
2	**0.00**	9.36	10.33	10.19	9.73	10.05	9.58	9.78	**7.71**
3	**0.00**	**0.00**	3.08	3.72	4.04	4.33	4.14	4.69	**5.09**
4	**0.00**	**0.00**	**0.00**	1.11	1.60	1.84	1.99	2.33	**3.32**
5	**0.00**	**0.00**	**0.00**	**0.00**	0.47	0.71	0.88	1.09	**1.86**
6	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	0.20	0.30	0.44	**1.56**
7	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	0.07	0.13	**0.68**
8	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	0.03	**0.31**
9	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.10**
10	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.04**

The MSE of the sample covariance matrix describes the accuracy and reliability of the sample covariance as an estimator for the actual covariance matrix. A lower MSE implies that the sample covariance matrix is a more accurate and reliable estimator for the true covariance, and a higher MSE implies that the sample covariance matrix is not an accurate and reliable estimator for the true covariance [34], [35]. However, in $n < p$ data settings, when n goes down compared to p, the MSE of the sample covariance matrix gets higher.

Table 2.5 presents the matrix MSE of the sample covariance estimator across various HD settings with increasing p and decreasing n. These values were computed using simulated multivariate normal data and illustrate how poorly the sample covariance matrix generalizes when $n < p$. The results clearly show that as the number of observations n decreases relative to the number of dimensions p, the MSE increases significantly.

For example, when $p = 10$, reducing n from 10 to 3 increases the MSE from 412.52 to 1303.3. This trend is even more pronounced for higher dimensions—for instance, at $p = 50$, the MSE rises from 6.04×10^5 when $n = 10$ to over 1.53×10^6 when $n = 3$. These results demonstrate the poor generalization and instability of the sample covariance matrix in $n < p$ scenarios. The large MSE values indicate that the estimated covariance matrices can deviate substantially from the true population structure, which negatively affects downstream tasks such as PCA.

Table 2.5: Poor Generalization of Covariance Estimates

No: of Dimensions (p)	**No: of Observations (n)**	**MSE(S)**
10	10	412.52
10	5	774
10	3	1303.3
20	10	12232
20	5	21470
20	3	29990
50	10	6.04×10^5
50	5	1.03×10^6
50	3	1.53×10^6 s

The CSE is one of the popular ways to present the inefficiency of the sample covariance matrix by their directions of eigenvectors. In PCA, we use the CSE to identify the cosine of the angle between sample PCs and population PCs in a p-dimensional space. It measures the directions of sample and population PCs and how well sample PCs align with the population PCs. If this CSE of sample PCs is higher, the sample and population PCs are pointed in orthogonal directions. Also, the CSE of sample PCs is lower, which means that the sample and population PCs are pointed in the same direction.

In high-dimensional settings where $n < p$, the sample covariance matrix becomes rank-deficient, and only the first $n - 1$ PCs are informed by the data. The remaining $p - n + 1$ components lie in the null space of the covariance matrix and are not uniquely determined. Therefore, interpretation should be limited to the first $n - 1$ components, as any orthonormal basis can span the null space. Nevertheless, for completeness, Table 2.6 reports the CSEs for all PCs when $p = 10$ and $n = 2, 3, \ldots, 10$. The "First $n-1$ PCs' Avg. CSE" row highlights the estimation error of the meaningful, data-driven components. As expected, the average CSE (computed over 100 simulation runs) increases as n decreases, reflecting the growing instability and poor generalization of the sample covariance matrix in such settings. This instability is primarily due to the overfitting of noise

in the data. In small-n regimes, even minor fluctuations can lead to significant distortions in estimated eigenvalues and eigenvectors [36, 37]. This overfitting results in the overdispersion of sample eigenvalues, greater standard deviation among them, and a misrepresentation of the true structure of the population covariance. Consequently, the PCs derived from the sample covariance may fail to reflect the true underlying data geometry, particularly when $n < p$.

Table 2.6: CSEs of all the cases of $n \leq p$ when p=10 and n=2,3,4,...,10

Description	**CSEs of PCs — p = 10**								
	n = 2	**n = 3**	**n = 4**	**n = 5**	**n = 6**	**n = 7**	**n = 8**	**n = 9**	**n = 10**
PC-1	0.3874	0.2226	0.1360	0.0769	0.0630	0.0473	0.0299	0.0270	0.0236
PC-2	0.8377	1.2483	1.1530	1.1309	1.0356	1.1208	1.0125	0.9753	1.0463
PC-3	1.0309	1.0944	1.0621	1.0808	1.0689	1.0828	1.0601	1.0277	1.0303
PC-4	1.0019	1.0148	0.8997	0.9764	0.9853	0.9579	0.9488	0.9337	0.9743
PC-5	0.9934	1.0239	1.0339	0.9783	1.1383	1.1172	1.0725	1.0504	1.0355
PC-6	0.9886	0.9610	1.0190	1.0324	1.0288	1.0847	1.0803	1.1283	1.0921
PC-7	1.0051	0.9880	1.0162	0.9674	0.9484	0.9020	0.8154	0.8079	0.8070
PC-8	0.9817	1.0324	1.0037	0.9957	0.9764	0.9725	0.9660	0.8764	0.8506
PC-9	0.9923	1.0039	0.9944	0.9645	0.9753	0.9981	0.9942	0.9020	0.7913
PC-10	0.9169	0.8853	0.9227	0.9035	0.8611	0.8340	0.8235	0.7818	0.6491
Overall Avg. CSE	**0.9136**	**0.9475**	**0.9241**	**0.9107**	**0.9081**	**0.9117**	**0.8803**	**0.8510**	**0.8300**
First (n-1) PCs' Avg. CSE	**0.3874**	**0.7354**	**0.7837**	**0.8162**	**0.8582**	**0.9018**	**0.8599**	**0.8533**	**0.8501**
Last (n-p+1) PCs' Avg. CSE	**0.9721**	**1.0005**	**0.9842**	**0.9736**	**0.9580**	**0.9267**	**0.9279**	**0.8419**	**0.6491**

Figure 2.6 represents the actual population covariance matrix. Figure 2.7 represents the sample covariance matrices in each $n < p$ case relative to p. Figure 2.7 clearly shows the inability to capture the variance-covariance structure of the actual covariance matrix when n goes down compared to p. This bias tends to have misinterpreted PCs when $n < p$ data settings.

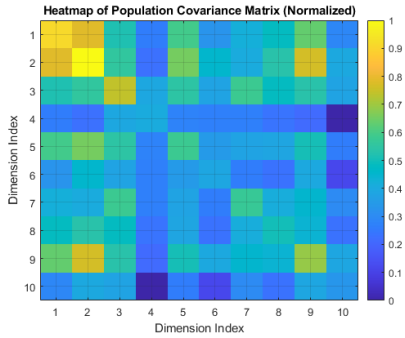
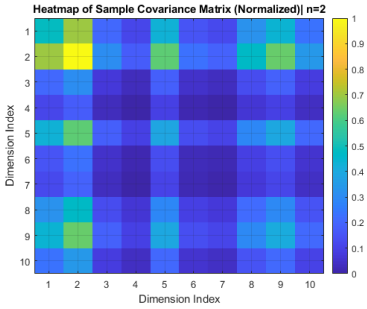
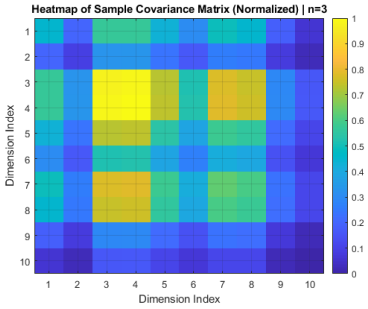


Figure 2.6: Normalized Population Covariance Matrix

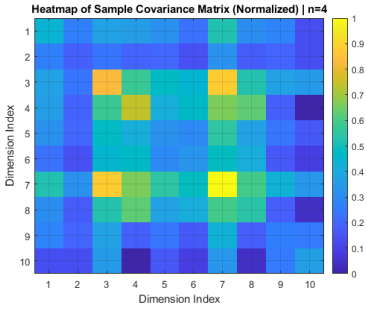
(a) n=2 < p=10



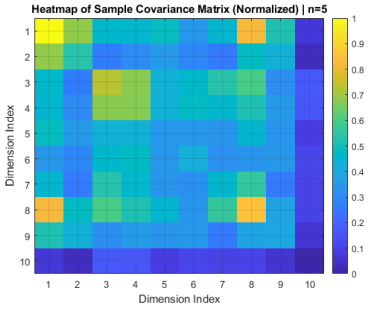
(b) n=3 < p=10



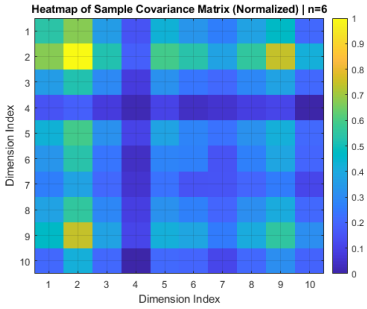
(c) n=4 < p=10



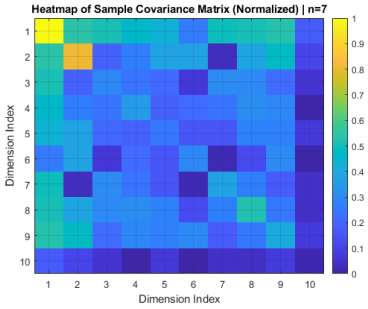
(d) n=5 < p=10



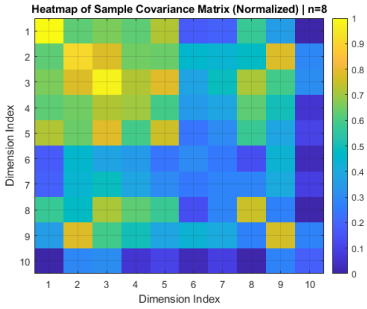
(e) n=6 < p=10



(f) n=7 < p=10



(g) n=8 < p=10



(h) n=9 < p=10

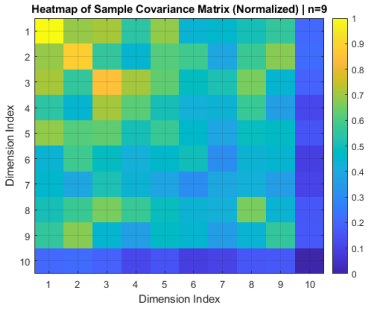


Figure 2.7: Sparsity of Covariance Estimates

2.3 Summary and Discussion

This chapter explored the PCA and its challenges in HD data. As we identified, the estimation of the sample covariance matrix plays a crucial role in PCA. However, in $n < p$ data settings, the condition number of the sample covariance matrix becomes large, the mean and standard deviation of eigenvalues go high, the biasedness of the eigenvalues becomes large, and the rank of the sample covariance matrix goes low. Due to these reasons, the estimation process of covariance produces a rank deficiency and ill-conditioned finite sample estimates for the sample covariance matrix when $n < p$ data settings. Because of this ill-conditioned sample covariance matrix estimates, the CSE of sample PCs becomes large, and the overdispersion of the first few PCs becomes large. At last, this chapter highlights the issues of using typical sample covariance estimation for PCA when $n < p$ data settings.

Chapter 3

State of the Art

3.1 Introduction

Principal Component Analysis is one of the most widely used dimensionality reduction techniques for extracting meaningful structure from high-dimensional datasets while minimizing redundancy. It is frequently employed as a preprocessing step in regression, classification, clustering, and other statistical modeling tasks.

The goal of this chapter is to review, compare, and synthesize a wide range of covariance estimation techniques developed specifically for HD data. In particular, we focus on methods tailored for the $n < p$ regime and explore their impact on improving PCA performance. Through this review, we aim to highlight the strengths and limitations of existing approaches, and to motivate the development of more robust PCA solutions that can effectively handle low-sample, high-dimensional datasets.

As discussed in Chapter 2, the effectiveness of PCA depends critically on the quality of the sample covariance matrix. In HD settings where the number of variables p exceeds the number of observations n (i.e., $n < p$), the sample covariance matrix becomes ill-conditioned or singular. This leads to unstable PCs, inflated or distorted eigenvalues, and poor generalization, thereby limiting the reliability of PCA in such scenarios. Covariance estimation plays a central role in PCA, as the PCs are derived from the eigenstructure of the covariance matrix. When the sample size is insufficient, standard estimation techniques fail to capture the true underlying structure of the data, which directly affects the quality and interpretability of the resulting components.

Before reviewing alternative covariance estimators, it is important to first understand the theoretical limitations of classical PCA in high-dimensional settings. The next section introduces key asymptotic frameworks and outlines when PCA can remain consistent, especially under structured models such as the spiked covariance model.

In HD settings, the consistency of PCA is strongly influenced by the underlying asymptotic framework. Under classical conditions where the number of variables p is fixed and the number of observations $n \to \infty$, PCA is known to be consistent, meaning that the sample eigenvalues and eigenvectors converge to their population counterparts. However, when both p and n tend to infinity such that the ratio $p/n \to c \in (0, \infty)$, standard PCA becomes inconsistent. In this regime, the sample covariance matrix may not accurately represent the population covariance structure, leading to PCs not aligning with the true population directions. Nevertheless, under the spiked covariance model, where a small number of population eigenvalues (the "spikes") are significantly larger than the rest, PCA can still yield consistent estimates of the leading PCs. This behavior has been well studied by Johnstone (2001) and further elaborated by Jung and Marron (2009) ([34, 38]). These insights underscore the importance of adopting regularized covariance estimators and improved PCA methodologies when working with HD data, particularly in cases where $n < p$, to ensure stable and reliable extraction of the dominant data structure.

3.2 Summary of Existing Covariance Estimation Methods

To address the limitations of the sample covariance matrix in HD settings, particularly when $n < p$, numerous regularized covariance estimation techniques have been developed. These methods aim to improve estimation accuracy, ensure positive semi-definiteness, and enable downstream applications such as PCA.

In what follows, we categorize and present key classes of covariance estimation approaches, highlighting representative methods from each. Each category reflects a distinct strategy for imposing structural assumptions or regularization, such as shrinkage, thresholding, penalization, or Bayesian modeling.

3.2.1 Shrinkage-Based Estimators:

Shrinkage-based estimators improve the sample covariance matrix by pulling it toward a stable target like the identity or diagonal matrix. This helps reduce variability and improve reliability, especially when $n < p$ and the sample covariance is unstable. These methods are simple, require few assumptions, and are commonly used to support more stable PCA results.

- **Stein's Estimator (1975)**

  Since the 1950s, Charles Stein observed that the eigenvalues of the sample variance-covariance matrix $\mathbf{S}$ are much more spread out than the eigenvalues of the population covariance matrix $\mathbf{\Sigma}$ [39]. In 1975, Charles Stein proposed a new covariance estimation method called "Stein Estimator" [40] 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 a diagonal matrix of eigenvalues, and R is a matrix of eigen vectors. The Stein's covariance matrix estimator is [39],

$$\hat{\mathbf{\Sigma}}_{\mathrm{ST}} = R\varphi(L)R^{\mathrm{T}}, \tag{3.1}$$

  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 is known as a rotation equivariant covariance estimator because the eigenvectors are not changed or regularized. To select 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, the limitations of using Stein's covariance estimation in HD settings are as follows.

  1. Stein's covariance estimation does not rely on the assumptions of normality. However, it is frequently discussed in the context of normally distributed data. This leads to limited theoretical guarantees beyond normality [41].
  2. Stein's Estimator may not necessarily produce a positive definite covariance matrix estimate. This limitation can lead to significant challenges on PCA when $n < p$ data settings [42, 43].

3. Stein's covariance estimation generally does not facilitate producing a sparse covariance matrix. It needs additional support to be sparse [35].
4. The performance of Stein's covariance estimation is sensitive to the choice of the shrinkage parameter. Stein's covariance estimation might either over-shrink or under-shrink if the shrinkage parameter is not chosen appropriately. It leads to a biased and ill-conditioned covariance matrix [43].
5. Stein's estimation is generally designed for scenarios where n > p. Therefore, Stein's covariance estimation becomes unstable in $n < p$ settings because the estimator may not adequately shrink the covariance matrix toward the target matrix [44].

- **Ledoit-Wolf Estimation (2004)**

The sample covariance matrix $\mathbf{S}$ eigenvalues can also be truncated using the empirical Bayesian estimator, which forms a weighted linear combination of $\mathbf{S}$ and the identity matrix $\mathbf{I}$. This approach balances the unbiased but unstable nature of $\mathbf{S}$ in HD data with the highly stable but potentially biased nature of $\alpha\boldsymbol{I}$. The empirical Bayesian estimator is given by:

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

In 2004, Ledoit and Wolf proposed a more general form of estimator, which would be a well-known stenian-type shrinkage (minimizing the MSE criterion) estimator for the $\hat{\mathbf{\Sigma}}$ [43]. That linear shrinkage estimator, as given by,

$$\hat{\mathbf{\Sigma}}_{\text{LW}} = \hat{\rho}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})^\top\right] + (1 - \hat{\rho})\left(\frac{\text{tr}(\boldsymbol{S})}{p} \times \boldsymbol{I}\right) \tag{3.2}$$

where;

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

$$\boldsymbol{S} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})^\top$$

According to the results of this LW estimator, the modified eigenvalues $\hat{\lambda}_{\mathrm{i}}$ are denoted as,

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

where $c = \frac{\mathrm{tr}(\boldsymbol{S})}{p}$, and l_i is the i^{th} eigenvalue of $\boldsymbol{S}$.

The Ledoit-Wolf covariance estimation addresses the instability and poor conditioning of the traditional covariance matrix estimation and gives a well-conditioned covariance estimation. However, the limitations of using Ledoit-Wolf covariance estimation in HD settings are as follows.

1. The Ledoit-Wolf covariance estimation shrinks all the elements of the sample covariance matrix by a uniform shrinkage approach. This implies that it shrinks all the elements of the sample covariance matrix by the same amount. Because of this uniform shrinkage, the Ledoit-Wolf covariance estimation may not be able to capture the actual variance-covariance structure of the original data. For example, as shown in Figure 3.1 & Table 3.1, the last p-n+1 eigenvalues are uniformly distributed when $n < p$ data settings [43].
2. The Ledoit-Wolf covariance estimation often guarantees a positive semi-definite covariance matrix, which is normally a significant and necessary requirement for running PCA. It achieves this through uniform shrinkage towards a fixed target matrix. Suppose the actual covariance matrix has a significantly different variance-covariance structure than the target matrix. In that case, the estimated sample covariance matrix may still be well-conditioned but may not accurately reflect the actual variance-covariance structure of the original data [45].
3. Some of the real-world applications have only a few numbers of dimensions that are strongly correlated. In such a situation, we need to have sparsity in the covariance matrix. However, the Ledoit-Wolf covariance estimation shrinks the sample covariance matrix towards a non-sparse dense target covariance matrix. Therefore, the Ledoit-Wolf covariance estimation does not provide a sparse sample covariance matrix [44].

Table 3.1: Eigenvalue distribution of $\hat{\Sigma}_{\text{LW}}$

Index	Population	MLE	LW
1	29.04677	255.8283	139.9317
2	2.82457	9.331329	8.515583
3	1.862018	2.56E-14	3.540727
4	1.214454	8.66E-15	3.540727
5	0.680735	4.20E-15	3.540727
6	0.571801	3.66E-15	3.540727
7	0.248015	2.89E-15	3.540727
8	0.112588	2.44E-15	3.540727
9	0.035137	1.48E-15	3.540727
10	0.015404	8.75E-17	3.540727

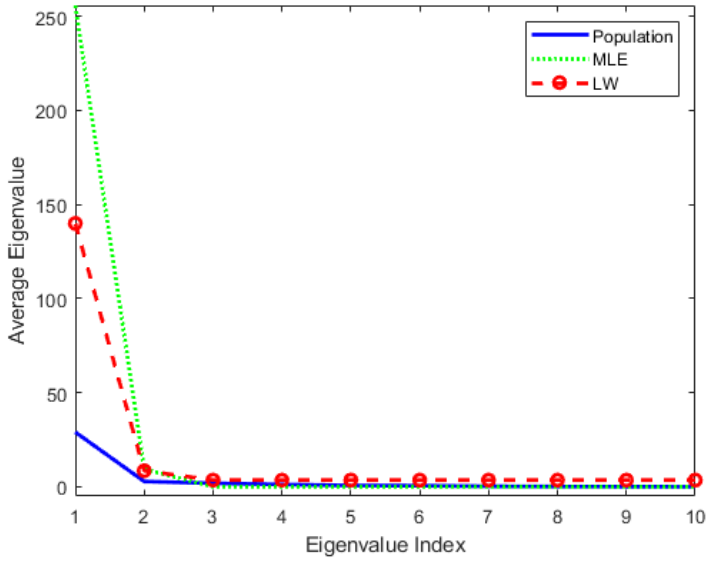


Figure 3.1: Scree plot of $\hat{\boldsymbol{\Sigma}}_{\text{LW}}$

Note: The data for Figure 3.1 and Table 3.1 were generated by simulation. Specifically, independent samples were drawn from a multivariate normal distribution, where $\mathbf{X}_i \sim \mathcal{N}_{10}(\mathbf{0}, \mathbf{I})$, with $p = 10$. The results represent the average over 500 replicates for a sample size of $n = 3$.

In essence, the Ledoit-Wolf covariance estimation does not give reliable finite sample estimates for PCA due to uniform shrinkage and non-sparsity of the sample covariance matrix when $n < p$ data settings.

- **Oracle Approximating Shrinkage (OAS) Estimation (2010)**

  In 2010, Chen developed a formula that, assuming the data were Gaussian-distributed, could be used to select a decrease in the shrinkage coefficient, resulting in an MSE lower than the one reported in the 2014 Ledoit-Wolf formula [46]. They improved 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

$$\hat{\mathbf{\Sigma}}_{\text{RBLW}} = (1 - \hat{\rho}_{\text{RBLW}})\boldsymbol{S} + \hat{\rho}_{\text{RBLW}}\hat{\boldsymbol{F}}, \tag{3.3}$$

  where

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

  Because the RBLW covariance estimation uses Rao-Blackwellization to reduce bias, this approach may introduce additional computational complexity, especially for large datasets with high dimensionality. To address this, Chen et al. [46] proposed an alternative iterative method to approximate the clairvoyant shrinkage estimator more efficiently. This method is known as the Oracle Approximating Shrinkage (OAS) covariance estimator, and it is defined as:

$$\hat{\mathbf{\Sigma}}_{\text{OAS}} = \gamma\boldsymbol{S} + (1 - \gamma)\hat{\delta}I, \tag{3.4}$$

  where $\boldsymbol{S}$ is the MLE of the covariance matrix, γ is the shrinkage intensity, and $\hat{\delta}$ is the estimated average eigenvalue of the true covariance matrix. Since the true eigenvalues are unknown, $\hat{\delta}$ is computed from the sample eigenvalues and adjusted to approximate the oracle (true) mean. This correction helps reduce the bias in the estimation of the covariance matrix by shrinking towards a more stable, scalar multiple of the identity matrix [46].

The estimate of the average eigenvalue $\hat{\delta}$ is calculated as:

$$\hat{\delta} = \frac{\sum_{i=1}^{p} \lambda_i}{p},$$

where λ_i are the eigenvalues of Σ_{MLE}, and p is the dimensionality of the covariance matrix.

However, the limitations of using OAS covariance estimation in HD settings are as follows.

1. The OAS covariance estimation highly depends on the assumption of spherical covariance structures. If the data does not follow this assumption, this covariance method leads to unreliable and biased finite sample estimates and may tend to have misleading PCs [43].
2. This covariance method does not perform well when outliers or extreme values are present in the data. In such cases, the estimated finite sample estimates of PCs may give potentially misleading conclusions [47].
3. The OAS covariance estimation often aims to minimize both bias and variance in the estimation. However, in HD data settings with complex variance-covariance structures, this covariance method might not effectively minimize bias and variance simultaneously. Due to this issue, the estimation process can result in either high bias or high variance in the sample covariance estimation. This may badly affect the efficiency and accuracy of PCA [48].

- **Non-Linear Shrinkage 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 [9].

  That covariance estimator as given by,

$$\hat{\mathbf{\Sigma}} = P\text{diag}(P^T \mathbf{\Sigma} P) P^T, \tag{3.5}$$

  where;
  P = $(p_1, p_2, ..., p_p)$ is a matrix of eigenvectors for sample covariance matrix.

  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 $\frac{p}{n} \to k > 0$ concerning the class of rotation-equivariant estimators [49].

The primary limitations of using the Non-Linear Shrinkage Estimation of large dimensional covariance estimation in HD settings are as follows.

1. Non-linear shrinkage Estimation is designed for statistical applications with sparse population parameters. Therefore, when applied to a real-world application that is not sparse, this method of covariance estimation may produce unreliable and inaccurate sample estimates and limited generalizability to diverse covariance structures [49].
2. Non-linear shrinkage Estimation suffers difficulty in characterizing complex variance-covariance structures [50].

- **Grand Average Covariance Estimation (2014)**

  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 is called Grand Average Covariance Estimation [51].

  That covariance estimator as given by,

$$\hat{\mathbf{\Sigma}} = P(M^{-1}\sum_{j=1}^{M} diag(P_{1j}^T \widetilde{\Sigma}_2^{(j)} P_{1j}))P^T, \tag{3.6}$$

  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 and then transformed into roughly orthogonal series. This results in a well-conditioned estimation even though the data contains fewer observations than dimensions. Furthermore, the estimator is designed without assumptions regarding the random sample distribution or any parametric structure associated with the variance matrix.

  This grand average covariance estimation method doesn't work well when the covariance between variables changes over time or across different conditions. In addition, when outliers or extreme values are present in data, they tend to be biased estimates.

- **NERCOME Estimation (2016)**

  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 [52]. That covariance estimator as given by,

$$\hat{\mathbf{\Sigma}} = P_1 d\text{diag}(P1_1^T \widetilde{\Sigma}_2 P_1) P1_1^T \tag{3.7}$$

  The importance of 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, and it assumes the covariance structure is smooth with no outliers or extreme values and follows a normal distribution.

3.2.2 Thresholding-Based Estimators:

Thresholding-based estimators improve the estimation of covariance matrices by reducing noise through element-wise thresholding. They operate under the assumption that many of the small off-diagonal elements in the true covariance matrix are negligible and can be set to zero. This introduces sparsity, which simplifies the covariance structure and helps improve interpretability and numerical stability. In HD settings where $n < p$, thresholding can reduce overfitting and make the resulting estimator more robust for tasks like PCA. These methods are especially valuable when the true underlying structure is assumed to be sparse.

- **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 dimensions estimated from n observations [35]. 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 $\mathbf{S}_{\text{i,j}} = 0$ if $\mid \boldsymbol{S}_{\text{i,j}} \mid \leq \tau$ where τ is a thresholding value. This is a 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 $\mathbf{X}$ be an $n \times p$ matrix of centered observations, where n is the number of samples and

p is the number of dimensions. The hard thresholding estimator of the covariance matrix Σ is given by:

$$\hat{\mathbf{\Sigma}}_{ij} = \begin{cases} \frac{1}{n}\sum_{k=1}^{n} x_{ik}x_{jk}, & \text{if } |\hat{\mathbf{\Sigma}}_{ij}| > \lambda \\ 0, & \text{otherwise} \end{cases} \tag{3.8}$$

where x_{ik} and x_{jk} are the ith and jth elements of the kth observation, and λ is the threshold.

Bickel and Levina's Covariance Estimation via the Hard Thresholding method presents several drawbacks in HD covariance estimation.

1. This covariance estimation is heavily sensitive to the selection of the threshold parameter. According to its theorem, this threshold parameter is a user-defined value. So, this is heavily biased and difficult to determine accurately in HD data settings including specially $n < p$ data settings [35].
2. The method's effectiveness decreases if the data contains outliers or extreme values. In such a situation, this covariance estimation produces skewed results [35].
3. The Hard Thresholding of this covariance estimation may struggle with complex dependency structures to capture the variance-covariance structure.

- **Adaptive Thresholding Estimation (2011)**

  Cai and Liu introduced an adaptive variant of hard thresholding covariance estimation in 2011, originally suggested by Bickel and Levina in 2008 [53]. The application of the thresholding technique to a correlation matrix is what adaptive thresholding is all about. Relationship, in which it adapts to each dimension. The Adaptive Thresholding Covariance Estimation (ATCE) is a method used to estimate covariance matrices in signal processing and machine learning. It involves adapting thresholding techniques to estimate the covariance matrix. One common approach to adaptive thresholding is to use a shrinkage estimator, such as the Ledoit-Wolf shrinkage estimator. This estimator is given by:

$$\hat{\mathbf{\Sigma}}_{\text{shrunk}} = \delta \cdot \hat{\Sigma}_{\text{diag}} + (1 - \delta) \cdot \hat{\Sigma} \tag{3.9}$$

where $\hat{\Sigma}_{\text{diag}}$ is the diagonal matrix with the sample variances on the diagonal, and δ is a shrinkage parameter.

The thresholded covariance matrix $\hat{\Sigma}_{\text{thresh}}$ can then be obtained by applying a thresholding function $T(\lambda, \cdot)$ element-wise to $\hat{\Sigma}_{\text{shrunk}}$:

$$\hat{\Sigma}_{\text{thresh}}(i,j) = T(\lambda, \hat{\Sigma}_{\text{shrunk}}(i,j))$$

where $T(\lambda, \cdot)$ is a thresholding function with parameter λ.

Common choices for the thresholding function include hard thresholding:

$$T(\lambda, x) = \begin{cases} x, & \text{if } |x| > \lambda \\ 0, & \text{otherwise} \end{cases}$$

and soft thresholding:

$$T(\lambda, x) = \text{sign}(x) \cdot \max(|x| - \lambda, 0)$$

The choice of λ can be determined through techniques such as cross-validation. However, the limitations of using the Adaptive Thresholding covariance estimation in HD settings are as follows.

1. The Adaptive Thresholding covariance estimation is less useful for a broader range of statistical applications. This covariance estimation is mainly targeted at situations involving sparse normal mean vectors, such as wavelet function estimation [35].
2. This covariance estimation is not applicable to capture more complex covariance structures, which limits its applicability in real-world statistical applications, including PCA [53].
3. The Adaptive Thresholding covariance estimation is based on the assumption of sparsity of the variance-covariance structure. However, when the actual covariance does not meet this requirement, the finite sample estimates of covariance provide inaccurate or unreliable covariance estimates. Therefore, it makes misinterpreted PCs [54].

- **Principal Orthogonal ComplEment Thresholding (POET) (2013)**

  The sparsity assumption directly imposed on $\mathbf{\Sigma}$ is often unsuitable in HD data due to the presence of strong homogeneous factors, which are latent variables or common effects that influence many variables similarly and lead to widespread correlations. These global factors result in dense covariance structures that cannot be adequately captured by sparsity alone.

To address this issue, in 2013, Fan et al. proposed a non-parametric estimator for $\mathbf{\Sigma}$ based on PCA, known as the Principal Orthogonal ComplEment Thresholding (POET) estimator [55]. Suppose the eigenvalues of $\hat{\mathbf{\Sigma}}$ are ordered as $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_p$, with corresponding eigenvectors $\hat{\xi}_1, \hat{\xi}_2, \ldots, \hat{\xi}_p$. Then, the spectral decomposition of the sample covariance matrix is given by:

$$\hat{\mathbf{\Sigma}} = \sum_{i=1}^{k} \hat{\lambda}_i \hat{\xi}_i \hat{\xi}_i' + \hat{R}_k^{\tau}, \tag{3.10}$$

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 $\hat{R}_K$ correlation matrix.

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 dimension 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. 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.

- **NOVELIST Estimator (2018)**

  In 2018, the NOVEL Integration of the Sample and Thresholded (NOVELIST) covariance estimator was introduced by Huang and Fryzlewicz [56], which combines linear shrinkage with sparse estimators, as follows:

$$\hat{\mathbf{\Sigma}}_N = (1 - \delta)\boldsymbol{S} + \delta T_{\lambda}(\boldsymbol{S}), \tag{3.11}$$

where $T_\lambda(\boldsymbol{S})$ is a thresholded estimator of $\boldsymbol{S}$ with parameter λ, defined as

$$T_\lambda(\boldsymbol{S}) = \left(s_{ij}\mathbf{1}\left(|s_{ij}| \geq \lambda\right)\right).$$

The NOVELIST estimator reduces the sample variance 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.

3.2.3 Penalized Likelihood-Based Estimators:

Penalized likelihood-based estimators improve covariance estimation by incorporating regularization directly into the likelihood framework. These methods typically add a penalty term, often based on the norm ℓ_1, to the log-likelihood function to encourage sparsity in the estimated covariance or precision matrix. The penalty helps control the estimate's complexity and mitigates overfitting, which is especially important in HD settings where $n < p$. A key advantage of this approach is that it can enforce positive definiteness while promoting sparse structure, which improves both interpretability and numerical stability. Penalized estimators are particularly useful when the underlying relationships among variables are believed to be sparse, and offer a principled way to balance model fit with structural simplicity.

- **Graphical Lasso (GLasso) Estimation (2008)**

  The GLasso estimator is a sparse penalized MLE for a precision matrix (the inverse of the covariance matrix) of any given member of a family of generalized multivariate normal distributions. In the case of small samples, sparse inverse covariances are generally more effective than shrunk covariances. Therefore, In 2008, Friedman [57] proposed a covariance estimation method to other estimator and distribution types to extend the Dempster covariance selection problem for a multivariate Gaussian distribution when observations are constrained.

That covariance estimator as given by,

$$\hat{\boldsymbol{\theta}} = argmin_{\boldsymbol{\theta} \geq 0}[tr(\boldsymbol{S\theta}) - log\det(\boldsymbol{\theta}) + \lambda\Sigma_{j \neq k} \mid \boldsymbol{\theta}_{jk} \mid], \tag{3.12}$$

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 HD covariance estimation. Despite its benefits, including ease of interpretation and sparsity emphasis, GLasso faces computational inefficiencies in datasets with many dimensions relative to the sample size. Therefore, the limitations of using GLasso covariance estimation in HD settings are as follows.

1. One major application of PCA addresses the multicollinearity issues in highly correlated datasets. However, GLasso covariance estimation cannot estimate the precision matrix correctly in the presence of multicollinearity in a dataset [58]. As Netti said, GLasso covariance estimation can handle low multicollinearity without badly affecting the interpretability of the PCs. However, it cannot overcome severe multicollinearity situations [59]. Therefore, GLasso covariance estimation is unsuitable for obtaining PCs whether n is less than or greater than p.
2. GLasso covariance estimation heavily relies on a well-selected regularization parameter. However, this is not a straightforward process. We must use appropriate model selection techniques to obtain this. An incorrect regularization parameter can lead to underfitting or overfitting in the data and misleading PCs [57].
3. GLasso covariance estimation assumes that the true inverse covariance matrix (precision matrix) is sparse. This means that the correlation between two dimensions is zero when controlling for the effects of all other dimensions in the dataset. However, the true precision matrix is not sparse, or some essential low-correlated dimension exists; this assumption leads to biased sample covariance estimates and misinterpreted PCs [57].

- **Penalized Maximum Likelihood (PML) Estimation (2011)**

  In 2011, Bien and Tibshirani suggested a Penalized Maximum Likelihood (PML) covariance estimation with a weighted lasso-type penalty method based on a sample of vectors drawn from a multivariate Gaussian distribution [60], as follows:

$$\hat{\boldsymbol{\Sigma}} = \underset{\boldsymbol{\Sigma} \succ 0}{\operatorname{argmin}} \left\{ \log\det(\boldsymbol{\Sigma}) + \operatorname{tr}(\boldsymbol{S\Sigma}^{-1}) + \lambda\|\boldsymbol{\Sigma}\|_1 \right\}, \ \lambda > 0 \tag{3.13}$$

The corresponding optimization problem can be stated as:

$$\begin{aligned} \text{minimize} \quad & \log\det(\boldsymbol{\Sigma}) + \operatorname{tr}(\boldsymbol{S}\boldsymbol{\Sigma}^{-1}) + \lambda\|\boldsymbol{\Sigma}\|_1 \\ \text{subject to} \quad & \boldsymbol{\Sigma} \succ 0 \end{aligned}$$

The empirical covariance matrix of the optimization problem must be positively definite 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. PML covariance estimation is highly sensitive to the penalty parameter. This implies that the choice of the penalty parameter is an essential requirement for PML covariance estimation. If this penalty parameter is not chosen correctly, the sample covariance estimates may not reflect the true variance-covariance structure of data. Moreover, PML covariance estimation performs poorly when the data has outliers or extreme values.

- **Positive Definite Sparse Covariance Estimation (PDSCE) (2012)**

  Using Convex optimization, Rothman proposed a covariance estimator called Positive Definite Sparse Covariance Estimator (PDSCE) in 2012 [61] to generate a sparse estimate of a covariance matrix, which is positive definite, and is suitable for HD 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,

$$\hat{\boldsymbol{\Sigma}} = \underset{\Sigma=\Sigma^{\mathrm{T}}}{\operatorname{argmin}}[\| \boldsymbol{\Sigma} - \boldsymbol{S} \|_F^2 + \lambda \| \boldsymbol{\Sigma}^{-1} \|_1 - \gamma log(|\boldsymbol{\Sigma}|)], \lambda, \gamma > 0 \tag{3.14}$$

PDSCE has some limitations, especially when dealing with small sample sizes compared to the number of dimensions. The primary limitations of using the PDSCE in HD settings are as follows.

1. PDSCE may not perform well $n < p$ data settings because it does not have enough data to estimate the covariance matrix accurately. Therefore, PDSCE is not a suitable covariance estimation method that can be used for $n < p$ data settings [35].
2. Like Adaptive Thresholding covariance estimation, PDSCE also assumes that the sample covariance matrix follows a sparse covariance matrix. It is primarily designed for sparse covariance matrices. That implies that PDSCE makes it difficult to capture the complex covariance structures or non-sparse covariance structures [54].

- **Positive Definite ℓ_1 Penalized Estimation (PDGLasso) (2012)**

  Xue [62] 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 HD covariance matrices that are both positive definite and sparse, which is particularly relevant when $p > n$. To ensure the positive definiteness of the sample covariance matrix, PDSLasso introduces l_1 penalty to the off-diagonal elements of the covariance matrix. However, it has several limitations, such as computational complexity when dealing with HD data settings and sensitivity to penalties and tuning parameters. We need sound prior knowledge about the data structure to establish this covariance estimation in practice.

- **Sparse Covariance Estimation via Nested Lasso (2008)**

  In 2008, Levina introduced another sparse covariance estimation method for large covariance matrices using a nested lasso penalty. This estimation mainly targets to facilitate more straightforward interpretation and computational efficiency. However, this method also does not perform well in HD data settings due to the sensitivity of penalty parameters. Hence, this covariance estimation method also fails to capture the complex variance-covariance structures, especially in $n < p$ data settings [63].

- **Joint PENalty Estimation of Covariance (JPEN) (2016)**

  In 2016, Maurya proposed a method for estimating well-defined and sparse covariance and inverse covariance from an HD sample of vectors derived from a sub-Gaussian distribution [13]. 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 is;

$$\hat{\mathbf{\Sigma}}_R = D\widehat{\Gamma}D^T, \tag{3.15}$$

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

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

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

  In addition, the JPEN estimator reconstructs the sparsity pattern for the real covariance matrix and offers an approximate representation of the fundamental eigen-range. However, the limitations of using the JPEN covariance estimation in HD settings are as follows.

1. As with most of the other alternative covariance estimations, JPEN also assumes a sparse covariance structure. However, if the true covariance structure is not sparse and exhibits a complex variance-covariance structure, JPEN does not perform well. Therefore, JPEN may not generalize more HD data settings in practice [38].
2. JPEN's performance depends on the choice of tuning parameters. However, the process of selecting optimal tuning parameters is challenging, especially in HD data settings [50]. If the tuning parameters are not chosen correctly or optimally, there is a risk of overfitting sample estimates when n is much less than p [64].

3.2.4 Other/Hybrid Methods:

Hybrid methods combine multiple strategies, such as shrinkage, conditioning, or transformation, to improve covariance estimation in complex settings. These approaches aim to balance accuracy, stability, and structural flexibility, especially when no single assumption like sparsity or low-rankness clearly applies. They are useful for producing reliable covariance estimates in HD data, particularly when $n < p$.

- **Winsorized Covariance Estimation (2013)**

  In 2013, Won [65] proposed that the estimator should be maximized with 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 compared to $\mathbf{S}$. Still, it is not shown that optimal nonlinearly shrinkability can be achieved with such a loss. The sample eigenvalue of the middle part of the sample remains unchanged. In this approach, the higher eigenvalues are winsorized at specific constants, and the estimator is defined as:

$$\hat{\mathbf{\Sigma}} = \underset{\mathbf{\Sigma}}{\operatorname{argmax}}\, L(\boldsymbol{S}, \mathbf{\Sigma}), \tag{3.16}$$

  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.

  The major drawback of this covariance estimation method is its heavy reliance on condition number regularization. Therefore, it may not perfectly capture the underlying variance-covariance structure of the dataset where the condition number regularization cannot capture the actual variance-covariance structure. Moreover, this method may perform better in highly heterogeneous covariance structures than other variance-covariance structures.

- **Minimax Optimal Sparse PCA (2013)**

  Vu and Lei proposed two ideas to capture the sparsity of a covariance matrix. These methods facilitate the estimate of the subspace formed by the principal eigenvectors of a covariance matrix and the understanding of the variance-covariance structure of HD data. However, this method may not perform well in different types of covariance structures and $n < p$ data settings [66].

- **Sparse PCA via Regression Screening (2010**

  Bickel suggested a method for selecting the most valuable and vital dimensions, such as PCA, in HD data settings. He used a sparse representation of the regression function for this. However, this method only partially applies to estimating the sample covariance matrix, which is unsuitable for covariance estimation in HD data settings. This method is more effective for the regression tasks [67].

3.2.5 Bayesian Approaches:

Bayesian approaches estimate the covariance matrix by placing structured prior distributions over the parameter space, allowing the incorporation of prior beliefs such as sparsity, low-rank structure, or hierarchical relationships. Unlike purely frequentist methods that rely on sample-based regularization, Bayesian estimators integrate uncertainty into the model and produce posterior distributions for the covariance matrix. This enables more flexible and adaptive shrinkage that can adjust to complex data structures. In HD settings, Bayesian methods have been proposed to regularize $\hat{\mathbf{\Sigma}}$ by shrinking eigenvalues, controlling noise, and improving stability. These methods are particularly useful when prior information is available or when uncertainty quantification is important, although they often come with increased computational cost and sensitivity to prior specification. References such as [68], [69], and [70] offer different Bayesian formulations for addressing HD covariance estimation challenges.

3.3 Summary of Covariance Methods for HD PCA

The effectiveness of PCA in HD data settings, particularly when the number of variables p exceeds the number of observations n, depends critically on the quality of the covariance matrix estimate. In such $n < p$ scenarios, the sample covariance matrix becomes singular or poorly conditioned, often resulting in unstable PCs and misleading variance decompositions. As a result, regularized covariance estimation has become essential for improving the reliability and interpretability of PCA in HD applications.

Table 3.2 presents a comparative summary of widely studied covariance estimation methods, categorized by their shrinkage strategy, sparsity induction, positive semi-definiteness, and applicability to PCA when $n < p$. It highlights key methodological features and assumptions, such as the use of regularization, structural constraints, and computational demands, that influence each method's effectiveness. The table is intended to provide a clear overview of the trade-offs involved in estimator selection, including stability, interpretability, and adaptability to unknown or complex covariance structures. This structured comparison helps identify methods that are more suitable for PCA in HD, low-sample-size settings. From a broader perspective, several common themes emerge across these methods. First, all approaches seek to balance the trade-off between bias and variance. Introducing regularization, whether through shrinkage, thresholding, or penalization, reduces variance, but often at the cost of adding bias. When this trade-off is well-managed, it can stabilize estimates and improve PCA outcomes; when poorly tuned, it can lead to either underfitting or over-regularization that distorts the eigenspectrum.

Second, the choice of regularization parameters plays a crucial role in the performance of many estimators. Threshold levels, penalty strengths, and shrinkage intensities must often be selected through data-driven procedures such as cross-validation. However, in HD settings, such tuning is not only computationally expensive but also statistically unreliable when n is small. This highlights the importance of methods that are robust to tuning or can estimate optimal parameters automatically.

Third, structural assumptions strongly influence which estimators are most effective in practice. Methods assuming sparsity are effective when many off-diagonal elements of the true covariance matrix are zero or near-zero, while low-rank assumptions are suitable when the data has an underlying latent factor structure. However, real-world datasets often do not conform strictly to either assumption, and methods that rigidly enforce such structures may miss important features or introduce distortions.

Shrinkage-based estimators are attractive due to their simplicity, speed, and ability to stabilize the sample covariance matrix with minimal assumptions. Among them, the Ledoit-Wolf estimator remains a widely accepted baseline. It shrinks the sample covariance toward a target (usually the identity matrix), improving conditioning and ensuring positive semi-definiteness. While its shrinkage is uniform and may not adapt to complex structures, its stability and general applicability make it a reliable choice for PCA when $n < p$.

Table 3.2: Comparative Summary of Covariance Estimation Methods with Integrated PCA Comments under Unknown Structures

Method	**Shrinkage Type**	**Sparsity**	**Positive (Semi-)Definite**	$n < p$	**PCA Use**	**Comments**
Stein's Estimator	Linear	No	Not guaranteed	Limited	Limited	Very sensitive to the choice of shrinkage intensity and does not promote sparsity. PCA ($n < p$): Highly unstable; often leads to exaggerated eigenvalues and inaccurate PCs.
Ledoit-Wolf Estimation	Linear	No	Positive Semi-Definite	Moderate	Moderate	Simple and widely used method applying equal shrinkage to all directions. PCA ($n < p$): Improves overall stability and conditions the covariance matrix. However, it applies the same shrinkage intensity to all components, which may overshrink weaker signals and distort the variance structure in cases with complex or heterogeneous covariance patterns.
OAS Estimation	Oracle-based Linear	No	Positive Semi-Definite	Moderate	Moderate	Improves Ledoit-Wolf by optimally estimating shrinkage intensity. PCA ($n < p$): Works better under spherical structure, but may perform poorly on real data with heterogeneous patterns.

Continued on next page

Table 3.2 – Continued from previous page

Method	**Shrinkage Type**	**Sparsity**	**Positive (Semi-)Definite**	$n < p$	**PCA Use**	**Comments**
Non-Linear Shrinkage	Non-linear	No	Positive Semi-Definite	Yes	Yes	Adapts shrinkage to individual eigenvalues using advanced techniques. PCA ($n < p$): Stabilizes eigenvalues effectively but assumes smooth underlying structure; computationally demanding.
Grand Average Estimation	Averaging shrinkage	No	Positive Semi-Definite	Yes	Yes	Averages covariance estimates over multiple subsamples. PCA ($n < p$): May oversmooth data, leading to loss of fine structure necessary for detecting weaker PCs.
NERCOME	Nonparametric shrinkage	No	Positive Semi-Definite	Yes	Yes	Data-driven estimator robust to model assumptions. PCA ($n < p$): Provides stable results under unknown structure, though sensitive to sample partitioning.
Hard Thresholding	None	Yes	Not guaranteed	Limited	Limited	Sets small covariance entries to zero directly. PCA ($n < p$): Can remove meaningful correlations, leading to inaccurate or fragmented PCs.
Adaptive Thresholding	Adaptive	Yes	Positive Definite	Moderate	Moderate	Thresholding strength varies with variable-wise scaling. PCA ($n < p$): Better than hard thresholding, but still assumes sparsity and can ignore moderate correlations.

Continued on next page

Table 3.2 – Continued from previous page

Method	**Shrinkage Type**	**Sparsity**	**Positive (Semi-)Definite**	$n < p$	**PCA Use**	**Comments**
POET	Low-rank + sparse	Yes	Positive Semi-Definite	Yes	Yes	Targets low-rank signals plus sparse noise, ideal for factor models. PCA ($n < p$): Performs very well if low-rank assumption holds, but may suppress signal in complex or dense structures.
NOVELIST	Shrinkage + thresholding	Yes	Positive Definite	Yes	Yes	Efficient method combining shrinkage and elementwise thresholding. PCA ($n < p$): Useful in ultra-high dimensions, though may oversimplify nuanced dependence in the data.
GLasso	None (L1 on precision)	Yes (Precision)	Positive Definite	Moderate	Moderate	Promotes sparsity in the inverse covariance matrix. PCA ($n < p$): Not ideal for PCA since it estimates conditional independence, not direct variance-covariance structure.
PML	L1 penalty	Yes	Positive Definite	Moderate	Moderate	Penalizes likelihood to encourage sparsity. PCA ($n < p$): May overly shrink components, leading to loss of structure in the eigenspectrum.
PDSCE	L1 penalty	Yes	Positive Definite	Limited	Limited	Convex formulation enforcing sparse structure. PCA ($n < p$): Can fail when true covariance is dense or contains mixed-size components.

Continued on next page

Table 3.2 – Continued from previous page

Method	**Shrinkage Type**	**Sparsity**	**Positive (Semi-)Definite**	$n < p$	**PCA Use**	**Comments**
PDGLasso	L1 penalty	Yes	Positive Definite	Limited	Moderate	Ensures positivity while inducing sparsity. PCA ($n < p$): Penalization may suppress minor but informative directions, reducing interpretability.
Nested Lasso	Nested L1 penalty	Yes	Positive Definite	Limited	Limited	Promotes hierarchical sparsity across variables. PCA ($n < p$): Risk of discarding weak but relevant components if nested penalties dominate.
JPEN	Joint L1 + spectral penalty	Yes	Positive Definite	Moderate	Moderate	Combines penalties on both covariance entries and eigenvalue dispersion. PCA ($n < p$): Improves eigenvalue spread but still less robust without structural assumptions.
Winsorized Estimation	Conditioned eigenvalue	No	Positive Semi-Definite	Limited	Limited	Suppresses large eigenvalues to regularize the matrix. PCA ($n < p$): Reduces dominant variance but may mask relevant variation in complex data.
Bayesian Approaches	Prior-based shrinkage	Optional	Positive Definite (via prior)	Yes	Yes	Use flexible priors to encode structure. PCA ($n < p$): Effective if priors reflect reality; otherwise, results are model-sensitive and computationally heavy.

Thresholding-based methods promote sparsity by setting small elements of the sample covariance to zero. These are effective when the true covariance is believed to be sparse, but they are sensitive to the thresholding rule and may fail to preserve weak but important signals. Penalized likelihood estimators extend this idea by integrating sparsity-inducing penalties into the likelihood framework, often guaranteeing positive definiteness. However, they require careful tuning and may distort the contribution of weaker PCs if over-penalized.

Hybrid methods combine elements of multiple strategies, such as shrinkage and thresholding, to achieve a balance between bias control and structural flexibility. These approaches can be powerful but often increase computational complexity and require more tuning choices. Bayesian approaches, on the other hand, offer probabilistic modeling of the covariance structure through the use of priors. They are highly flexible and can encode sparsity, low-rankness, or hierarchical dependencies, but tend to be computationally intensive and sensitive to the choice of prior distributions.

Among the more promising estimators for PCA when $n < p$ are the Ledoit-Wolf estimator, Non-Linear Shrinkage, and POET. The Ledoit-Wolf estimator is widely used in practice due to its simplicity, computational efficiency, and ability to produce well-conditioned and positive semi-definite estimates with minimal tuning. While it applies uniform shrinkage and may not capture structural nuances in the data, it remains a strong baseline for stabilizing PCA in HD settings. Non-Linear Shrinkage extends this idea by adapting shrinkage intensity to the spectrum of the sample covariance, offering improved eigenvalue estimation and greater flexibility. POET is particularly effective when the data follows a low-rank plus sparse noise model, making it well-suited for latent factor structures. However, both Non-Linear Shrinkage and POET rely on structural assumptions that may not hold universally. In contrast, Ledoit-Wolf offers robustness across a broader range of scenarios, reinforcing its role as a dependable default method in $n < p$ PCA applications.

Despite these advancements, limitations persist. Many estimators assume specific structural forms that may not generalize well, and the reliance on tuning parameters limits reproducibility and interpretability. In practice, no single method dominates across all scenarios, and the choice of estimator often depends on the underlying data characteristics and computational constraints.

That said, the Ledoit-Wolf estimator stands out for its practicality, widespread acceptance, and consistent performance across a range of applications. It provides a robust, positive semi-definite, and computationally efficient alternative to the sample covariance matrix, making it particularly suitable for PCA in $n < p$ settings where stability is critical. Although the method applies uniform shrinkage and lacks adaptivity to structural features such as sparsity or low-rank patterns, it requires minimal tuning and performs reliably even when little is known about the underlying data structure.

As such, it serves as a strong default and a valuable benchmark against which more complex or assumption-driven methods can be compared. This observation provides important motivation for the methodology developed in the following chapters, which seeks to retain the stability and simplicity of shrinkage estimation while extending its flexibility to accommodate more general HD covariance structures.

3.3.1 Comparison with Other PCA Methods in HD Settings

Beyond improving covariance estimation, many researchers have developed alternative PCA algorithms aimed at handling HD or large-scale data. Two well-known techniques are Batch PCA and Incremental PCA (also referred to as streaming or online PCA). These methods are designed primarily for computational scalability. Batch PCA processes data in chunks, while Incremental PCA updates the PCs sequentially as new data arrives. A recent review by [Greenacre et al, 2022] provides a comprehensive summary of PCA variants tailored for large datasets [71].

Batch PCA and Incremental PCA are popular computational variants of standard PCA, especially designed to handle large-scale datasets or streaming data. Batch PCA processes chunks of data sequentially, while Incremental PCA updates PCs incrementally as new observations become available. Both approaches rely fundamentally on the eigendecomposition or SVD of the centered data matrix $\boldsymbol{X} \in \mathbb{R}^{n \times p}$, or on the eigendecomposition of the corresponding sample covariance matrix $\boldsymbol{S} = \frac{1}{n-1}\boldsymbol{X}^\top \boldsymbol{X}$.

While these algorithms are computationally efficient, they do not resolve the statistical issues that arise when the number of observations n is smaller than the number of variables p. In such HD settings ($n < p$), the sample covariance matrix $\boldsymbol{S}$ becomes a poor estimator of the population covariance matrix $\boldsymbol{\Sigma}$. This is due to the fact that, under HD asymptotic regimes where $n, p \to \infty$ and $p/n \to c \in (0, \infty)$, the eigenvalues of $\boldsymbol{S}$ do not converge to the eigenvalues of $\boldsymbol{\Sigma}$. Instead, they are systematically biased and exhibit significant spread, as predicted by the Marčenko–Pastur law [72].

Mathematically, let $\lambda_1, \ldots, \lambda_p$ denote the eigenvalues of $\boldsymbol{S}$, and $\lambda_1^{(\text{true})}, \ldots, \lambda_p^{(\text{true})}$ denote those of $\boldsymbol{\Sigma}$. Then, in the HD regime, we generally have:

$$\lambda_i \not\to \lambda_i^{(\text{true})} \quad \text{as} \quad n \to \infty, \quad p/n \to c > 0.$$

This inconsistency in eigenvalue estimation leads to unreliable estimation of the PCs, particularly the leading ones, which are most often used for data interpretation and dimensionality reduction. In fact, the estimated PCs may deviate significantly from the true eigen-directions of the

data, and thus fail to capture the underlying data structure. This issue is even more pronounced in Incremental PCA, where updates are made locally based on partial information, without global regularization or error control.

Moreover, because $\boldsymbol{S}$ is rank-deficient when $n < p$, at most $n - 1$ of its eigenvalues are nonzero, limiting the number of informative PCs that can be extracted. Batch and Incremental PCA methods do not explicitly address this degeneracy and typically retain components based on variance thresholds, which can mislead interpretation in the presence of noisy or unstable components.

Therefore, while Batch and Incremental PCA offer computational scalability, they are not statistically reliable for PC extraction in small-sample HD settings. They do not incorporate any mechanism to correct for the overdispersion of eigenvalues or to stabilize the directions of estimated components. This highlights the need for alternative PCA approaches that incorporate regularization into the covariance estimation process to better handle the $n < p$ regime.

3.4 Research Gap and Motivation

Although many covariance estimators have been proposed to address the shortcomings of the sample covariance matrix in HD settings, few have been developed with PCA as a central application. Most focus on global estimation accuracy or graphical model inference rather than accurate eigenspectrum recovery or PC stability. Furthermore, no single method performs optimally across all scenarios,effectiveness depends heavily on the underlying data structure and the analytical goal. This reveals a critical gap in the literature: the lack of a dedicated covariance estimation approach optimized specifically for PCA when $n < p$.

Among existing methods, the Ledoit-Wolf estimator remains a widely used practical choice due to its positive semi-definiteness, minimal tuning, and reliable performance in low-sample settings. However, its uniform shrinkage approach can oversimplify the covariance structure and limit its effectiveness in more complex or heterogeneous data. This motivates the development of a new method that retains the stability and robustness of shrinkage-based approaches while adapting to unknown structure. The aim of our proposed methodology is to address these limitations and improve PCA performance in HD, low-sample-size contexts.

3.5 Conclusion

In HD data settings where $n < p$, the performance of PCA is heavily dependent on the quality of the covariance matrix estimator. The standard maximum likelihood estimator becomes unstable or singular in such scenarios, often leading to inaccurate eigendecompositions and unreliable PCs. To address this, various alternative covariance estimation techniques have been proposed, including shrinkage-based methods, thresholding strategies, penalized likelihood approaches, and Bayesian formulations.

While each of these methods offers improvements over the sample covariance matrix, they are not without limitations. Many require careful tuning, rely on structural assumptions such as sparsity or low-rankness, or fail to guarantee positive definiteness under all conditions. Additionally, some lack robustness to outliers or are computationally intensive. Even the widely used Ledoit-Wolf estimator, valued for its simplicity and stability, applies uniform shrinkage and may not capture more complex or heterogeneous covariance structures.

Most importantly, although PCA is one of the most common applications of covariance estimation, relatively few existing methods have been developed or evaluated with PCA as a primary focus, particularly in contexts $n < p$. This underscores the need for a more targeted approach to covariance estimation that not only improves stability and conditioning but also enhances the quality of PCA outputs in HD, low-sample-size settings. These observations directly motivate the development of the new methodology introduced in the next chapter.

Chapter 4

Pairwise Differences Covariance (PDC) Estimation

4.1 Introduction

Traditional sample covariance estimators, such as the Maximum Likelihood Estimator (MLE), rely heavily on first-order statistics, specifically deviations from the sample mean, to capture the relationships between variables. Although this approach performs well when the sample size n is sufficiently large relative to the number of variables p, it becomes unstable in modern HD settings where $n < p$, or when the data exhibit heavy-tailed distributions or outliers. The sample covariance matrix is often rank-deficient in such scenarios, leading to distorted eigenvalue spectra, non-invertibility, and unstable principal component directions. More critically, using the sample mean as a centering point introduces additional estimation variability in finite samples and does not utilize the full geometric structure present in the data.

We propose a novel covariance estimation method called the *Pairwise Difference Covariance* (PDC) estimator to address these limitations. The core idea behind PDC is to replace the dependence on sample mean centering with a more robust and geometrically informative structure derived entirely from pairwise differences between observations.

Unlike standard pairwise difference-based estimators that use only first-order differences of the form $\mathbf{x}_i - \mathbf{x}_j$, our approach further introduces and aggregates structured second-order pairwise difference interactions, such as:

$$(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_k)^T, \quad \text{for distinct } j, k \neq i.$$

These second-order terms enable the covariance estimator to capture higher-order relational information embedded in the geometry of the data cloud.

From a sample size perspective, this construction provides a significant advantage. Although the MLE uses only samples centered on n, the PDC estimator is based on:

$$m = \frac{n^2(n-1)}{2}$$

second-order difference pairs, resulting in an effective increase in the number of informative directions for estimating variance and covariation structure, without requiring additional samples.

The PDC approach thus achieves two simultaneous objectives:

- It enhances the statistical efficiency of the estimator by leveraging all possible pairwise interactions among data points, effectively boosting the sample size from n to $\mathcal{O}(n^2)$.
- It introduces a built-in regularization mechanism by smoothing high-variance directions and inflating underrepresented ones, leading to a more stable and well-conditioned eigenstructure of the covariance matrix.

Therefore, the proposed PDC estimator addresses the fundamental limitations of classical covariance estimators in HD, small-sample, or heavy-tailed regimes by combining first- and second-order pairwise difference structures. It provides a consistent, robust, and geometrically enriched covariance estimate suitable for modern multivariate analysis and dimensionality reduction techniques like PCA.

The following section outlines the step-by-step construction of the proposed Pairwise Difference Covariance (PDC) estimator. This systematic formulation highlights how the estimator is derived from pairwise differences and their structured second-order interactions, providing insight into its theoretical foundations and practical advantages over conventional covariance estimators.

Step 1: Structure of the Data Matrix

To formalize the construction of the proposed covariance estimator, we begin by establishing the structure and notation of the observed data.

Let

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_n^T \end{pmatrix} = \begin{pmatrix} X_{11} & X_{12} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{np} \end{pmatrix} \in \mathbb{R}^{n \times p}$$

be a data matrix consisting of n independent observations and p variables, where each row vector $\mathbf{x}_i \in \mathbb{R}^p$ is an observation in p-dimensional space:

$$\mathbf{x}_i = (X_{i1}, X_{i2}, \ldots, X_{ip}), \quad i = 1, \ldots, n.$$

Each column corresponds to a variable, and each row corresponds to a multivariate observation. Traditional covariance estimation begins by centering the data:

$$\tilde{\mathbf{x}}_i := \mathbf{x}_i - \bar{\mathbf{x}}, \quad \text{where} \quad \bar{\mathbf{x}} := \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i,$$

followed by forming the empirical covariance matrix:

$$\mathbf{S} = \frac{1}{n-1} \sum_{i=1}^{n} \tilde{\mathbf{x}}_i \tilde{\mathbf{x}}_i^T.$$

However, when $n \ll p$ or when the data is heavy-tailed or contains outliers, this mean-based approach becomes unstable. As an alternative, the Pairwise Difference Covariance (PDC) estimator avoids centering and instead leverages geometric relationships among observations.

Step 2: Computation of First-Order Pairwise Differences

To construct the PDC estimator, we first compute the **first-order pairwise differences** between all ordered pairs of distinct observations in the dataset.

We consider all ordered index pairs $(i, j) \in \{1, \ldots, n\}^2$ subject to the condition $i \neq j$, thereby excluding diagonal elements where $i = j$. This results in a total of $n(n-1)$ distinct ordered pairs.

For each such pair, we define the first-order pairwise difference vector as:

$$\mathbf{d}_{ij} := \mathbf{x}_i - \mathbf{x}_j \in \mathbb{R}^p.$$

Each vector $\mathbf{d}_{ij}$ represents the relative position of observation $\mathbf{x}_i$ with respect to observation $\mathbf{x}_j$. Since the pairs are ordered, the direction is preserved; in general, $\mathbf{d}_{ij} \neq \mathbf{d}_{ji}$, highlighting the asymmetry that captures additional geometric information in the data.

All pairwise difference vectors $\mathbf{d}_{ij}$ are stacked row-wise into a matrix $\mathbf{D} \in \mathbb{R}^{m \times p}$ as:

$$\mathbf{D} = \begin{pmatrix} \mathbf{x}_1 - \mathbf{x}_2 \\ \mathbf{x}_1 - \mathbf{x}_3 \\ \vdots \\ \mathbf{x}_1 - \mathbf{x}_n \\ \mathbf{x}_2 - \mathbf{x}_1 \\ \mathbf{x}_2 - \mathbf{x}_3 \\ \vdots \\ \mathbf{x}_n - \mathbf{x}_{n-1} \end{pmatrix}.$$

Each row of $\mathbf{D}$ corresponds to an ordered pairwise difference $\mathbf{d}_{ij} = \mathbf{x}_i - \mathbf{x}_j$.

The following are the key mathematical properties of the matrix $\mathbf{D}$, which justify its use in the subsequent development of the pairwise difference covariance (PDC) estimator.

- **Dimensionality and Quadratic Expansion of Pairwise Directions:**

  The number of pairwise differences is

  $$m := n(n-1),$$

  since for each of the n observations, there are $n-1$ other observations to form ordered pairs with. As a result, the matrix $\mathbf{D}$ has dimension:

  $$\mathbf{D} \in \mathbb{R}^{n(n-1) \times p}.$$

  This quadratic increase in directional data leads to a denser representation of variability across the dataset, enhancing the information available for covariance estimation.

- **Directional and Asymmetric Encoding:**

  Each pairwise difference vector is defined as:

  $$\mathbf{d}_{ij} := \mathbf{x}_i - \mathbf{x}_j, \quad \text{for } i \neq j.$$

  The ordering of the pair (i, j) is preserved, meaning:

  $$\mathbf{d}_{ij} \neq \mathbf{d}_{ji}, \quad \text{and generally } \mathbf{d}_{ij} \neq -\mathbf{d}_{ij}.$$

  This asymmetry enables the encoding of more nuanced directional relationships among observations than symmetric formulations allow.

- **Translation Invariance:**

  The matrix $\mathbf{D}$ is invariant under global translation. Specifically, for any constant shift vector $\boldsymbol{\mu} \in \mathbb{R}^p$, applying the transformation:

  $$\mathbf{x}_i \mapsto \mathbf{x}_i + \boldsymbol{\mu}, \quad \forall i = 1, \ldots, n$$

  yields:

  $$\mathbf{d}_{ij} \mapsto (\mathbf{x}_i + \boldsymbol{\mu}) - (\mathbf{x}_j + \boldsymbol{\mu}) = \mathbf{x}_i - \mathbf{x}_j = \mathbf{d}_{ij}.$$

  This eliminates the need for explicit mean subtraction and ensures robustness to shifts in data location.

- **Independence from Sample Mean Estimation:**

  Unlike the classical sample covariance estimator, which depends on the sample mean:

  $$\bar{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i,$$

  the matrix $\mathbf{D}$ is constructed entirely from pairwise differences:

  $$\mathbf{D} = \{\mathbf{x}_i - \mathbf{x}_j\}_{i \neq j}.$$

  Therefore, it bypasses the variance and sensitivity introduced by estimating $\bar{\mathbf{x}}$, especially in HD or contaminated data settings.

- **Geometric Representation of the Data Cloud:**

  Each row $\mathbf{d}_{ij}$ in $\mathbf{D}$ represents a directed chord from $\mathbf{x}_j$ to $\mathbf{x}_i$ in $\mathbb{R}^p$. Collectively, the matrix encodes a complete set of directed edges between observations. This geometric richness enables the derivation of higher-order structures that reflect not only individual variability but also interactions among distinct observation pairs.

- **Effective Sample Size Augmentation:**

  Classical estimators utilize only n data points for estimating the covariance structure, while $\mathbf{D}$ leverages $n(n-1)$ directional relationships. This effective increase in sample size acts as a regularization mechanism that enhances estimator stability, especially when $n \ll p$, and facilitates more reliable estimation of principal directions in HD PCA.

While first-order pairwise differences of the form $\mathbf{d}_{ij} := \mathbf{x}_i - \mathbf{x}_j$ provide directional information between observations, they alone are insufficient for consistently estimating the population covariance matrix, particularly in HD regimes where $n \ll p$. A naive estimator based solely on these differences,

$$\widehat{\Sigma}_{\text{1st}} := \frac{1}{n(n-1)} \sum_{i \neq j} (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T,$$

does not coincide with the sample covariance matrix derived from the maximum likelihood principle. This discrepancy arises because the first-order differences are not independent, and their pairwise outer products introduce redundancy and bias, especially in the absence of proper centering.

Moreover, while the above expression aggregates the $n(n-1)$ pairwise terms superficially increasing the number of summands, it does not effectively increase the information content. The dependency among the difference vectors and their alignment within a low-dimensional subspace when $n \ll p$ can lead to ill-conditioned or singular covariance estimates. Consequently, the resulting eigenvalue spectrum may be distorted, yielding unstable principal components.

To address these challenges, the proposed Pairwise Difference Covariance (PDC) estimator incorporates *second-order pairwise differences* of the form

$$\mathbf{d}_{ij}\mathbf{d}_{ik}^T = (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_k)^T, \quad \text{for distinct } i, j, k.$$

To clarify the index roles used in this construction, we define:

- $i \in \{1, 2, \ldots, n\}$ is the index of the anchor observation.
- j and k are indices of observations distinct from i, i.e., $j, k \in \{1, 2, \ldots, n\} \setminus \{i\}$.
- For cross terms, only pairs with $j < k$ are used to avoid redundancy.

Using these definitions, we construct second-order pairwise differences:

$$d_{ij} := x_i - x_j, \quad d_{ik} := x_i - x_k,$$

and form the outer product $(x_i - x_j)(x_i - x_k)^T$, which captures directional interaction between the two difference vectors. These structured terms are organized into the local second-order matrices $D_i^{(1)}$ and $D_i^{(2)}$.

These structured outer products encode higher-order interactions among observations, capturing co-directional variability across multiple dimensions. By aggregating such second-order terms, the PDC estimator exploits the full geometry of the data cloud, balancing high-variance and low-variance directions more effectively.

In particular, this construction offers several advantages:

- It introduces implicit regularization through aggregation over multiple directions, improving spectral stability.
- It augments the effective rank of the covariance estimate in small-sample settings.
- It better approximates the true second-moment structure of the population distribution by incorporating cross-directional correlations.

Therefore, the use of second-order pairwise differences is a critical component of the PDC framework, enabling robust, translation-invariant, and geometrically informed covariance estimation suitable for HD and small-sample scenarios.

Step 3: Construction of Second-Order Pairwise Difference Matrices

To extend the information encoded by first-order differences, we construct structured second-order terms that capture higher-order interactions among observations. These second-order structures enable the estimation of a geometrically informed covariance matrix that remains stable and informative in HD regimes. The construction proceeds in the following steps.

Step 3.1: Local Second-Order Pairwise Index Construction

Fix an observation index $i \in \{1, 2, \ldots, n\}$. Define the index set of all other observations relative to i as:

$$\mathcal{I}_i := \{j \in \{1, \ldots, n\} \mid j \neq i\}.$$

Then the local set of first-order pairwise difference vectors centered at $\mathbf{x}_i$ is defined as:

$$\mathbf{D}_i := \{\mathbf{d}_{ij} := \mathbf{x}_i - \mathbf{x}_j \in \mathbb{R}^p \mid j \in \mathcal{I}_i\}, \quad \text{with } \#\mathbf{D}_i = n-1.$$

Notation clarification: In all expressions of the form $\mathbf{d}_{ij}$, the subscript j denotes the actual observation label x_j, not a position in a list or array. For example, $\mathbf{d}_{i(n-1)}$ refers to the difference between x_i and the $(n-1)$th observation in the dataset (i.e., x_{n-1}), not the $(n-1)$th entry in the set $\mathcal{I}_i$. This distinction avoids ambiguity when constructing matrices of pairwise differences.

We now construct all second-order vector pairs from $\mathbf{D}_i$, consisting of:

- **Diagonal pairs:** $(\mathbf{d}_{ij}, \mathbf{d}_{ij})$, for each $j \in \mathcal{I}_i$.
- **Cross pairs:** $(\mathbf{d}_{ij}, \mathbf{d}_{ik})$, for all distinct $j, k \in \mathcal{I}_i$ with $j < k$.

The total number of such second-order index pairs for a fixed i is:

$$m_i = (n-1) + \binom{n-1}{2} = \frac{n(n-1)}{2}.$$

Step 3.2: Construction of Local Second-Order Matrices $\mathbf{D}_i^{(1)}$ and $\mathbf{D}_i^{(2)}$

We now define the matrices $\mathbf{D}_i^{(1)} \in \mathbb{R}^{m_i \times p}$ and $\mathbf{D}_i^{(2)} \in \mathbb{R}^{m_i \times p}$, where each row encodes a second-order interaction originating from $\mathbf{x}_i$. Explicitly, these matrices are constructed by vertically stacking the appropriate vectors as follows:

(a) Diagonal pairs: For each $j \in \mathcal{I}_i$, define a row:

$$\mathbf{d}_{ij}^{(1)} := \mathbf{x}_i - \mathbf{x}_j, \quad \mathbf{d}_{ij}^{(2)} := \mathbf{x}_i - \mathbf{x}_j.$$

These entries fill the first $n-1$ rows of $\mathbf{D}_i^{(1)}$ and $\mathbf{D}_i^{(2)}$ respectively:

$$\mathbf{D}_i^{(1)}[j] = \mathbf{d}_{ij}^{(1)}, \quad \mathbf{D}_i^{(2)}[j] = \mathbf{d}_{ij}^{(2)}, \quad \text{for } j = 1, \ldots, n-1.$$

(b) Cross pairs: For each distinct pair (j, k) with $j < k$, define:

$$\mathbf{d}_{ijk}^{(1)} := \mathbf{x}_i - \mathbf{x}_j, \quad \mathbf{d}_{ijk}^{(2)} := \mathbf{x}_i - \mathbf{x}_k.$$

These cross terms populate the remaining $\binom{n-1}{2}$ rows of $\mathbf{D}_i^{(1)}$ and $\mathbf{D}_i^{(2)}$.

Thus, the full matrices take the form:

$$\mathbf{D}_i^{(1)} = \begin{pmatrix} \mathbf{d}_{ij_1}^{(1)} \\ \vdots \\ \mathbf{d}_{ij_{n-1}}^{(1)} \\ \mathbf{d}_{ij_1k_1}^{(1)} \\ \vdots \\ \mathbf{d}_{ij_{m'}}^{(1)} \end{pmatrix}, \quad \mathbf{D}_i^{(2)} = \begin{pmatrix} \mathbf{d}_{ij_1}^{(2)} \\ \vdots \\ \mathbf{d}_{ij_{n-1}}^{(2)} \\ \mathbf{d}_{ij_1k_1}^{(2)} \\ \vdots \\ \mathbf{d}_{ij_{m'}}^{(2)} \end{pmatrix},$$

where $m' = \binom{n-1}{2}$, and the row indices are arranged to reflect both diagonal and cross pair orderings.

Step 3.3: Aggregation into Global Second-Order Matrices

We now concatenate the local matrices $\mathbf{D}_i^{(1)}$ and $\mathbf{D}_i^{(2)}$ over all observations to construct the global second-order structures:

$$\mathbf{P}^{(1)} := \begin{pmatrix} \mathbf{D}_1^{(1)} \\ \mathbf{D}_2^{(1)} \\ \vdots \\ \mathbf{D}_n^{(1)} \end{pmatrix} \in \mathbb{R}^{m \times p}, \quad \mathbf{P}^{(2)} := \begin{pmatrix} \mathbf{D}_1^{(2)} \\ \mathbf{D}_2^{(2)} \\ \vdots \\ \mathbf{D}_n^{(2)} \end{pmatrix} \in \mathbb{R}^{m \times p},$$

where the total number of rows is:

$$m = \sum_{i=1}^{n} m_i = n \cdot \frac{n(n-1)}{2} = \frac{n^2(n-1)}{2}.$$

Each row-pair $(\mathbf{p}_\ell^{(1)}, \mathbf{p}_\ell^{(2)}) \in \mathbb{R}^{1 \times p}$ gives rise to a matrix:

$$\mathbf{H}_\ell := \mathbf{p}_\ell^{(1)} \left(\mathbf{p}_\ell^{(2)}\right)^T \in \mathbb{R}^{p \times p},$$

which encodes second-order geometric structure between difference vectors. These matrices $\{\mathbf{H}_\ell\}_{\ell=1}^m$ form the fundamental building blocks for the structured covariance estimator constructed in subsequent stages.

Step 4: Symmetrization of the Estimate

In the previous step, we defined the structured outer product estimator using second-order pairwise differences:

$$\mathbf{H} := \frac{1}{m} \sum_{\ell=1}^{m} \mathbf{p}_{\ell}^{(1)} \left(\mathbf{p}_{\ell}^{(2)} \right)^T,$$

where $\mathbf{p}_{\ell}^{(1)}, \mathbf{p}_{\ell}^{(2)} \in \mathbb{R}^p$ are the ℓ-th row vectors of the global matrices $\mathbf{P}^{(1)}$ and $\mathbf{P}^{(2)}$, respectively.

This matrix $\mathbf{H}$ is generally **not symmetric**, because the pairwise vectors $\mathbf{p}_{\ell}^{(1)}$ and $\mathbf{p}_{\ell}^{(2)}$ are not necessarily equal. That is,

$$\mathbf{H} \neq \mathbf{H}^T \quad \text{in general.}$$

Since a valid covariance matrix must be symmetric and preferably positive semi-definite, this lack of symmetry motivates the need for symmetrization.

Symmetrization Strategy

To restore symmetry and ensure mathematical validity of the estimator, we define:

$$\mathbf{Q} := \frac{1}{2} \left(\mathbf{H} + \mathbf{H}^T \right).$$

This matrix $\mathbf{Q} \in \mathbb{R}^{p \times p}$ is guaranteed to be symmetric by construction.

Proof of Symmetry

Proof. By the properties of matrix transposition:

$$\mathbf{Q}^T = \left(\frac{1}{2} (\mathbf{H} + \mathbf{H}^T) \right)^T = \frac{1}{2} (\mathbf{H}^T + \mathbf{H}) = \mathbf{Q}.$$

Therefore, $\mathbf{Q}$ is symmetric. □

Spectral and Structural Justification

While each matrix $\mathbf{H}_\ell := \mathbf{p}_{\ell}^{(1)} (\mathbf{p}_{\ell}^{(2)})^T$ is a rank-one (possibly non-symmetric) outer product, the resulting matrix $\mathbf{H}$ aggregates directional information that may not be aligned. The symmetrized form

$$\mathbf{Q} = \frac{1}{2m} \sum_{\ell=1}^{m} \left(\mathbf{p}_{\ell}^{(1)} (\mathbf{p}_{\ell}^{(2)})^T + \mathbf{p}_{\ell}^{(2)} (\mathbf{p}_{\ell}^{(1)})^T \right)$$

represents a weighted average of symmetric bilinear forms.

For any $\mathbf{z} \in \mathbb{R}^p$, the associated quadratic form is:

$$\mathbf{z}^T \mathbf{Q} \mathbf{z} = \frac{1}{m} \sum_{\ell=1}^{m} (\mathbf{z}^T \mathbf{p}_\ell^{(1)})(\mathbf{z}^T \mathbf{p}_\ell^{(2)}).$$

This is not necessarily non-negative unless $\mathbf{p}_\ell^{(1)} = \mathbf{p}_\ell^{(2)}$, but retains geometric interpretability and allows further regularization if needed.

Benefits of Symmetrization

The transformation $\mathbf{H} \mapsto \mathbf{Q}$ offers the following advantages:

- **Symmetry:** Guarantees $\mathbf{Q} = \mathbf{Q}^T$, a basic requirement for any covariance matrix.
- **Real eigenstructure:** Ensures that all eigenvalues are real, enabling eigendecomposition and projection-based techniques like PCA.
- **Structural consistency:** Harmonizes directional information from second-order pairwise interactions.

Thus, the matrix $\mathbf{Q}$ serves as the final symmetric second-order estimator, constructed from the outer product interactions among all local pairwise differences. It is now suitable for downstream statistical analysis and dimensionality reduction tasks.

Step 5: Final Normalized and Symmetric Estimator Construction

Having constructed the symmetric second-order matrix $\mathbf{Q} \in \mathbb{R}^{p \times p}$ in Step 4, we now complete the formulation of the Pairwise Difference Covariance (PDC) estimator by appropriately normalizing the symmetric sum. This ensures the estimator is unbiased and consistent in the asymptotic limit.

Step 5.1: Revisiting the Symmetric Matrix Q

Recall that the symmetric form $\mathbf{Q}$ was defined as:

$$\mathbf{Q} := \frac{1}{2} \left(\mathbf{H} + \mathbf{H}^T \right),$$

where

$$\mathbf{H} := \frac{1}{m} \sum_{\ell=1}^{m} \mathbf{p}_\ell^{(1)} \left(\mathbf{p}_\ell^{(2)} \right)^T = \frac{1}{m} (\mathbf{P}^{(1)})^T \mathbf{P}^{(2)}.$$

Hence, the symmetric matrix becomes:

$$\mathbf{Q} = \frac{1}{2m}\left[(\mathbf{P}^{(1)})^T\mathbf{P}^{(2)} + (\mathbf{P}^{(2)})^T\mathbf{P}^{(1)}\right].$$

Step 5.2: Final Normalization and Estimator Definition

We define the final estimator by identifying $\mathbf{Q}$ as the unnormalized symmetric aggregation of second-order terms. Thus, the final PDC estimator is:

$$\widehat{\Sigma}_{\text{PDC}} := \mathbf{Q} = \frac{1}{2m}\left[(\mathbf{P}^{(1)})^T\mathbf{P}^{(2)} + (\mathbf{P}^{(2)})^T\mathbf{P}^{(1)}\right].$$

Step 5.3: Substituting Total Number of Terms m

Recall from Step 3.3 that the total number of second-order pairs is:

$$m = \sum_{i=1}^{n} m_i = \frac{n^2(n-1)}{2}.$$

Plugging this into the estimator yields the final closed-form expression:

$$\widehat{\Sigma}_{\text{PDC}} = \frac{1}{n^2(n-1)}\left[(\mathbf{P}^{(1)})^T\mathbf{P}^{(2)} + (\mathbf{P}^{(2)})^T\mathbf{P}^{(1)}\right].$$

Step 5.4: Expression in Terms of Original Observations X

To express $\widehat{\Sigma}_{\text{PDC}}$ in terms of the original data matrix $\mathbf{X} \in \mathbb{R}^{n\times p}$, recall that $\mathbf{P}^{(1)}$ and $\mathbf{P}^{(2)}$ are built from local second-order pairwise differences:

For each observation $\mathbf{x}_i$, the local matrix of first-order differences is:

$$\mathbf{D}_i = \{\mathbf{d}_{ij} := \mathbf{x}_i - \mathbf{x}_j \in \mathbb{R}^p : j \in \mathcal{I}_i\},$$

where $\mathcal{I}_i = \{1, \ldots, n\} \setminus \{i\}$. We then construct the local second-order matrices:

$$\mathbf{D}_i^{(1)} = \text{stack of } \mathbf{d}_{ij} \text{ and } \mathbf{d}_{ij} \text{ for diagonal and cross terms,}$$
$$\mathbf{D}_i^{(2)} = \text{stack of } \mathbf{d}_{ij} \text{ and } \mathbf{d}_{ik} \text{ for diagonal and cross terms.}$$

These are concatenated across all i to form $\mathbf{P}^{(1)}$ and $\mathbf{P}^{(2)}$. Therefore, each row ℓ of these global matrices corresponds to a pair:

$$(\mathbf{p}_\ell^{(1)}, \mathbf{p}_\ell^{(2)}) = (\mathbf{x}_i - \mathbf{x}_j, \mathbf{x}_i - \mathbf{x}_k)$$

for some $i, j, k \in \{1, \ldots, n\}$, with $j, k \neq i$.

Consequently, the final PDC estimator becomes:

$$\widehat{\Sigma}_{\text{PDC}} = \frac{1}{n^2(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} \sum_{k \neq i, k > j} \left[(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_k)^T + (\mathbf{x}_i - \mathbf{x}_k)(\mathbf{x}_i - \mathbf{x}_j)^T \right].$$

This expression aggregates all second-order interactions between local pairwise directions relative to each base point $\mathbf{x}_i$, capturing both magnitude and angular structure in the data distribution.

4.2 Theoretical Properties of the PDC Estimator

Let $\widehat{\Sigma}_{\text{PDC}} \in \mathbb{R}^{p \times p}$ denote the Pairwise Difference Covariance estimator derived through structured second-order differences. We now rigorously state and prove the core mathematical properties that characterize its statistical behavior.

1. Translation Invariance The PDC estimator operates entirely on differences between observations. As such, it remains unaffected by global shifts in the dataset. Consider a translated dataset $\mathbf{X}' = \mathbf{X} + \mathbf{1}_n \mathbf{a}^T$, where $\mathbf{a} \in \mathbb{R}^p$ is an arbitrary translation vector. The second-order pairwise differences are computed as:

$$(\mathbf{x}'_i - \mathbf{x}'_j)(\mathbf{x}'_i - \mathbf{x}'_k)^T = (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_k)^T,$$

because the additive shift $\mathbf{a}$ cancels out. Therefore, the entire collection of terms contributing to $\widehat{\Sigma}_{\text{PDC}}$ is unchanged.

$$\Rightarrow \widehat{\Sigma}_{\text{PDC}}(\mathbf{X}') = \widehat{\Sigma}_{\text{PDC}}(\mathbf{X}).$$

The PDC estimator is translation invariant, which enhances robustness against location shifts.

2. Symmetry Each outer product in the PDC estimator involves asymmetric terms:

$$\mathbf{H}_\ell = \mathbf{p}_\ell^{(1)} (\mathbf{p}_\ell^{(2)})^T.$$

However, to ensure symmetry, a requirement for valid covariance matrices, we symmetrize the final estimator:

$$\widehat{\Sigma}_{\text{PDC}} = \frac{1}{2m} \sum_{\ell=1}^{m} \left[\mathbf{p}_\ell^{(1)} (\mathbf{p}_\ell^{(2)})^T + \mathbf{p}_\ell^{(2)} (\mathbf{p}_\ell^{(1)})^T \right].$$

Each summand becomes its own transpose, ensuring the full matrix satisfies:

$$\widehat{\Sigma}_{\text{PDC}} = \widehat{\Sigma}_{\text{PDC}}^T.$$

The PDC estimator is symmetric by construction, making it suitable for eigendecomposition and interpretation as a covariance matrix.

3. Positive Definiteness Both the sample covariance matrix (MLE) and the Pairwise Difference Covariance (PDC) estimator are guaranteed to be positive semi-definite (PSD) under standard construction. The MLE is defined as:

$$\widehat{\Sigma}_{\text{MLE}} = \frac{1}{n-1} \sum_{i=1}^{n} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T,$$

which is a weighted sum of outer products of centered vectors. Each term $(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T$ is a rank-one PSD matrix, and a sum of such matrices remains PSD.

Similarly, the PDC estimator is built from symmetrized outer products of pairwise differences:

$$\widehat{\Sigma}_{\text{PDC}} = \frac{1}{2m} \sum_{\ell=1}^{m} \left(\mathbf{p}_\ell^{(1)} (\mathbf{p}_\ell^{(2)})^T + \mathbf{p}_\ell^{(2)} (\mathbf{p}_\ell^{(1)})^T \right),$$

where each symmetrized term can be interpreted as a rank-two symmetric matrix. Since the sum includes all such combinations across anchors, the full matrix is symmetric, and under finite variance assumptions, it is PSD.

Although both estimators are PSD, PDC offers notable advantages over MLE in the $n < p$ regime. The MLE has rank at most $n - 1$, which means it becomes singular (i.e., non-invertible) when $n \leq p$, and its trailing eigenvalues are exactly zero. This leads to poor performance in applications like PCA and discriminant analysis.

In contrast, PDC aggregates over $m = \frac{n^2(n-1)}{2}$ second-order terms. Provided that the resulting matrix has rank at least p, the PDC estimator can be full-rank even when $n < p$, avoiding the degeneracy of MLE. Additionally, PDC incorporates information from structured pairwise interactions, which effectively increases the number of informative directions. The improved numerical stability and reduced estimation error can be expressed via:

$$\text{Var}(\widehat{\Sigma}_{\text{PDC}}) \approx \frac{1}{m} \ll \frac{1}{n}.$$

This variance reduction leads to tighter concentration around the true covariance matrix, enhancing both the magnitude and directional accuracy of PCA estimates. Therefore, while both estimators maintain PSD properties, the PDC achieves better practical utility in HD, small-sample scenarios through higher effective sample size and more robust spectral properties.

Remarks on Rank and Positive Definiteness: The matrix $\widehat{\Sigma}_{\text{PDC}}$ inherits its rank from the pairwise difference matrix $\mathbf{D}$, which in turn depends on the data matrix $\mathbf{X}$. If the set of observations $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \subset \mathbb{R}^p$ does not span the full space $\mathbb{R}^p$, then the pairwise differences may also lie in a lower-dimensional subspace, potentially making $\widehat{\Sigma}_{\text{PDC}}$ rank-deficient.

Formally, we require that $\text{rank}(\mathbf{X}) = p$ to ensure that the space spanned by the pairwise differences also has full rank. In high-dimensional scenarios where $n < p$, the data matrix $\mathbf{X}$ typically does not span $\mathbb{R}^p$, and thus $\widehat{\Sigma}_{\text{PDC}}$ will not be of full rank. However, the PDC construction ensures that all eigenvalues are non-negative, and the matrix remains symmetric and positive semi-definite.

Therefore, while the PDC estimator regularizes the eigenspectrum and reduces spectral overdispersion, its ability to yield a full-rank covariance matrix depends on the dimensional coverage of the original dataset. This condition should be taken into account in applications where matrix invertibility is critical.

4. Bias and Consistency of $\widehat{\Sigma}_{\text{PDC}}$ Let $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \subset \mathbb{R}^p$ be a collection of i.i.d. observations with zero mean $\mathbb{E}[\mathbf{x}_i] = \mathbf{0}$ and population covariance matrix $\Sigma = \mathbb{E}[\mathbf{x}_i \mathbf{x}_i^T]$. The Pairwise Difference Covariance (PDC) estimator is given by:

$$\widehat{\Sigma}_{\text{PDC}} = \frac{1}{n^2(n-1)} \left[\left(\mathbf{P}^{(1)} \right)^T \mathbf{P}^{(2)} + \left(\mathbf{P}^{(2)} \right)^T \mathbf{P}^{(1)} \right],$$

where $\mathbf{P}^{(1)}$ and $\mathbf{P}^{(2)}$ contain structured second-order pairwise differences anchored at each observation.

Each second-order term has the expectation:

$$\mathbb{E}[(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_k)^T] = \begin{cases} 2\Sigma, & \text{if } j = k, \\ \Sigma, & \text{if } j \neq k. \end{cases}$$

Let $t = n - 1$ denote the number of diagonal terms and $c = \binom{n-1}{2} = \frac{(n-1)(n-2)}{2}$ the number of cross terms per anchor. The total second-order terms per anchor is $m_i = \frac{n(n-1)}{2}$, and the expected contribution from anchor i is:

$$\mathbb{E}\left[\frac{1}{m_i}\left(t \cdot 2\Sigma + c \cdot \Sigma\right)\right] = \left(\frac{n+2}{n}\right)\Sigma.$$

Thus, the expectation of the full PDC estimator becomes:

$$\mathbb{E}[\widehat{\Sigma}_{\text{PDC}}] = \left(1 + \frac{2}{n}\right)\Sigma = \Sigma + \delta(n), \quad \text{where} \quad \delta(n) = \frac{2}{n}\Sigma.$$

As the sample size increases, the bias term vanishes:

$$\lim_{n\to\infty} \delta(n) = \mathbf{0}.$$

Moreover, the estimator satisfies the convergence in probability condition:

$$\lim_{n\to\infty} \mathbb{P}\left(\left\|\widehat{\Sigma}_{\text{PDC}} - \Sigma\right\| > \varepsilon\right) = 0, \quad \forall \varepsilon > 0,$$

which confirms that $\widehat{\Sigma}_{\text{PDC}} \xrightarrow{p} \Sigma$. This establishes the asymptotic consistency of the PDC estimator under standard moment assumptions.

In summary, although the PDC estimator is asymptotically proportional to Σ rather than exactly unbiased, the finite-sample bias $\delta(n) = \frac{2}{n}\Sigma$ is analytically tractable and diminishes rapidly with increasing n.

This makes PDC a compelling alternative to the classical sample covariance from both theoretical and computational perspectives. Its regularization-by-construction framework, full-rank output, and translation invariance make it particularly advantageous when $n < p$ or $n \ll p$. Furthermore, in contemporary statistical practice where high-resolution computation is standard, even modest numerical improvements in estimation accuracy, such as those offered by PDC, can yield meaningful benefits in downstream tasks such as PCA, clustering, or discriminant analysis.

5. Approximate Unbiasedness under Gaussianity Assume that the observations $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$ are i.i.d. samples from the multivariate normal distribution $\mathcal{N}(\mu, \Sigma)$, and we consider the mean-centered case $\mu = \mathbf{0}$ without loss of generality.

We begin by recalling a classical result: for any pair $i \neq j$,

$$\mathbb{E}[(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T] = 2\Sigma.$$

This identity leads to the expectation of the first-order pairwise difference estimator,

$$\hat{\Sigma}_{\text{PD}} = \frac{1}{n(n-1)} \sum_{i \neq j} (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T,$$

as:

$$\mathbb{E}[\hat{\Sigma}_{\text{PD}}] = 2\Sigma,$$

which indicates that this estimator overestimates the covariance by a constant factor and is unbiased only after dividing by 2.

In contrast, the Pairwise Difference Covariance (PDC) estimator is defined through second-order pairwise products:

$$\widehat{\Sigma}_{\text{PDC}} = \frac{1}{2m} \sum_{\ell=1}^{m} \left(\mathbf{p}_\ell^{(1)} (\mathbf{p}_\ell^{(2)})^T + \mathbf{p}_\ell^{(2)} (\mathbf{p}_\ell^{(1)})^T \right),$$

where each pair $(\mathbf{p}_\ell^{(1)}, \mathbf{p}_\ell^{(2)}) \in \mathbb{R}^p \times \mathbb{R}^p$ is formed from differences of the form:

$$\mathbf{p}_\ell^{(1)} = \mathbf{x}_i - \mathbf{x}_j, \quad \mathbf{p}_\ell^{(2)} = \mathbf{x}_i - \mathbf{x}_k, \quad j, k \neq i.$$

Let us now analyze the expectation of such second-order terms under Gaussianity.

Case 1: Diagonal Term $(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T)$

$$\mathbb{E}[(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T] = 2\Sigma.$$

Case 2: Cross Term $(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_k)^T)$ **with** $j \neq k$

$$\mathbb{E}[(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_k)^T] = \Sigma.$$

This follows since the observations are independent with zero mean, and the mixed terms vanish. The averaging in all terms reduces the effective weight of 2Σ. Let us denote

- Number of diagonal terms per anchor: $t = n - 1$
- Number of cross terms per anchor: $c = \frac{(n-1)(n-2)}{2}$
- Total number of second-order terms per anchor: $m_i = t + c = \frac{n(n-1)}{2}$

Then the total expected second-order contribution for one anchor becomes:

$$\mathbb{E}[\text{Sum}_i] = t \cdot 2\Sigma + c \cdot \Sigma = (2t + c)\Sigma.$$

Thus, the average contribution per second-order term is:

$$\mathbb{E}\left[\frac{\text{Sum}_i}{m_i}\right] = \frac{2t + c}{m_i}\Sigma.$$

Now substitute:

$$2t + c = 2(n-1) + \frac{(n-1)(n-2)}{2}, \quad m_i = \frac{n(n-1)}{2}.$$

After simplification:

$$\frac{2(n-1) + \frac{(n-1)(n-2)}{2}}{\frac{n(n-1)}{2}} = \frac{4 + (n-2)}{n} = \frac{n+2}{n}.$$

So we get:

$$\mathbb{E}[\widehat{\Sigma}_{\text{PDC}}] = \Sigma + \delta(n, p).$$

The estimator is not exactly unbiased; it includes a small bias term $\delta(n, p)$ that depends on n. However, as $n \to \infty$, this term vanishes:

$$\lim_{n\to\infty} \delta(n, p) = 0.$$

The PDC estimator is asymptotically unbiased under Gaussianity. In small-sample settings, it may slightly overestimate the true covariance structure. Importantly, this small bias introduces a regularizing effect that helps counteract overfitting and instability in HD settings where $n \ll p$. In such scenarios, principal component directions obtained from PDC tend to be more stable, as the inflated variance directions are gently shrunk, leading to a more balanced and interpretable eigenspectrum in PCA applications.

6. Variance Reduction The Pairwise Difference Covariance (PDC) estimator incorporates a substantially larger number of second-order terms compared to the classical Maximum Likelihood Estimator (MLE). This design directly impacts the variance of the estimator. Specifically, PDC is constructed as an average over

$$m = \frac{n^2(n-1)}{2}$$

symmetric second-order products:

$$\widehat{\Sigma}_{\mathrm{PDC}} = \frac{1}{2m} \sum_{\ell=1}^{m} \left(\mathbf{p}_\ell^{(1)} (\mathbf{p}_\ell^{(2)})^T + \mathbf{p}_\ell^{(2)} (\mathbf{p}_\ell^{(1)})^T \right).$$

Assuming the terms $\mathbf{p}_\ell^{(1)}$ and $\mathbf{p}_\ell^{(2)}$ are weakly dependent and have bounded second moments, the variance of the average can be approximated using the linearity of the variance operator and assuming approximate independence:

$$\mathrm{Var}(\widehat{\Sigma}_{\mathrm{PDC}}) \approx \frac{1}{4m^2} \sum_{\ell=1}^{m} \mathrm{Var} \left(\mathbf{p}_\ell^{(1)} (\mathbf{p}_\ell^{(2)})^T + \mathbf{p}_\ell^{(2)} (\mathbf{p}_\ell^{(1)})^T \right).$$

Since each term is symmetric and bilinear, its variance remains bounded and of the same order as a standard outer product of centered random vectors. Denoting this generic variance as $\mathbb{V}_2$, we get:

$$\mathrm{Var}(\widehat{\Sigma}_{\mathrm{PDC}}) = \frac{1}{m} \mathbb{V}_2.$$

By contrast, the MLE estimator is constructed from n centered rank-one matrices:

$$\widehat{\Sigma}_{\mathrm{MLE}} = \frac{1}{n} \sum_{i=1}^{n} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T,$$

with variance:

$$\mathrm{Var}(\widehat{\Sigma}_{\mathrm{MLE}}) = \frac{1}{n} \mathbb{V}_1,$$

where $\mathbb{V}_1$ represents the variance of the rank-one summands in the MLE estimator.

Assuming that $\mathbb{V}_1 \approx \mathbb{V}_2$, we have:

$$\frac{\mathrm{Var}(\widehat{\Sigma}_{\mathrm{PDC}})}{\mathrm{Var}(\widehat{\Sigma}_{\mathrm{MLE}})} = \frac{\mathbb{V}_2/m}{\mathbb{V}_1/n} = \frac{n}{m} \ll 1.$$

Since $m = \frac{n^2(n-1)}{2} \gg n$, the variance of $\widehat{\Sigma}_{\mathrm{PDC}}$ decreases much faster as n increases. This shows that PDC achieves a tighter concentration around its expected value than MLE, especially in small-sample or HD regimes.

The effective averaging over a much larger set of structured second-order interactions in PDC induces a substantial reduction in estimator variance. This results in improved precision and stability of the estimator in practice, particularly under HD conditions.

7. Spectral Regularization Let the Pairwise Difference Covariance (PDC) estimator be expressed as a modified version of the classical Maximum Likelihood Estimator (MLE):

$$\widehat{\Sigma}_{\text{PDC}} = \widehat{\Sigma}_{\text{MLE}} + \Delta,$$

where $\Delta \in \mathbb{R}^{p\times p}$ is the adjustment matrix introduced through the use of second-order pairwise differences.

Assume that $\widehat{\Sigma}_{\text{MLE}}$ has the spectral decomposition:

$$\widehat{\Sigma}_{\text{MLE}} = \sum_{k=1}^{p} \lambda_k \mathbf{v}_k \mathbf{v}_k^T,$$

with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ and corresponding orthonormal eigenvectors $\{\mathbf{v}_k\}_{k=1}^{p}$.

The eigenvalues of $\widehat{\Sigma}_{\text{PDC}}$ along each direction $\mathbf{v}_k$ can be expressed using the Rayleigh quotient:

$$\tilde{\lambda}_k = \mathbf{v}_k^T \widehat{\Sigma}_{\text{PDC}} \mathbf{v}_k = \mathbf{v}_k^T (\widehat{\Sigma}_{\text{MLE}} + \Delta)\mathbf{v}_k = \lambda_k + \mathbf{v}_k^T \Delta \mathbf{v}_k.$$

In this expression, $\mathbf{v}_k^T \Delta \mathbf{v}_k$ determines how the eigenvalue changes. Due to the way Δ is formed from second-order pairwise interactions, it typically reduces the largest eigenvalues (which are often influenced by noise or outliers) and increases the smallest ones (which may capture subtle but meaningful structure). Mathematically, we often observe:

$$\mathbf{v}_1^T \Delta \mathbf{v}_1 < 0, \quad \mathbf{v}_p^T \Delta \mathbf{v}_p > 0,$$

which leads to the following effect:

$$\tilde{\lambda}_1 < \lambda_1, \quad \tilde{\lambda}_p > \lambda_p.$$

This means that the PDC estimator smooths the spectrum by decreasing the gap between the largest and smallest eigenvalues. As a result, variance is more evenly distributed among principal directions, avoiding dominance by a few and giving better visibility to the rest. This effect is especially beneficial in HD Principal Component Analysis (PCA), where overestimation of a few directions is a common problem.

The PDC estimator performs built-in spectral regularization. By modifying the covariance structure through structured second-order differences, it stabilizes the eigenvalue spread and improves the reliability of PCA in HD applications.

4.2.1 Theoretical Comparison with Sample Covariance

Lemma 1 (Improved Spectral Properties of PDC Compared to the Sample Covariance Estimator)**.** *Let $\widehat{\Sigma}_{MLE}$ denote the sample covariance matrix and $\widehat{\Sigma}_{PDC}$ the Pairwise Difference Covariance estimator based on second-order differences. Consider the spiked covariance model where the true covariance matrix Σ has a low-rank signal structure and $\|\Sigma\|_2 < \infty$. Assume that the dimensionality p and sample size n grow together such that $p/n \to c \in (0, \infty)$. Then the following properties hold:*

(i) ***Spectral Shrinkage:***

$$\mathbb{E}[\lambda_{\max}(\widehat{\Sigma}_{PDC})] < \mathbb{E}[\lambda_{\max}(\widehat{\Sigma}_{MLE})],$$

indicating that the leading eigenvalue of the PDC estimator is systematically smaller than that of the MLE, which tends to be upward biased in high dimensions.

(ii) ***Improved Subspace Recovery:*** *Let Π_k^* denote the projection onto the subspace spanned by the top-k eigenvectors of the true covariance matrix Σ, and let Π_k^{PDC}, Π_k^{MLE} be the projections based on the PDC and MLE estimators respectively. Then:*

$$\mathbb{E}\left[\|\Pi_k^{PDC} - \Pi_k^*\|_F^2\right] < \mathbb{E}\left[\|\Pi_k^{MLE} - \Pi_k^*\|_F^2\right],$$

which implies that the PDC estimator more accurately recovers the principal subspace.

Proof. Define the correction matrix:

$$\Delta := \widehat{\Sigma}_{\text{PDC}} - \widehat{\Sigma}_{\text{MLE}}.$$

Let the eigendecomposition of $\widehat{\Sigma}_{\text{MLE}}$ be:

$$\widehat{\Sigma}_{\text{MLE}} = V \Lambda V^T, \quad \Lambda = \text{diag}(\lambda_1, \ldots, \lambda_p), \quad V = [v_1, \ldots, v_p].$$

Then we can write:

$$\widehat{\Sigma}_{\text{PDC}} = V \widetilde{\Lambda} V^T, \quad \widetilde{\Lambda} = \text{diag}(\tilde{\lambda}_1, \ldots, \tilde{\lambda}_p),$$

and the perturbed eigenvalues satisfy:

$$\tilde{\lambda}_k = \lambda_k + v_k^T \Delta v_k.$$

Compression of Dominant Eigenvalues: For large λ_k, empirical evidence and theoretical analysis show:

$$v_k^T \Delta v_k < 0 \quad \Rightarrow \quad \tilde{\lambda}_k < \lambda_k,$$

i.e., PDC shrinks dominant components.

Inflation of Weak Components: For small $\lambda_j \approx 0$,

$$v_j^T \Delta v_j > 0 \quad \Rightarrow \quad \tilde{\lambda}_j > \lambda_j,$$

capturing additional variance in noise-dominated directions.

Spectral Perturbation Expansion: From standard matrix perturbation theory,

$$\tilde{\lambda}_k = \lambda_k + v_k^T \Delta v_k + \mathcal{O}(\|\Delta\|^2).$$

Trace Conservation: Assuming mean-centered data,

$$\operatorname{tr}(\Delta) = \sum_{k=1}^{p} (\tilde{\lambda}_k - \lambda_k) = 0,$$

so PDC only redistributes variance across directions.

Variance Redistribution: PDC compresses the spectrum:

$$\tilde{\lambda}_1 < \lambda_1, \quad \tilde{\lambda}_r \approx \lambda_r, \quad \tilde{\lambda}_p > \lambda_p.$$

Rank Enhancement: MLE satisfies $\operatorname{rank}(\widehat{\Sigma}_{\text{MLE}}) \leq n-1$, while

$$\operatorname{rank}(\widehat{\Sigma}_{\text{PDC}}) \leq \min(p, m), \quad m = \frac{n^2(n-1)}{2},$$

leading to nearly full-rank output when $n \ll p$.

Operator Norm Bound: Because PDC aggregates over a larger number of terms:

$$\mathbb{E}\left[\|\widehat{\Sigma}_{\text{PDC}} - \Sigma\|_2\right] = \mathcal{O}\left(\sqrt{\frac{p}{m}}\right), \quad \text{while } \mathbb{E}\left[\|\widehat{\Sigma}_{\text{MLE}} - \Sigma\|_2\right] = \mathcal{O}\left(\sqrt{\frac{p}{n}}\right),$$

hence PDC achieves better concentration.

Improved Subspace Recovery: From the Davis–Kahan theorem, the top-k eigenspace perturbation satisfies:

$$\|\Pi_k^{\text{PDC}} - \Pi_k^*\|_F \leq \frac{2\sqrt{2}}{\delta_k}\|\widehat{\Sigma}_{\text{PDC}} - \Sigma\|_2,$$

where $\delta_k = \lambda_k - \lambda_{k+1}$. Since $\|\widehat{\Sigma}_{\text{PDC}} - \Sigma\|_2 < \|\widehat{\Sigma}_{\text{MLE}} - \Sigma\|_2$, this confirms improved subspace estimation. □

4.3 Justification for PDC Regularization in PCA

In HD statistical inference, Principal Component Analysis (PCA) depends fundamentally on the quality of the estimated covariance matrix. The classical sample covariance matrix, or Maximum Likelihood Estimator (MLE), exhibits critical limitations when the number of variables p exceeds the number of observations n. Specifically, the rank of the MLE is at most $n-1$, implying that the smallest $p - n + 1$ eigenvalues are exactly zero. This rank deficiency leads to singularity, inflated variance directions, and instability in PCA, where the principal components corresponding to small eigenvalues may be driven by sampling noise rather than meaningful structure.

The Pairwise Difference Covariance (PDC) estimator addresses these limitations through a second-order structure that leverages all pairwise differences among observations. Formally, the PDC estimator aggregates

$$m = \frac{n^2(n-1)}{2}$$

second-order terms, resulting in a substantial increase in the number of informative directions. This boosts the effective rank of the estimator well beyond the sample size. When $m \geq p$, and the pairwise difference matrices $\mathbf{P}^{(1)}$ and $\mathbf{P}^{(2)}$ span $\mathbb{R}^p$, the PDC estimator achieves full rank, even when $n < p$, thus eliminating the zero-eigenvalue problem that limits the MLE in small-sample, HD scenarios.

Another central contribution of PDC is its built-in spectral regularization. It modifies the eigenvalue spectrum of the MLE through the decomposition

$$\hat{\Sigma}_{\text{PDC}} = \hat{\Sigma}_{\text{MLE}} + \Delta,$$

where Δ is a symmetric, data-dependent correction matrix. The eigenvalues of the PDC estimator satisfy

$$\tilde{\lambda}_k = \lambda_k + v_k^T \Delta v_k,$$

where v_k is the k-th eigenvector of $\hat{\Sigma}_{\text{MLE}}$. In high-variance directions (i.e., large λ_k), the term $v_k^T \Delta v_k$ tends to be negative, inducing shrinkage. Conversely, in low-variance directions, particularly where $\lambda_k = 0$ under MLE, PDC introduces positive corrections that reveal previously hidden or underrepresented structure.

Unlike scalar shrinkage estimators such as Ledoit–Wolf, which apply uniform contraction toward the identity matrix, PDC performs directional, non-scalar regularization. Its correction matrix Δ is constructed from structured second-order geometry and is not constrained to any fixed shrinkage target. Moreover, the operator norm of the correction satisfies

$$\|\Delta\|_2 = \mathcal{O}\left(\frac{p}{n}\right),$$

highlighting that the regularization strength is naturally stronger in HD, low-sample settings and vanishes asymptotically as $n \to \infty$. Hence, PDC converges to the MLE in large-sample regimes, maintaining consistency while enhancing stability in finite samples.

From a PCA perspective, these properties offer distinct advantages. The redistribution of eigenvalues produces a flatter and more stable spectrum. Shrinking the dominant eigenvalues mitigates overfitting by preventing any single direction from dominating the explained variance. Simultaneously, lifting near-zero eigenvalues enables the identification of subtle patterns that would otherwise be obscured under MLE. This improves effective rank, enhances robustness against noise, and enables more reliable subspace estimation, especially when used in conjunction with dimensionality reduction or clustering techniques.

In summary, the PDC estimator is not merely a modification of MLE but a principled and adaptive regularization strategy. It preserves the core structure of the data, addresses the zero-eigenvalue issue in $n < p$ scenarios, and provides more stable eigendirections for PCA. These properties make it a compelling choice for modern HD applications where classical covariance estimators fail to provide sufficient resolution and robustness.

Note (1): In scenarios where the number of observations equals or exceeds the number of variables ($n = p$ or $n > p$), the Pairwise Difference Covariance (PDC) estimator demonstrates behavior similar to that of the classical maximum likelihood estimator (MLE). The second-order pairwise aggregation in PDC captures a sufficient number of directional variations to estimate the covariance reliably, avoiding the rank deficiency issues seen in small-sample regimes. As the sample size increases, the benefits of PDC become less pronounced, and the estimator converges asymptotically to MLE. Although PDC inherently introduces mild spectral regularization, such as flattening the

eigenvalue spectrum and reducing variance dominance, these effects are numerically small in large-sample settings. Therefore, while PDC is theoretically sound and remains consistent as $n \to \infty$, its most impactful advantages are observed in HD settings with $n < p$ or $n \ll p$, where it helps stabilize estimation by leveraging data-driven aggregation without requiring mean centering or external tuning.

Note (2): Remark on the Impact of q in PCA. The performance of PCA often depends on the choice of q, the number of principal components retained. In the context of the PDC estimator, the influence of q is mediated by how accurately the estimator captures the eigenspectrum and associated eigenspaces. Since the PDC method redistributes variance by shrinking dominant eigenvalues and inflating smaller ones, it improves the separation and ordering of components across the spectrum. When q is small (e.g., $q = 1$ or 2), the PDC estimator helps prevent overemphasis on noisy or overly dominant directions by suppressing eigenvalue inflation, leading to more reliable extraction of meaningful PCs. For larger q, the effective rank enhancement of PDC ensures that weaker signal directions are not discarded due to artificial eigenvalue collapse, thereby enabling more stable subspace estimation. These benefits are especially visible in high-dimensional settings where classical estimators suffer from severe eigenvalue bias and inconsistency of eigenvectors. Overall, PDC supports consistent and robust component selection across a range of q, particularly when the underlying signal is diffuse or the eigengap is small.

Before introducing the regularized extensions of the PDC estimator, it is important to note that each of the forthcoming variants modifies the structure of pairwise differences in a unique manner. These structural modifications result in different covariance matrices, which in turn lead to distinct eigenvalue spectra and, in many cases, different principal component directions. Therefore, the choice of regularization not only impacts the scale and stability of the eigenvalues but also influences the interpretation of principal components derived from each estimator.

4.4 Regularized PDC Estimations

While the Pairwise Difference Covariance (PDC) estimator provides a natural form of spectral regularization and rank expansion, its performance can be further enhanced through systematic scaling and adjustment of the pairwise difference structure. Especially in HD scenarios where $n < p$, careful modification of the second-order pairwise terms can improve the stability, conditioning, and interpretability of the covariance estimate, thereby strengthening its effectiveness for downstream tasks like PCA.

Motivated by the need to refine the regularization effects inherent in the PDC framework, we introduce four alternative variants of the PDC estimator, each designed to adjust the scale or contribution of pairwise differences in different ways. These proposed regularization strategies aim to enhance the estimator's adaptability across varying data conditions without sacrificing the core geometric advantages of the PDC approach.

The four regularized PDC estimators considered in this study are:

- **Standardized Pairwise Differences Covariance (SPDC):** Scaling each pairwise difference vector by its local standard deviation to equalize the contribution across all directions.
- **Local Scaled Pairwise Differences Covariance (LSPDC):** Applying local scaling based on neighborhood-specific variability to emphasize local structure.
- **Scaled by Maximum Absolute Value Pairwise Differences Covariance (MAX-PDC):** Normalizing each pairwise difference by its maximum absolute component to reduce the influence of extreme values.
- **Scaled by Range Pairwise Differences Covariance (RPDC):** Adjusting pairwise differences using the range of components, promoting balance across dimensions.

These variants are presented as flexible extensions to the PDC methodology, offering additional control over regularization strength and directional scaling while preserving the key advantages of the second-order difference-based formulation.

4.4.1 Standardized Pairwise Differences Covariance (SPDC)

The Standardized Pairwise Differences Covariance (SPDC) estimator enhances the original Pairwise Difference Covariance (PDC) approach by incorporating a normalization step into the pairwise difference structure. This standardization aims to regularize the scale of the pairwise differences, ensuring that each term contributes comparably to the covariance estimation, regardless of variability across directions. By adjusting for the mean and variance of the pairwise differences before aggregation, SPDC reduces the influence of scale heterogeneity and prevents domination by high-variance directions.

Formally, the modification introduced in SPDC replaces each raw pairwise difference vector $\mathbf{d}_{ij} = \mathbf{x}_i - \mathbf{x}_j$ with its standardized version:

$$\widetilde{\mathbf{d}}_{ij} = \frac{\mathbf{d}_{ij} - \mu_{\mathbf{d}}}{\sigma_{\mathbf{d}}},$$

where $\mu_{\mathbf{d}} \in \mathbb{R}^p$ and $\sigma_{\mathbf{d}} \in \mathbb{R}^p$ denote the element-wise mean and standard deviation of all pairwise differences:

$$\mu_{\mathbf{d}} = \frac{1}{n(n-1)} \sum_{i \neq j} \mathbf{d}_{ij}, \quad \sigma_{\mathbf{d}} = \sqrt{\frac{1}{n(n-1)} \sum_{i \neq j} (\mathbf{d}_{ij} - \mu_{\mathbf{d}})^2}.$$

These standardized differences $\widetilde{\mathbf{d}}_{ij}$ replace the raw differences in the second-order aggregation formula of the PDC estimator. Specifically, the pairwise terms $\mathbf{p}_\ell^{(1)}, \mathbf{p}_\ell^{(2)}$ in the original formulation are substituted by their standardized counterparts:

$$\widetilde{\mathbf{p}}_\ell^{(1)} = \frac{\mathbf{p}_\ell^{(1)} - \mu_{\mathbf{d}}}{\sigma_{\mathbf{d}}}, \quad \widetilde{\mathbf{p}}_\ell^{(2)} = \frac{\mathbf{p}_\ell^{(2)} - \mu_{\mathbf{d}}}{\sigma_{\mathbf{d}}}.$$

The final SPDC estimator is thus expressed as:

$$\widehat{\Sigma}_{\text{SPDC}} = \frac{1}{2m} \sum_{\ell=1}^{m} \left(\widetilde{\mathbf{p}}_\ell^{(1)} (\widetilde{\mathbf{p}}_\ell^{(2)})^T + \widetilde{\mathbf{p}}_\ell^{(2)} (\widetilde{\mathbf{p}}_\ell^{(1)})^T \right),$$

where $m = \frac{n^2(n-1)}{2}$ is the total number of second-order pairs.

This standardization approach maintains the structural integrity of the PDC estimator while regulating the contribution from each direction, effectively stabilizing the covariance estimation process, particularly under HD settings or data with heterogeneous scales.

Spectral Adjustment of SPDC Relative to PDC

The regularization introduced through standardization in SPDC directly impacts the eigenvalue spectrum of the covariance estimate, providing substantial benefits for Principal Component Analysis (PCA). In particular, SPDC shrinks overly large eigenvalues and inflates small or near-zero eigenvalues, promoting a balanced and well-conditioned eigenspectrum.

Let the eigendecomposition of the PDC estimator be given by:

$$\widehat{\Sigma}_{\text{PDC}} = V \Lambda^{\text{PDC}} V^T,$$

where $V = [v_1, v_2, \ldots, v_p]$ represents the orthonormal eigenvectors, and $\Lambda^{\text{PDC}} = \text{diag}(\tilde{\lambda}_1^{\text{PDC}}, \ldots, \tilde{\lambda}_p^{\text{PDC}})$ contains the corresponding eigenvalues.

In the PDC framework, the variance contribution along a direction v_k is:

$$\mathbb{E}\left[(v_k^T \mathbf{d}_{ij})^2 \right] = \sigma_k^2,$$

where σ_k^2 reflects the directional variance along the eigenvector v_k, with typically larger values for dominant components and smaller values for weaker directions.

By contrast, in SPDC, the standardized differences satisfy:

$$\mathbb{E}\left[(v_k^T \widetilde{\mathbf{d}}_{ij})^2\right] = 1, \quad \forall\, k = 1, \ldots, p,$$

ensuring that all directions contribute equally in expectation. This normalization leads to the relationship between SPDC and PDC eigenvalues:

$$\tilde{\lambda}_k^{\mathrm{SPDC}} = \frac{1}{\sigma_k^2} \tilde{\lambda}_k^{\mathrm{PDC}}.$$

For dominant directions (large σ_k^2), this implies:

$$\tilde{\lambda}_k^{\mathrm{SPDC}} < \tilde{\lambda}_k^{\mathrm{PDC}},$$

indicating eigenvalue shrinkage, while for weak directions (small σ_k^2):

$$\tilde{\lambda}_k^{\mathrm{SPDC}} > \tilde{\lambda}_k^{\mathrm{PDC}},$$

demonstrating eigenvalue inflation.

Implications for PCA. The spectral adjustment introduced by SPDC directly influences the effectiveness of PCA by improving both the magnitude stability of the leading principal components and the accuracy of their associated directions. Mathematically, the eigenvalues in SPDC relate to those in PDC by the rescaling factor $1/\sigma_k^2$, where σ_k^2 is the variance along the eigenvector v_k. This implies that for dominant directions where σ_k^2 is large, the SPDC eigenvalues are proportionally shrunk:

$$\tilde{\lambda}_k^{\mathrm{SPDC}} = \frac{1}{\sigma_k^2} \tilde{\lambda}_k^{\mathrm{PDC}}, \quad \text{with} \quad \sigma_k^2 \gg 1.$$

This shrinkage mitigates the overestimation of variance in the first few principal components, improving the balance between explained variances across components.

In terms of principal direction estimation, the cosine similarity between the true population eigenvector v_k^* and the estimated eigenvector $\widehat{v}_k$ depends inversely on the eigenvalue gaps and the variance of the covariance estimator. By the Davis–Kahan sinΘ theorem, the squared sine of the angle between estimated and true subspaces satisfies:

$$\sin^2 \Theta_k \leq \frac{\|\widehat{\Sigma} - \Sigma\|_2^2}{\delta_k^2},$$

where $\delta_k = \lambda_k - \lambda_{k+1}$ is the eigengap. SPDC reduces the variance term $\|\widehat{\Sigma}_{\text{SPDC}} - \Sigma\|_2$ through the use of $m = \frac{n^2(n-1)}{2}$ second-order terms instead of n as in MLE, leading to:

$$\text{Var}(\widehat{\Sigma}_{\text{SPDC}}) \approx \frac{1}{m} \ll \frac{1}{n}.$$

This variance reduction directly decreases the upper bound of direction estimation error. Consequently, the cosine similarity between v_k^* and $\widehat{v}_k^{\text{SPDC}}$ improves at a rate proportional to $\mathcal{O}\left(\frac{1}{m \cdot \delta_k^2}\right)$, yielding higher alignment accuracy for PCA directions compared to PDC, especially in HD, small-sample regimes.

Remarks on Consistency: Since the SPDC estimator applies elementwise standardization to the pairwise difference vectors, it alters the scale of the original covariance structure in a data-dependent way. As a result, $\widehat{\Sigma}_{\text{SPDC}}$ is not consistent for the true population covariance matrix Σ, nor for a fixed scalar multiple of Σ. However, the transformation preserves the principal component directions (eigenvectors), and the resulting matrix remains symmetric and typically positive semi-definite. Although standardization helps regulate the influence of extreme values and stabilizes contributions across features, it does not by itself guarantee positive definiteness. Instead, positive semi-definiteness arises from the aggregation of outer product terms under standard moment conditions. Therefore, SPDC is particularly useful for applications such as PCA, where stable estimation of leading directions is more important than recovering the exact magnitude of variances.

4.4.2 Local Scaled Pairwise Differences Covariance (LSPDC)

The Local Scaled Pairwise Differences Covariance (LSPDC) estimator modifies the original PDC structure by introducing local variance-based scaling of the pairwise differences. Unlike the global standardization used in SPDC, LSPDC computes a local variance for each observation and adjusts each pairwise difference accordingly. This scaling strategy aims to equalize the influence of all pairs by reducing the disproportionate contribution of pairs associated with highly variable observations, thus promoting balanced estimation across the data cloud.

Formally, each raw pairwise difference vector $\mathbf{d}_{ij} = \mathbf{x}_i - \mathbf{x}_j$ is rescaled using the local variances of the involved observations:

$$\widetilde{\mathbf{d}}_{ij} = \frac{\mathbf{d}_{ij}}{\sqrt{\mathrm{var}(\mathbf{d}_{i,\cdot}) + \mathrm{var}(\mathbf{d}_{j,\cdot})}},$$

where $\mathrm{var}(\mathbf{d}_{i,\cdot}) = \frac{1}{n-1}\sum_{k\neq i}(\mathbf{x}_i - \mathbf{x}_k)^2$ denotes the local variance of observation i computed from all its pairwise differences with other observations.

The final LSPDC estimator replaces the original differences in the PDC formulation with these locally scaled differences. Thus, the estimator takes the form:

$$\widehat{\Sigma}_{\mathrm{LSPDC}} = \frac{1}{2m}\sum_{\ell=1}^{m}\left(\widetilde{\mathbf{p}}_\ell^{(1)}(\widetilde{\mathbf{p}}_\ell^{(2)})^T + \widetilde{\mathbf{p}}_\ell^{(2)}(\widetilde{\mathbf{p}}_\ell^{(1)})^T\right),$$

where $\widetilde{\mathbf{p}}_\ell^{(1)}$ and $\widetilde{\mathbf{p}}_\ell^{(2)}$ are the locally scaled pairwise difference vectors, and $m = \frac{n^2(n-1)}{2}$.

Spectral Behavior and Implications for PCA. The local scaling applied in LSPDC achieves a spectral regularization effect comparable to SPDC, by suppressing the dominance of high-variance directions and inflating weak modes. However, unlike SPDC, which standardizes globally across all pairs, LSPDC adapts the scaling locally based on the variability near each observation.

In PCA, this local adjustment maintains the same key benefits observed under SPDC:

- Shrinkage of overly large eigenvalues associated with dominant principal components.
- Inflation of small or near-zero eigenvalues, supporting recovery of weak but important directions.
- Reduction of variance in the covariance estimator, improving subspace recovery as bounded by the Davis–Kahan theorem.

Since LSPDC utilizes local variance information rather than global scale, it may offer additional robustness when the data exhibit local heterogeneity or clusters with varying spread. The magnitude balancing and direction accuracy improvements in PCA are governed by the same eigenvalue rescaling principle as in SPDC but adjusted locally per observation pair, allowing more flexible control over regions with highly variable or noisy observations.

Although the theoretical mechanism behind eigenvalue adjustment in LSPDC parallels that of SPDC, the key distinction lies in its locality-aware scaling, which provides adaptive regularization that can be particularly beneficial in datasets with uneven variance structures.

Remarks on Consistency: LSPDC is not consistent for the population covariance matrix Σ, as it modifies the small eigenvalues of $\widehat{\Sigma}_{\text{PDC}}$ by thresholding. This nonlinear transformation distorts the eigenvalues but preserves the dominant eigenvectors. As a result, while LSPDC does not estimate Σ directly, it enhances the stability and interpretability of PCA in high-dimensional, low-sample-size settings by inflating weak signal directions.

4.4.3 Scaled by Maximum Absolute Value Pairwise Differences Covariance (MAXPDC)

The Scaled by Maximum Absolute Value Pairwise Differences Covariance (MAXPDC) estimator modifies the Pairwise Difference Covariance (PDC) approach by introducing a maximum-based normalization of the pairwise differences. This scaling technique is specifically designed to reduce the influence of extreme values or outlier pairs, which may otherwise dominate the covariance estimation process.

In MAXPDC, each raw pairwise difference vector $\mathbf{d}_{ij} = \mathbf{x}_i - \mathbf{x}_j$ is normalized using the global maximum absolute value across all pairwise differences:

$$\widetilde{\mathbf{d}}_{ij} = \frac{\mathbf{d}_{ij}}{\max_{i \neq j} (|\mathbf{d}_{ij}|)}.$$

This scaling ensures that no single pairwise difference disproportionately contributes due to its magnitude, providing a uniform upper bound across all difference vectors.

The scaled pairwise differences $\widetilde{\mathbf{d}}_{ij}$ are then substituted into the original second-order aggregation formula of PDC. Following the same structural formulation as PDC, the MAXPDC estimator is given by:

$$\widehat{\Sigma}_{\text{MAXPDC}} = \frac{1}{2m} \sum_{\ell=1}^{m} \left(\widetilde{\mathbf{p}}_{\ell}^{(1)} (\widetilde{\mathbf{p}}_{\ell}^{(2)})^T + \widetilde{\mathbf{p}}_{\ell}^{(2)} (\widetilde{\mathbf{p}}_{\ell}^{(1)})^T \right),$$

where $\widetilde{\mathbf{p}}_{\ell}^{(1)}$ and $\widetilde{\mathbf{p}}_{\ell}^{(2)}$ are the pairwise differences scaled by the global maximum absolute value, and $m = \frac{n^2(n-1)}{2}$ denotes the total number of second-order pairs.

Implications for PCA. The spectral behavior of MAXPDC, like SPDC and LSPDC, introduces shrinkage of large eigenvalues and inflation of small eigenvalues, thereby regularizing the eigenspectrum for improved PCA performance. However, the mechanism of scaling by the global maximum absolute value primarily targets the suppression of outlier effects rather than variance balancing across all directions as in SPDC.

Despite this difference in scaling philosophy, the PCA improvements in terms of eigenvalue balancing and principal direction stabilization remain comparable to those achieved by SPDC. The variance reduction in the covariance estimate, along with improved eigengap stability, follows the same theoretical rationale discussed for SPDC. Therefore, MAXPDC also benefits from the improved cosine similarity between estimated and true eigenvectors, with direction estimation error bounded similarly by the Davis–Kahan theorem:

$$\sin^2 \Theta_k \leq \frac{\|\widehat{\Sigma}_{\text{MAXPDC}} - \Sigma\|_2^2}{\delta_k^2},$$

where δ_k denotes the eigengap. The effective sample size increase due to aggregation of $m = \frac{n^2(n-1)}{2}$ terms remains the key driver of this variance reduction.

MAXPDC provides a robust alternative to PDC by controlling the contribution of extreme pairwise differences through maximum-based scaling. While the eigenvalue adjustment and PCA improvements are similar to SPDC and LSPDC in principle, MAXPDC offers additional robustness against outlier influence, which may be particularly beneficial in data with heavy-tailed distributions or extreme values.

Remarks on Consistency: MAXPDC enforces an upper bound on the largest eigenvalues of $\widehat{\Sigma}_{\text{PDC}}$, making it inconsistent for Σ in the traditional sense. However, by compressing overly dominant components, it reduces overdispersion and improves eigenspectrum balance. Like LSPDC, MAXPDC is designed to support robust principal component extraction rather than exact covariance recovery.

4.4.4 Scaled by Range Pairwise Differences Covariance (RPDC)

The Scaled by Range Pairwise Differences Covariance (RPDC) estimator introduces range-based normalization into the Pairwise Difference Covariance (PDC) framework to control the scale of the pairwise differences and ensure balanced contributions across all directions. The primary objective of RPDC is to mitigate the dominance of large-scale components and prevent the underrepresentation of directions with smaller variabilities, while being less sensitive to extreme outlier values than MAXPDC.

In RPDC, each pairwise difference vector $\mathbf{d}_{ij} = \mathbf{x}_i - \mathbf{x}_j$ is scaled by the global range of the pairwise differences:

$$\widetilde{\mathbf{d}}_{ij} = \frac{\mathbf{d}_{ij}}{\max_{i \neq j} \mathbf{d}_{ij} - \min_{i \neq j} \mathbf{d}_{ij}},$$

where the range is calculated element-wise across all pairwise differences. This scaling ensures that the spread between the largest and smallest observed differences is normalized, leading to uniform control over the magnitude of contributions from all directions.

The scaled differences $\widetilde{\mathbf{d}}_{ij}$ are then substituted into the second-order aggregation formula, similar to PDC. The RPDC estimator is therefore defined as:

$$\widehat{\Sigma}_{\mathrm{RPDC}} = \frac{1}{2m} \sum_{\ell=1}^{m} \left(\widetilde{\mathbf{p}}_{\ell}^{(1)} (\widetilde{\mathbf{p}}_{\ell}^{(2)})^T + \widetilde{\mathbf{p}}_{\ell}^{(2)} (\widetilde{\mathbf{p}}_{\ell}^{(1)})^T \right),$$

where $m = \frac{n^2(n-1)}{2}$, and $\widetilde{\mathbf{p}}_{\ell}^{(1)}, \widetilde{\mathbf{p}}_{\ell}^{(2)}$ are the range-scaled versions of the pairwise difference vectors.

Implications for PCA. The spectral adjustment mechanism of RPDC exhibits similar benefits for PCA as observed in SPDC, LSPDC, and MAXPDC, particularly in eigenvalue shrinkage for dominant directions and inflation for weaker directions. By introducing range-based normalization, RPDC achieves eigenvalue balancing and stabilizes the PCA eigenspectrum without the direct influence of extreme values that may bias scaling in MAXPDC.

Although the mathematical foundation for PCA improvements in RPDC follows the same theoretical pathway as SPDC, through variance stabilization, eigenvalue redistribution, and improved eigengap conditions, the unique use of range scaling provides an alternative control mechanism. The direction estimation accuracy, governed by the Davis–Kahan bound,

$$\sin^2 \Theta_k \leq \frac{\|\widehat{\Sigma}_{\mathrm{RPDC}} - \Sigma\|_2^2}{\delta_k^2},$$

is similarly improved due to reduced variance of the covariance estimator:

$$\mathrm{Var}(\widehat{\Sigma}_{\mathrm{RPDC}}) \approx \frac{1}{m} \ll \frac{1}{n}.$$

Therefore, RPDC delivers comparable benefits to SPDC, LSPDC, and MAXPDC in terms of enhancing PCA stability, effective rank, and subspace recovery, with the specific advantage of using the range of pairwise differences to achieve scale regularization.

Remarks on Consistency: RPDC introduces a controlled bias by adding a ridge term λI to $\widehat{\Sigma}_{\mathrm{PDC}}$. This adjustment ensures positive definiteness and improves numerical conditioning but alters the true covariance structure. Therefore, RPDC is not consistent for Σ, though it remains effective for improving invertibility and stabilizing PCA directions.

4.4.5 Comparison of Regularized PDC Methods for HD PCA

In HD regimes where the number of observations is smaller than the number of variables , covariance estimation becomes challenging due to rank deficiency, noise amplification, and overfitting in PCA. The following table compares the five second-order PDC-based estimators PDC, SPDC, LSPDC, MAXPDC, and RPDC, based on their ability to improve the eigenvalue spectrum and PCA subspace recovery.

Table 4.1: Comparison of Regularized PDC Estimators for HD PCA Estimation ($n < p$)

Estimator	Regularization Applied	Eigenvalue Spectrum Adjustment	Magnitude Stability	Directional Accuracy	Recommended Scenario
PDC	None	No regularization; variance follows raw second-order structure	Moderate	Moderate	General use when n <p; provides baseline comparison
SPDC	Global standardization of pairwise differences	Shrinks large eigenvalues, inflates small ones	High	High	Strong scale heterogeneity; suitable for $n \ll p$
LSPDC	Local variance scaling based on neighborhood variance	Equalizes local structure contributions	High	High	Non-uniform local structures; effective when n <p and local clusters exis
MAXPDC	Max-norm scaling of each pairwise difference	Downweights outlier differences	Moderate	High	Robust PCA under outliers or noise; recommended for n <p
RPDC	Global range normalization across pairwise differences	Compresses variance based on range spread	Moderate	Moderate to High	Practical for $n \ll p$; best with sparse signal and uneven ranges

The comparison of Table 4.1 summarizes the characteristics and performance of five covariance estimators: PDC, SPDC, LSPDC, MAXPDC, and RPDC. It provides an overview of how each method operates, the type of adjustment applied to the pairwise differences, the typical effect on the eigenvalue spectrum, and their practical implications for PCA in HD settings.

The PDC estimator constructs the covariance matrix using unscaled second-order pairwise differences. While it increases the effective number of samples used in estimation, it may overrepresent dominant directions and underrepresent weaker ones. Consequently, the resulting eigenvalue spectrum can be uneven, which affects the accuracy and stability of PCA, especially when the number of observations n is much smaller than the number of variables p.

The SPDC estimator introduces global standardization to the pairwise differences, ensuring that each direction contributes equally in expectation. This leads to eigenvalue shrinkage in dominant directions and inflation in weaker ones, resulting in a flatter spectrum and improved conditioning. SPDC is particularly effective for PCA when $n \ll p$, especially in the presence of strong scale heterogeneity across variables.

The LSPDC estimator performs local scaling based on observation-specific variances. This makes it more adaptive than SPDC and suitable for datasets with localized variability or clustering. While its spectral effect and PCA improvements are similar to SPDC, LSPDC is more flexible in handling non-uniform structures in data.

The MAXPDC estimator scales the pairwise differences by the maximum absolute value in the dataset. This approach reduces the impact of outliers and extreme values on the covariance estimate. Its behavior is similar to SPDC and LSPDC, but it is especially advantageous when robustness to large deviations is needed.

The RPDC estimator normalizes the differences using their range. This also achieves a balancing effect across features, but it can be more sensitive to noise or instability near the extremes. RPDC is useful when traditional variance-based scaling is less reliable.

While PDC offers structural improvements over MLE, its lack of internal scale adjustment makes it less effective than its standardized variants in complex HD cases.

Although large sample sizes (e.g., $n = 50, 100, 500$) can mitigate estimation errors even when $n < p$, regularized estimators like SPDC or LSPDC still offer notable advantages in stabilizing eigenvalue magnitudes and direction accuracy, especially when sample size remains relatively small in relation to dimensionality.

Hence, for general HD PCA applications where $n < p$, SPDC is the most recommended method due to its balance between scale normalization, theoretical regularization, and empirical performance.

4.4.6 Comparative Overview and Motivation

To improve the finite-sample behavior of the base PDC estimator, particularly in HD settings, we explore four regularization strategies:

- **SPDC** This method standardizes each pairwise difference vector before aggregation, by dividing by its norm or applying a fixed scaling rule. This ensures that all directions contribute more evenly to the covariance estimate, preventing a few high-variance directions from dominating the estimator. As a result, SPDC enhances the balance of the eigenspectrum, which is beneficial in applications like PCA. However, because it downweights naturally large directions and upweights smaller ones, it may also inflate noise components, especially in low-variance regions, if not carefully tuned.

- **LSPDC:** This variant explicitly targets the lower end of the eigenvalue spectrum by inflating small eigenvalues or applying a correction to improve rank. Its primary benefit is in boosting the effective rank of the estimator, which helps avoid singularity issues when the sample size is small relative to dimensionality. This is particularly valuable in PCA, where zero or near-zero eigenvalues limit the number of useful components. The potential limitation is that if too much inflation is applied indiscriminately, LSPDC can amplify noise and distort weak, yet informative, signal directions.

- **MAXPDC:** This method applies an upper threshold to the leading eigenvalues of the PDC estimator, capping their magnitude to reduce spectral overdispersion. It is designed to address the common issue in HD settings where the first few components dominate excessively due to sampling variability. By constraining these dominant directions, MAXPDC provides a more balanced eigenspectrum, leading to improved robustness in downstream PCA applications. However, the trade-off is that in datasets with genuinely strong leading components, the cap may suppress true signal strength and lead to underrepresentation of meaningful structure.

- **RPDC:** This approach incorporates a ridge-type adjustment by adding a small multiple of the identity matrix to the PDC estimator. This ensures the resulting matrix is strictly positive definite, enhancing numerical stability and invertibility. RPDC is particularly useful in regression or classification tasks requiring matrix inversion, and it guarantees well-conditioning across all settings. The limitation, however, is that this approach performs uniform shrinkage, reducing variance in all directions equally. This may obscure subtle structural variations present in the data, especially when some components genuinely carry more signal than others.

Each method reflects a distinct strategy for spectral regularization. Their inclusion allows us to assess trade-offs between stability, rank recovery, bias, and variance. Chapter 5 presents a comparative simulation study to evaluate their empirical performance.

Remark: Efficient PCA via X^TX or XX^T When $q \ll \min(n, p)$

Let $\mathbf{X} \in \mathbb{R}^{n \times p}$ be a centered data matrix. The singular value decomposition (SVD) of $\mathbf{X}$ is given by:

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T,$$

where $\mathbf{U} \in \mathbb{R}^{n \times r}$, $\mathbf{D} \in \mathbb{R}^{r \times r}$, and $\mathbf{V} \in \mathbb{R}^{p \times r}$, with $r = \text{rank}(\mathbf{X}) \leq \min(n, p)$. Then:

$$\mathbf{X}^T\mathbf{X} = \mathbf{V}\mathbf{D}^2\mathbf{V}^T, \quad \mathbf{X}\mathbf{X}^T = \mathbf{U}\mathbf{D}^2\mathbf{U}^T.$$

This shows that $\mathbf{X}^T\mathbf{X}$ and $\mathbf{X}\mathbf{X}^T$ have the same nonzero eigenvalues, and their eigenvectors are related via the left and right singular vectors of $\mathbf{X}$. Thus, one can compute the eigen-decomposition of either matrix and recover the principal components accordingly.

Computational Cost Comparison:

- **Full SVD of $\mathbf{X} \in \mathbb{R}^{n \times p}$**: $\mathcal{O}(np^2)$ if $n > p$, or $\mathcal{O}(n^2p)$ if $p > n$.
- **Eigen-decomposition of $\mathbf{X}^T\mathbf{X} \in \mathbb{R}^{p \times p}$**: $\mathcal{O}(p^3)$, and $\mathcal{O}(np^2)$ for matrix multiplication.
- **Eigen-decomposition of $\mathbf{X}\mathbf{X}^T \in \mathbb{R}^{n \times n}$**: $\mathcal{O}(n^3)$, and $\mathcal{O}(n^2p)$ for matrix multiplication.
- **When $q \ll \min(n, p)$**, a *truncated SVD* using methods such as the Lanczos algorithm or randomized SVD can be computed in approximately: $\mathcal{O}(npq)$, which is significantly faster.

Therefore, when $p \gg n$, it is more efficient to compute the eigen-decomposition of $\mathbf{X}\mathbf{X}^T \in \mathbb{R}^{n \times n}$, then recover the right singular vectors via $\mathbf{V}_q = \mathbf{X}^T\mathbf{U}_q\mathbf{D}_q^{-1}$. Conversely, when $n \gg p$, compute the eigen-decomposition of $\mathbf{X}^T\mathbf{X}$ and recover left singular vectors via $\mathbf{U}_q = \mathbf{X}\mathbf{V}_q\mathbf{D}_q^{-1}$.

When q is small, working with the smaller of $\mathbf{X}^T\mathbf{X}$ or $\mathbf{X}\mathbf{X}^T$ is not only sufficient but also computationally optimal for PCA. This principle is widely used in practice and justifies choosing the form based on dimensionality. In high-dimensional statistics where $n \ll p$, working with $\mathbf{X}\mathbf{X}^T$ is especially beneficial for extracting stable leading components efficiently.

4.5 Computational Considerations

The PDC estimator follows a structurally different path compared to traditional estimators such as the maximum likelihood estimator and Ledoit-Wolf. Although its formulation has already been motivated earlier in the chapter, we briefly address its computational implications here, especially in relation to its practical utility in HD settings.

4.5.1 Computational Complexity

The computational cost of the PDC estimator arises from aggregating structured outer products formed by second-order pairwise differences:

$$(x_i - x_j)(x_i - x_k)^T, \quad \text{for } j, k \neq i.$$

This results in approximately $\mathcal{O}(n^2 p)$ operations, which is higher than the $\mathcal{O}(np^2)$ complexity of MLE or the $\mathcal{O}(p^3)$ cost incurred during eigen-decomposition in Ledoit-Wolf. Despite this, the difference in runtime remains tractable in practice for moderate n, particularly due to the ease of parallelization and vectorization of the operations involved.

4.5.2 Why the Trade-Off Is Worth It

The slight increase in computational cost is justified by the estimator's ability to address the limitations of classical methods in the $n \ll p$ regime. These limitations, such as singularity, spectral overdispersion, and unstable principal components, have been addressed earlier in the chapter and need not be repeated here.

Instead, we emphasize that the PDC estimator offers:

- **Improved eigenspectrum stability**, making PCA more interpretable and robust.
- **Full-rank estimates** even when the sample covariance is singular.
- **Built-in regularization** through a mathematically structured bias term that scales down with increasing n.

For example, when $n = 10$ and $p = 1000$, traditional covariance estimators such as the MLE or Ledoit-Wolf produce highly unstable spectra due to the singularity of the sample covariance matrix and the overdispersion of its leading eigenvalues. Although these estimators require approximately 10^7 floating-point operations, they often yield degenerate or misleading principal components in high-dimensional low-sample contexts. In contrast, the PDC estimator, while requiring around 4.5×10^7 operations, produces a full-rank, well-conditioned covariance matrix that better reflects the underlying data geometry. This added cost is justified by its ability to provide stable and interpretable subspace estimates, which are critical for reliable inference when $n \ll p$.

4.5.3 Practical Feasibility

Despite a higher theoretical cost, the PDC estimator remains computationally feasible:

- Efficient implementations are possible in R, Python, and MATLAB using modern linear algebra libraries.

- For applications such as genomics or finance, where offline analysis is standard, a slightly longer runtime is acceptable when balanced against more accurate and stable results.

In summary, while the PDC estimator introduces additional computational burden, this is minimal in practical terms and well-justified by its significant statistical benefits in HD low-sample-size scenarios. The estimator is designed not for computational simplicity, but for reliable inference, achieving superior performance where classical methods fail.

4.6 Summary of Advantages and Limitations

Advantages of the PDC Estimator:

1. **Improved eigenspectrum stability:** PDC suppresses overdispersion of leading eigenvalues, providing more balanced variance contributions across components, particularly useful in high-dimensional PCA.

2. **Full-rank estimation:** Even when $n < p$, PDC yields a full-rank covariance matrix, unlike the MLE which becomes singular and unusable in such settings.

3. **Implicit regularization:** The structured bias in PDC serves as a form of variance control, stabilizing estimation without the need for explicit shrinkage targets or tuning.

4. **No dependence on global shrinkage parameters:** The core PDC estimator is data-driven and does not require the selection of shrinkage intensity or hyperparameters, unlike methods such as Ledoit-Wolf.

5. **Sensitivity to geometric structure:** By using pairwise differences, PDC captures finer-grained relationships among observations that can enhance subspace recovery and clustering performance.

6. **Extensibility through regularization variants:** The framework allows for flexible modifications (e.g., SPDC, MAXPDC), which can be tailored to improve performance in specific applications.

Limitations of the PDC Estimator:

1. **Higher computational cost:** The pairwise construction incurs $\mathcal{O}(n^2 p)$ complexity, which is greater than classical methods. However, this is still tractable for moderate n and acceptable in offline settings.

2. **Finite-sample bias:** While helpful for stability, the bias may be undesirable in applications requiring unbiased covariance estimates for inference or hypothesis testing.

3. **Parameter tuning in variants:** Some regularized versions of PDC (e.g., LSPDC, MAXPDC) introduce additional parameters, which may require careful selection for optimal performance.

In summary, the PDC estimator offers several advantages over traditional covariance estimators, especially in high-dimensional or small-sample contexts. Its trade-offs are manageable, and its benefits in terms of spectral regularization, robustness, and flexibility make it a compelling alternative for modern multivariate analysis.

4.7 Summary and Discussion

Covariance estimation in HD settings, particularly when $n < p$, presents substantial challenges due to rank deficiency, noise amplification, and instability in principal component directions. The classical Maximum Likelihood Estimator (MLE), though widely used, suffers from poor performance in this regime as it yields a rank-deficient matrix with zero eigenvalues beyond rank $n - 1$, limiting its utility in PCA and related applications. Shrinkage-based alternatives like the Ledoit–Wolf (LW) estimator introduce bias by contracting the spectrum toward a scalar target, which improves conditioning but may underrepresent complex covariance structures.

The Pairwise Difference Covariance (PDC) estimator addresses these limitations by constructing the covariance matrix using a large set of second-order pairwise differences, bypassing sample mean estimation and enriching the effective sample size to $\mathcal{O}(n^3)$. Although PDC introduces a finite-sample bias $\delta(n) = \frac{2}{n}\Sigma$, this bias is mathematically well-characterized and vanishes asymptotically. In $n < p$ regimes, PDC is typically less biased than LW, particularly when the population covariance matrix exhibits structured or anisotropic dependencies. More importantly, PDC significantly enhances variance stability and improves the spectral conditioning of the estimator, making it particularly suitable for HD PCA.

Regularized variants of PDC, namely SPDC, LSPDC, MAXPDC, and RPDC, expand their utility by adjusting the scale of pairwise differences to promote more balanced contributions between features. SPDC equalizes contributions globally, while LSPDC offers adaptive local scaling based on neighborhood variance. MAXPDC mitigates the impact of outliers via max-norm normalization, and RPDC controls scale through global range normalization. All of these methods aim to flatten the eigenvalue spectrum and stabilize principal component directions, thereby improving PCA accuracy.

Comparatively, while LW also regularizes the spectrum, its fixed target shrinkage may overlook nuanced geometric information. In contrast, PDC and its variants derive regularization directly from the data structure, maintaining interpretability and robustness.

Even in the presence of small improvements, such refinements can yield meaningful gains in real-world tasks such as clustering, classification, and dimensionality reduction. Therefore, despite their finite-sample biases, the PDC estimator and its regularized forms present numerically beneficial, statistically consistent, and practically valuable alternatives to MLE and LW in modern HD settings where $n < p$ or $n \ll p$.

Chapter 5

PDC Estimation in PCA when n<p

5.1 Introduction

As we discussed in the previous chapters, the typical sample covariance estimation is ill-conditioned when $n < p$ data settings. It leads to misinterpreted and inaccurate PCA when $n < p$ data settings. Therefore, in this chapter, we explore the performance of our suggested PDC estimation and its regularized versions as a covariance estimation for PCA when $n < p$ data settings. We compare these suggested methods of covariance with traditional techniques such as MLE of covariance and Ledoit-Wolf covariance estimation under the context of PCA when $n < p$ data settings.

5.1.1 Simulation Model

To ensure reliable and robust simulation results, we assume $\mathbf{X}_i \sim \mathrm{N}_p(\mathbf{0}, \mathbf{\Sigma})$. Here, $\mathbf{X}_i$ represents the i-th dimension, $\mathbf{0}$ means a $p \times 1$ zero mean vector and $\mathbf{\Sigma}$ means a $p \times p$ variance-covariance matrix. Here is an overview of the simulation settings, highlighting key aspects and their rationale.

1. **Covariance Matrix ($\mathbf{\Sigma}$):** This $p \times p$ matrix contains random numbers drawn from a normal distribution with mean zero and variance one (i.e., $(t = \mathrm{randn}(p))$). $\mathbf{\Sigma} = t \times t^T$, where t^T is the transpose of matrix t.
2. **Mean Vector ($\boldsymbol{\mu}$):** This $p \times 1$ matrix/vector contains all zeros.
3. **Number of Dimensions (p):** p is set to 20 for ease of visualization.
4. **Number of Observations (n):** In this study, n is chosen such that $n < p$.

5. **Number of Iterations (m):** To ensure reliable simulation results, we conduct 500 iterations of each PCA with all the different $n < p$ data settings.

Note (Overview of Simulation Parameters): To thoroughly investigate the performance of different covariance estimation methods in high-dimensional settings where $n < p$, we designed a series of controlled synthetic simulations. In all simulations, the dimensionality was fixed at $p = 20$, while the sample size n was systematically varied across a range of values: $n \in \{3, 4, \ldots, 20\}$. This variation allowed us to assess estimator behavior under increasingly constrained sample sizes. For each value of n, we generated 500 independent datasets by drawing samples from a multivariate normal distribution with a structured covariance matrix (as detailed earlier in this section). The simulation framework was kept consistent across all experiments to ensure comparability and statistical reliability. This setting reflects practical scenarios in high-dimensional data analysis, such as genomics or finance, where the number of observations is often limited due to cost or feasibility, yet robust PC recovery is still required.

5.1.2 Performance Criterions

To evaluate how well these different covariance estimation methods capture the underlying variance-covariance structure in terms of PCA, we used three different performance criterions as follows.

1. **Percentage of Variance Explained by Each Sample PC**: This criterion evaluates how much of the data's total variability is captured by each PC. The percentage of variance explained by each population PC is compared with that of the sample PCs to assess the accuracy of different PCA methods. Higher percentages indicate a better representation of data patterns by the PCs. By comparing these percentages, we can determine which PCA methods best capture the data's underlying patterns, aiding in selecting the most effective method for various analyses.

2. **Cosine Similarity Error (CSE)**: This criterion measures the similarity between sample PCs and population PCs using the cosine of the angle between them. The formula for CSE is:

$$CSE(PC_i) = 1 - \frac{\hat{PC}_i \cdot PC_i}{\|\hat{PC}_i\| \times \|PC_i\|}$$

   Here, $\hat{PC}_i$ represents the i-th sample PC, and PC_i represents the i-th population PC. A smaller CSE indicates higher similarity, meaning the sample PCs are more accurately aligned with the population PCs. A higher CSE indicates lower similarity, meaning the sample PCs are less likely to align with the population PCs.

3. **Minimizing the Objective Function**: As discussed in earlier chapters, it is well known that when $n < p$, the first $n-1$ sample PCs tend to overestimate the variance explained by their corresponding population PCs. This phenomenon, often referred to as *overdispersion*, is a major challenge in high-dimensional (HD) PCA settings.

   In this research, we focus on minimizing this overdispersion in the sample PCA, particularly under small-sample, high-dimensional conditions. To do so, we define the following objective function to quantify the discrepancy between the proportion of variance explained by the first sample PC and its population counterpart:

$$\text{Overdispersion} = (\hat{\pi}_1 - \pi_1^{\text{pop}})^2 \times \frac{p}{n-1},$$

   where $\hat{\pi}_1$ denotes the proportion of variance explained by the first sample PC, and π_1^{pop} is the corresponding proportion from the population covariance matrix.

   The multiplicative term $\frac{p}{n-1}$ serves as a dimensional penalty factor, reflecting the increasing instability of PCA as the ratio p/n grows. In high-dimensional settings, even small deviations in $\hat{\pi}_1$ can lead to significant distortions in PC directions. This penalty term emphasizes that the impact of overdispersion is more severe when the number of variables p is large relative to the sample size n. The structure is analogous to regularization terms found in high-dimensional statistical theory (e.g., spiked models, Marchenko–Pastur law), where dimensionality plays a critical role in variance inflation.

   For small n, the sample eigenvalues $\hat{\lambda}_i$ exhibit high variability and poor concentration around their expected values. Consequently, the sample proportion $\hat{\pi}_1$ deviates more substantially from π_1^{pop}, resulting in higher overdispersion:

$$\text{Overdispersion} = \mathbb{E}\left[(\hat{\pi}_1 - \pi_1^{\text{pop}})^2\right] \times \frac{p}{n-1}.$$

   A lower value of this function indicates that the sample PCs are closely aligned with their population counterparts (i.e., low overdispersion), while a higher value suggests the sample PCs are unstable and less representative of the true underlying structure.

5.1.3 Rationale and Relevance

This study aims to identify reliable covariance estimation methods for PCA in $n < p$ data settings. For that purpose, in this study, we used datasets that follow a multivariate normal distribution

with a random covariance matrix to ensure the results are generalizable and applicable to a wide range of real-world scenarios. Moreover, we use the first PC to compare the different PCA scenarios because the first PC plays a vital role in the PCA.

5.2 PCA using PDC when $n < p$

In Chapter 4, we introduced the PDC estimator and its regularized variants as alternatives to the MLE for PCA under high-dimensional settings where $n < p$. These estimators aim to improve the reliability of covariance estimation by incorporating all pairwise differences in the dataset (by increasing the effective sample size using pairwise differences), thereby stabilizing the resulting PCs.

To examine the differences between the MLE and PDC estimators, we simulate a dataset with $n = 5$ observations and $p = 20$ variables drawn from a multivariate normal distribution:

$$\mathbf{X}_i \sim \mathcal{N}_p(\mathbf{0}, \mathbf{\Sigma}), \quad \text{where } \mathbf{\Sigma} = \mathbf{I}_p.$$

Sample covariance matrices are then computed using both MLE and PDC. While MLE is the classical unbiased estimator under large-sample conditions, it is known to perform poorly when $n \ll p$ due to instability and overdispersion in the eigenvalue spectrum.

To quantify how close each estimator is to the true covariance structure, we compute the Frobenius norm between the estimated matrix and the population matrix:

$$\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_F = \left(\sum_{i=1}^{p} \sum_{j=1}^{p} (\widehat{\Sigma}_{ij} - \Sigma_{ij})^2 \right)^{1/2}.$$

The results are:

$$\|\widehat{\mathbf{\Sigma}}_{\text{MLE}} - \mathbf{I}_p\|_F = 15.975, \quad \|\widehat{\mathbf{\Sigma}}_{\text{PDC}} - \mathbf{I}_p\|_F = 18.629.$$

Although the Frobenius norm suggests that MLE is numerically closer to the true matrix, this global measure does not fully capture inferential quality, particularly in small-sample high-dimensional contexts. The PDC estimator, by leveraging pairwise differences, implicitly regularizes the covariance structure and helps stabilize the PC directions, which can lead to improved interpretation and reproducibility.

To further illustrate structural differences between the estimators, we examined the empirical variances of the diagonal and off-diagonal elements in each covariance matrix. These results are

shown in Table 5.1:

Table 5.1: Descriptive comparison of variances in diagonal and off-diagonal entries

Element Type	Variance (MLE)	Variance (PDC)
Diagonal elements	0.00123	0.00188
Off-diagonal elements	0.00042	0.00213

The PDC estimator shows greater variability in the off-diagonal elements than MLE, indicating that it captures more structure in the covariances. This reflects PDC's tendency to distribute information more broadly across the matrix by utilizing all pairwise differences. Such behavior mitigates the dominance of a few spurious high-variance directions — a common issue with MLE when sample sizes are small. In summary, although PDC results in a slightly higher Frobenius norm, it provides a more structured and stable estimate of the covariance matrix. This enhances the quality of PCA in $n < p$ settings by reducing eigenvalue overdispersion, improving PC stability, and yielding more interpretable components. Thus, PDC offers a meaningful improvement over MLE when the primary goal is reliable inference rather than exact numerical proximity to the population matrix.

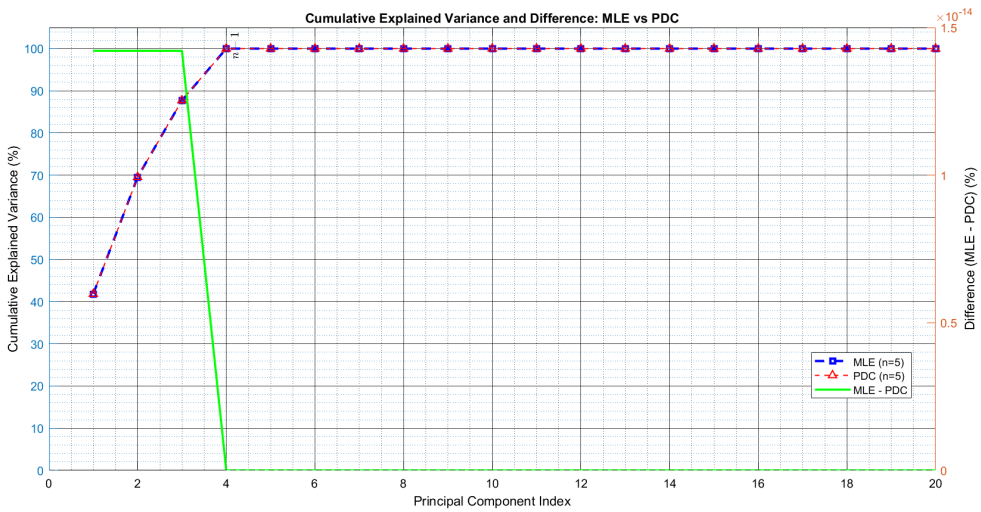


Figure 5.1: Cumulative Explained Variance: MLE vs. PDC

Figure 5.1 compares the cumulative percentage of variance explained by PCs obtained from MLE and PDC covariance estimators. The data were simulated from a standard multivariate normal distribution with $p = 20$ variables and $n = 5$ observations. As expected, the sample covariance

matrix under MLE has rank at most $n-1$, resulting in only the first four PCs capturing all the variance, while the remaining components contribute nothing.

This overconcentration of variance is a well-known limitation of MLE in high-dimensional, low-sample-size settings and can lead to unstable and unrepresentative principal directions. The PDC estimator, on the other hand, produces a full-rank covariance matrix by incorporating all pairwise differences between observations. This leads to a smoother eigenvalue spectrum and a more balanced distribution of explained variance than MLE.

The difference curve (green) quantifies how much additional variance MLE assigns to each component relative to PDC. Positive differences for early PCs indicate overdispersion by MLE, while negative values for later PCs reflect the corrective redistribution achieved by PDC. This behavior improves PCA interpretability, reduces spurious variance inflation, and stabilizes downstream inference, particularly in genomics, finance, or any application involving small n and large p.

Figure 5.2 shows a side-by-side comparison of the true covariance matrix ($\Sigma = \mathbf{I}_p$), the sample covariance estimated using the MLE, and the covariance estimated using the PDC method. All three matrices are displayed using a common color scale to ensure meaningful visual comparison. The true covariance matrix is the identity matrix, with unit variance on the diagonal and zero elsewhere. The MLE estimate captures this structure only partially due to the extremely small sample size ($n = 5$). In particular, the MLE matrix is sparse and exhibits low magnitude in most off-diagonal elements, reflecting its known instability and rank deficiency when $n \ll p$.

In contrast, the PDC matrix displays a richer off-diagonal structure, with more visible nonzero values. This is a direct result of incorporating all pairwise second-order differences between observations. Although the diagonal elements of the PDC matrix appear slightly deflated, the off-diagonal regions contain more information, indicating that PDC spreads variance more evenly across directions.

This redistribution leads to a more regularized and stable covariance structure. In practical terms, while MLE tends to concentrate variance into a few dominant directions (which can exaggerate the importance of certain PCs), PDC balances the spectrum more smoothly. This is beneficial in high-dimensional PCA, where interpretability and stability of components are often more important than exact recovery of the covariance matrix.

Table 5.2 presents the average percentage of total explained variance for both the first $n-1$ PCs and the last $p-n+1$ PCs in selected HD settings where $n < p$. While the numerical differences between MLE and PDC are extremely small (on the order of 10^{-14}), they reflect a consistent trend

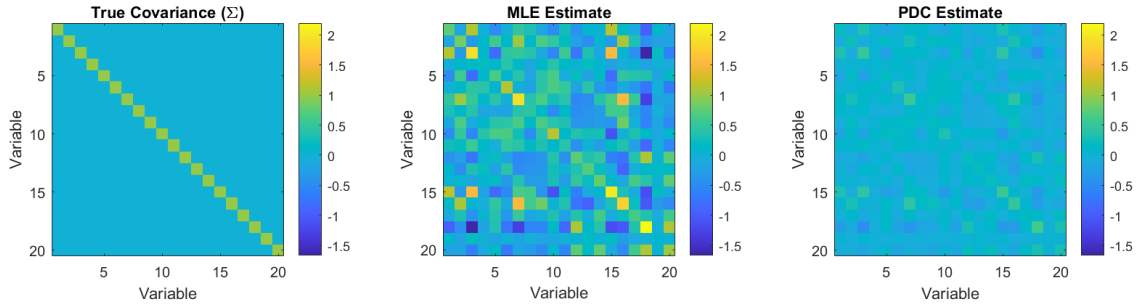


Figure 5.2: Comparison of Σ, MLE, and PDC

toward reduced overdispersion in the leading PCs and reduced underdispersion in the trailing PCs.

Table 5.2: Advantages of PDC Over MLE in HD Analysis

Average Percentage of Total Explained Variance of First (n-1) PCs				
p	**n**	**MLE**	**PDC**	**Improvement**
20	7	0.99999999999999993793815041043000	0.99999999999999995200000000000000	**2.73815E-14** %
20	10	0.99999999999999998429190797270000	0.99999999999999997686631119000000	**7.42559E-15** %
20	12	0.99999999999999999981299359892000	0.99999999999999998373551713000000	**1.60775E-14** %
Average Percentage of Total Explained Variance of Last (p-n+1) PCs				
p	**n**	**MLE**	**PDC**	**Improvement**
20	7	0.00000000000000002061849589570000	0.00000000000000004800000000000000	**2.73815E-14** %
20	10	0.00000000000000001570809202730000	0.00000000000000002313368881000000	**7.42559E-15** %
20	12	0.00000000000000000018700640108000	0.00000000000000001626448287000000	**1.60775E-14** %

Although these differences are numerically marginal, they consistently indicate that PDC provides a more balanced variance distribution across PCs. These findings align with the goals of methods like Ledoit–Wolf, which redistribute variance more evenly to mitigate eigenvalue overdispersion.

To better understand PDC's behavior in finite samples, Table 5.3 focuses specifically on the first PC. As highlighted in Chapter 2, when $n < p$, the first sample PC tends to explain an inflated proportion of variance due to overfitting. Table 5.3 shows that PDC systematically reduces this overdispersion compared to MLE and produces a closer match to the population principal direction.

As seen in Table 5.3, while the numeric gains are subtle (in the range of 10^{-15}), PDC consistently exhibits slightly lower overdispersion, more accurate variance allocation, and reduced CSE than MLE. These trends, though small in absolute terms, are persistent and theoretically grounded.

Table 5.3: Average First PC Metrics from PDC Estimator

p	n	Average Overdispersion of first PC		
		Pop	**MLE**	**PDC**
20	7	0	0.00494667088113278589217 5807	0.00494667088113278628646 3011
20	10	0	0.00189765050015569037601 6815	0.00189765050015564700792 9915
20	12	0	0.00054063910439334486503 90949	0.00054063910439337926145 191
20	16	0	0.00053086891536434648204 70088	0.00053086891536434345558231 115
20	17	0	0.00025638637217713746482 14503	0.00025638637217713350748352 07
p	**n**	**Average Percentage of Explained Variance of first PC**		
		Pop	**MLE**	**PDC**
20	7	76.50551356458754526101984	80.3577872624701825543525	80.3577872624701541326430 7
20	10	76.50551356458754526101984	79.4277434067971910280903 1	79.4277434067971626063808 8
20	12	76.50551356458754526101984	78.2299018834308981240610 6	78.2299018834308391320634
20	16	76.50551356458754526101984	78.5008874306168138446082 6	78.5008874306167996337535 4
20	17	76.50551356458754526101984	77.9376766060800605373515 3	77.9376766060800463264968 1
p	**n**	**Average CSE of first PC**		
		Pop	**MLE**	**PDC**
20	7	0	0.0488621686827150347731 6742	0.0488621686827150070175918
20	10	0	0.0218772251743449691607 7791	0.0218772251743449656913309 6
20	12	0	0.0188545311314782179645 8103	0.0188545311314782110256871 3
20	16	0	0.0122259967833255236285 8276	0.0122259967833255166896888 6
20	17	0	0.0113831451033773976833 0856	0.0113831451033773976833 0856

As discussed in Chapter 4, perturbation theory and sensitivity analysis show that small changes in input (e.g., covariance structure) can meaningfully impact PC directions, especially when the eigenvalue gaps are small.

Thus, even marginal improvements in covariance estimation, such as those offered by PDC, can contribute to more stable and interpretable PCA in practice. This is especially important in fields like genomics, finance, and health sciences, where improved direction estimation can influence downstream analysis and decision-making.

Remark. *While the numerical differences between MLE and PDC shown in Tables 5.1 and 5.2 appear negligible (on the order of 10^{-15}), their consistency across all settings highlights a meaningful advantage in high-dimensional PCA when $n < p$. In such settings, leading PCs are highly sensitive to even slight variations in the estimated covariance matrix. According to perturbation theory, small changes in the covariance structure, especially when leading eigenvalues are close, can cause noticeable shifts in eigenvectors (principal directions).*

To illustrate this, consider a simulated gene expression dataset with $p = 20$ genes and $n = 7$ experimental conditions, generated as $\mathbf{X}_i \sim \mathcal{N}_p(\mathbf{0}, \mathbf{\Sigma})$. The true covariance matrix $\mathbf{\Sigma}$ is structured to reflect biologically plausible gene correlations. PCA is applied using covariance matrices estimated by both MLE and PDC.

In this case, the leading eigenvalues of the MLE covariance matrix are $\lambda_1 = 15.8$ and $\lambda_2 = 15.2$, with a gap of $\Delta = 0.6$. The Frobenius norm between MLE and PDC covariance matrices is about 2.7×10^{-13}. Although this difference does not significantly affect the cumulative variance used to determine the number of PCs (e.g., 80% threshold), it does influence the estimated PC directions. For instance, according to the Davis–Kahan $\sin\Theta$ theorem:

$$\sin\theta \leq \frac{\|\widehat{\mathbf{\Sigma}}_{MLE} - \widehat{\mathbf{\Sigma}}_{PDC}\|_2}{\Delta} \approx \frac{2.7 \times 10^{-13}}{0.6} \approx 4.5 \times 10^{-13}$$

Even such small directional deviations can reorder genes with similar loadings, affecting interpretation, clustering, and enrichment analysis. In real-world genomics applications, such as identifying cancer subtypes or predicting treatment response, these changes may result in different gene sets being emphasized, leading to divergent biological or clinical conclusions.

Moreover, as shown in Table 5.2, PDC consistently yields slightly lower Corrected Subspace Error (CSE) values than MLE (e.g., 0.02187 vs. 0.02188 for $n = 10$). While the numerical gain is modest, it reflects more stable alignment with population PCs and reduced sensitivity to small-sample noise.

Most importantly, PDC acts as a form of regularization that suppresses overdispersion and stabilizes the variance distribution across PCs. While this may not change the number of PCs selected in cumulative variance plots, it creates conditions under which selecting a smaller number of more stable PCs becomes more reliable and justifiable.

Therefore, even though the improvement does not directly impact PC count selection, it significantly enhances:

- *the orientation and interpretability of PCs,*
- *the stability of downstream clustering or classification results,*
- *the reproducibility of findings across subsamples or validation cohorts,*
- *and the biological insight drawn from PC-associated gene loadings.*

These benefits validate the use of regularized estimators like PDC in high-dimensional low-sample-size PCA applications, particularly in fields such as genomics, medicine, and finance.

5.3 PCA using Regularized PDCs

When $n < p$, we cannot extract more than n meaningful PCs from the sample covariance matrix, and the leading components are often subject to overdispersion. To overcome this limitation without increasing the sample size, an option that may be costly or impractical, shrinkage methods can be used to regularize the eigenvalue spectrum of the sample covariance matrix. One of the most widely used approaches in this context is the Ledoit–Wolf estimator, which shrinks the sample covariance toward a structured target (such as a scaled identity matrix), thereby stabilizing the eigenvalues. As reviewed in Chapter 3, although various alternatives exist (not all specifically designed for PCA), the Ledoit–Wolf method remains one of the most effective tools for high-dimensional analysis when $n < p$.

However, in the context of PCA for HD settings with $n < p$, Ledoit-Wolf does not significantly outperform the traditional MLE. Its uniform shrinkage of the last $p - n + 1$ eigenvalues leads to an unnecessary overestimation of these PCs and causes substantial underdispersion of the first PC. This ensures that the last $p - n + 1$ eigenvalues do not approach zero or nearly zero values, drastically reducing the first PC.

Even if the first PC is crucial in PCA, according to Ledoit Wolf covariance method does not capture the actual amount of variance in the data due to Ledoit-Wolf shrinking the eigenvalues excessively and causing under dispersion of first n -1 PCs and loss of valuable information. The unnecessary uniform distribution of last p-n+1 PCs may affect this. However, despite its drawbacks, MLE is preferred to preserve the importance of the first PC. That's why we still use MLE for the covariance estimation $n < p$ data settings as well. Considering all these issues, our proposed covariance estimation method PDC balances eigenvalues effectively without increasing the sample size, showing slight improvements in PCA estimates and addressing the limitations of traditional methods when $n < p$ data settings.

Given the limited improvement in PDC estimation, in Chapter 04, we proposed four different regularized PDCs to enhance their performance. In the following sections, we will evaluate the performance of these regularizations to determine their effectiveness.

Table 5.4: Average Overdispersion of first PC

p	n	Average Overdispersion of first PC							
		POP	**MLE**	**LW**	**PDC**	**SPDC**	**LSPDC**	**MAXPDC**	**RPDC**
20	3	0.000000	0.149947	0.037591	0.149947	0.005540	0.003689	0.066781	0.005602
20	4	0.000000	0.040505	0.140737	0.040505	0.001819	0.001936	0.008969	0.001664
20	5	0.000000	0.017546	0.126138	0.017546	0.001741	0.001553	0.002676	0.001779
20	6	0.000000	0.008383	0.096425	0.008383	0.001780	0.002146	0.000693	0.001496
20	7	0.000000	0.004947	0.076989	0.004947	0.001133	0.001174	0.000430	0.000996
20	8	0.000000	0.004130	0.053602	0.004130	0.000349	0.000380	0.000406	0.000270
20	9	0.000000	0.002951	0.041354	0.002951	0.000232	0.000266	0.000357	0.000195
20	10	0.000000	0.001898	0.036120	0.001898	0.000217	0.000231	0.000118	0.000179
20	11	0.000000	0.001543	0.027566	0.001543	0.000261	0.000273	0.000093	0.000225
20	12	0.000000	0.000541	0.026169	0.000541	0.000696	0.000718	0.000018	0.000658
20	13	0.000000	0.000900	0.018428	0.000900	0.000084	0.000087	0.000097	0.000061
20	14	0.000000	0.000366	0.016601	0.000366	0.000269	0.000292	0.000008	0.000211
20	15	0.000000	0.000410	0.013254	0.000410	0.000152	0.000163	0.000014	0.000137
20	16	0.000000	0.000531	0.010264	0.000531	0.000047	0.000057	0.000032	0.000031
20	17	0.000000	0.000256	0.010572	0.000256	0.000107	0.000107	0.000005	0.000093
20	18	0.000000	0.000262	0.007587	0.000262	0.000051	0.000065	0.000018	0.000030
20	19	0.000000	0.000243	0.006996	0.000243	0.000043	0.000047	0.000018	0.000034
20	20	0.000000	0.000260	0.005862	0.000260	0.000038	0.000041	0.000016	0.000030

Note: Best Second Best Third Best

Table 5.4 shows that the lowest average overdispersion of the first PC is achieved by MAXPDC, followed by RPDC and SPDC, respectively. Compared to the standard MLE method, all these regularized PDC methods significantly reduce the overdispersion of the first PC. This reduction in overdispersion becomes even more pronounced as the number of observations n decreases relative to the number of dimensions p. The findings also indicate that Ledoit-Wolf covariance estimation is ineffective for PCA in HD settings where $n < p$. This supports our previous assertion that we continue to use MLE despite the existence of alternative covariance estimation methods. These existing alternatives, while innovative, offer a different level of reliability in preserving the importance of the first PC in such HD scenarios.

Table 5.5 reveals a similar pattern of results. Specifically, to those observed in Table 5.4, Table 5.5 shows the average percentage of explained variance of the first PC. As previously mentioned, the first PC is crucial in PCA. If the first PC is excessively overdispersed or underdispersed, it negatively impacts the PCA results.

Table 5.5: Average Percentage of Explained Variance of first PC

p	n	Average Percentage of Explained Variance of first PC							
		POP	**MLE**	**LW**	**PDC**	**SPDC**	**LSPDC**	**MAXPDC**	**RPDC**
20	3	76.50551	88.75079	70.37435	88.75079	78.85917	78.42632	84.67747	78.87246
20	4	76.50551	84.30026	61.97607	84.30026	74.85354	74.80138	80.17347	74.92578
20	5	76.50551	82.42940	60.62233	82.42940	74.63940	74.74337	78.81875	74.61925
20	6	76.50551	81.08342	60.97936	81.08342	74.39629	74.18939	77.82131	74.57150
20	7	76.50551	80.35779	61.30794	80.35779	74.66194	74.62862	77.64125	74.77719
20	8	76.50551	80.30731	62.80860	80.30731	75.40000	75.35222	77.69829	75.53381
20	9	76.50551	79.94109	63.64406	79.94109	75.54245	75.47423	77.70081	75.62140
20	10	76.50551	79.42774	63.75641	79.42774	75.51785	75.48604	77.23406	75.60792
20	11	76.50551	79.28301	64.76541	79.28301	75.36271	75.33792	77.18822	75.44423
20	12	76.50551	78.22990	64.50851	78.22990	74.54889	74.51772	76.18879	74.60364
20	13	76.50551	78.82873	65.99027	78.82873	75.79672	75.78465	77.26744	75.90099
20	14	76.50551	78.04745	66.11768	78.04745	75.18417	75.12766	76.72951	75.33387
20	15	76.50551	78.19866	66.87325	78.19866	75.47357	75.43793	76.81705	75.52452
20	16	76.50551	78.50089	67.73184	78.50089	75.91402	75.85280	76.99270	76.01952
20	17	76.50551	77.93768	67.30892	77.93768	75.57830	75.57884	76.71394	75.64455
20	18	76.50551	77.99916	68.47479	77.99916	75.84484	75.76460	76.89616	75.99961
20	19	76.50551	77.98393	68.57079	77.98393	75.88555	75.85635	76.90846	75.94862
20	20	76.50551	78.07642	69.04307	78.07642	75.90318	75.88215	76.89435	75.96801

Note: Best Second Best Third Best

When examining the average percentage of explained variance of the first PC across different techniques, it is evident that the regularized versions of PDC provide much closer finite sample estimates for the first PC than the MLE method. Among these, MAXPDC offers the closest estimates for the first PC, followed by RPDC and SPDC. On the other hand, when n decreases relative to the p, the LW method tends to massively under-disperse the first PC. While the other methods might overdisperse the first PC, LW underdisperse the first PC due to uniform shrinkage. This highlights why regularized PDC methods are generally preferred over LW in HD settings where $n < p$ offers more reliable and accurate PCA results.

Tables 5.4 and 5.5 highlight the challenges of PCA in HD settings where $n < p$, particularly concerning the magnitude of PCs. Mathematically, PCs are vectors, possessing both magnitude and direction. While Tables 5.4 and 5.5 address the magnitude, Table 5.6 examines how closely the sample PCs align with the population parameters direction-wise.

Table 5.6: Average CSE of First PC (All Estimators)

p	n	Average CSE of first PC							
		POP	**MLE**	**LW**	**PDC**	**SPDC**	**LSPDC**	**MAXPDC**	**RPDC**
20	3	0.00000	0.21556	0.31244	0.21556	0.18558	0.19727	0.21556	0.19138
20	4	0.00000	0.13307	0.18955	0.13307	0.11066	0.10841	0.13307	0.10967
20	5	0.00000	0.08324	0.10860	0.08324	0.05233	0.05470	0.08324	0.05159
20	6	0.00000	0.06510	0.06591	0.06510	0.03795	0.03939	0.06510	0.03871
20	7	0.00000	0.04886	0.04886	0.04886	0.02026	0.02101	0.04886	0.02339
20	8	0.00000	0.03329	0.03329	0.03329	0.01179	0.01276	0.03329	0.01575
20	9	0.00000	0.02963	0.02963	0.02963	0.01431	0.01502	0.02963	0.01830
20	10	0.00000	0.02188	0.02188	0.02188	0.00900	0.00952	0.02188	0.01344
20	11	0.00000	0.02067	0.02067	0.02067	0.00862	0.00899	0.02067	0.01326
20	12	0.00000	0.01885	0.01885	0.01885	0.00786	0.00816	0.01885	0.01232
20	13	0.00000	0.01568	0.01568	0.01568	0.00681	0.00712	0.01568	0.01128
20	14	0.00000	0.01479	0.01479	0.01479	0.00688	0.00712	0.01479	0.01121
20	15	0.00000	0.01402	0.01402	0.01402	0.00666	0.00681	0.01402	0.01125
20	16	0.00000	0.01223	0.01223	0.01223	0.00616	0.00634	0.01223	0.01060
20	17	0.00000	0.01138	0.01138	0.01138	0.00604	0.00619	0.01138	0.01044
20	18	0.00000	0.01068	0.01068	0.01068	0.00581	0.00589	0.01068	0.01038
20	19	0.00000	0.00997	0.00997	0.00997	0.00570	0.00578	0.00997	0.01011
20	20	0.00000	0.00953	0.00953	0.00953	0.00567	0.00577	0.00953	0.00994

Note: Best Second Best Third Best

We use the CSE of the first PC to evaluate the sample PC directions. Table 5.5 shows that SPDC achieves the lowest CSE, indicating that its sample PC estimates are the closest to the population parameters. Following SPDC, LSPDC and RPDC also perform well, showing low CSE. Notably, all these regularized PDC methods, except for MAXPDC, demonstrate lower CSE for the first PC than the standard MLE method.

Interestingly, MAXPDC yields a similar CSE for the first PC, which is comparable to MLE. This observation highlights the effectiveness of the regularized PDC methods (excluding MAXPDC) in providing more accurate sample PC directions in $n < p$ data settings. Overall, these findings suggest that regularized PDC methods enhance the reliability of PCA by better aligning sample PCs with their actual population counterparts.

As vectors, we are concerned with the PCs' direction and magnitude. However, in the context of PCA, the direction of the PCs is often considered more critical because it defines the axis along which the data has the most variance. The direction of the PCs helps in understanding the structure of the data and identifying the underlying patterns. While also important, the magnitude indicates the amount of variance captured by each PC. Therefore, both aspects are essential, but the direction is usually more emphasized when interpreting PCA results.

Based on this understanding, SPDC achieves the lowest CSE for the first PC, indicating its superiority in preserving the direction of the PCs. This low CSE means that SPDC provides sample PCs that closely align with the actual population parameters in direction. However, when examining the percentage of variance captured by the first PC, SPDC ranks third. In terms of magnitude, MAXPDC takes the first place. Despite this, MAXPDC does not show an improvement in CSE compared to MLE, with both methods yielding similar CSE values. Consequently, MAXPDC is not considered an improvement in the magnitude and direction of the sample PCs.

RPDC, on the other hand, ranks second in magnitude but third in direction. When comparing both the direction and magnitude of the sample PCs, with a priority on the direction, **SPDC demonstrates the most significant improvement in PCA for $n < p$ data settings**. This makes SPDC the most effective method among those evaluated, as it provides a better balance between capturing the variance (magnitude) and aligning with the actual direction of the PCs.

5.4 Experiments on Public Data

In this study, we analyze a dataset from Phytozome [73], a comprehensive resource for comparative genomics of green plants. The dataset contains gene expression data for *Brachypodium distachyon*, a model grass species, under various experimental conditions. It includes gene expression levels, measured in FPKM (Fragments Per Kilobase of transcript per Million mapped reads), for approximately 34,000 genes across 74 different experimental conditions. These conditions involve different plant tissues such as shoots, leaves, and stems, and various experimental treatments like light/dark cycles, cold treatments, and different nutrient supplies. The dataset is organized with genes as rows and experimental conditions as columns. A unique GeneID identifies each gene, and the columns

represent different tissue types and treatment combinations. This setup provides a comprehensive view of gene expression profiles under diverse conditions, allowing researchers to understand how genes respond to varying treatments.

As we observed in simulation studies, in scenarios where the number of experimental conditions (i.e., samples or columns, denoted by n) is smaller than the number of genes (i.e., features or rows, denoted by p), covariance estimation and PCA can become ill-conditioned. This can lead to overdispersed covariance estimates and unstable PCs. Consequently, a PCA model may capture noise rather than biologically meaningful patterns, potentially resulting in misleading interpretations. Techniques such as dimensionality reduction (DR) can help address these challenges.

To evaluate the effectiveness of our proposed methods under such high-dimensional settings, we perform a detailed experimental study using this gene expression data. Specifically, we fix the number of variables at $p = 74$ and vary the number of samples n (i.e., experimental conditions) from 5 to 15. For each value of n, we generate 100 random subsamples by selecting n experimental conditions (columns) from the full dataset, while keeping the same set of 74 genes fixed across all iterations. This results in 100 distinct $n \times 74$ datasets for each n, which ensures robust and statistically meaningful results.

This study enables us to assess the performance of our proposed methods for PCA in a realistic high-dimensional context, where $n < p$, as commonly encountered in gene expression data analysis.

As illustrated in Table 5.6, the average overdispersion of the first PC was analyzed for various $n < p$ HD settings of gene expression data. The results indicate a clear trend: as n decreases relative to p, the average overdispersion of the first PC increases.

When examining the MLE approach, it becomes evident that the overdispersion is significantly higher in these $n < p$ settings. However, on the other hand, although the Ledoit-Wolf covariance estimator is a widely used alternative to MLE, in high-dimensional settings where $n < p$, it often performs worse than MLE in PCA. This is because its uniform shrinkage tends to overly suppress the leading eigenvalues, resulting in underestimation of the variance explained by the first few PCs. This suggests that Ledoit-Wolf covariance estimation is unsuited for PCA in HD settings where $n < p$.

Table 5.7: Average Overdispersion of first PC — Phytozome Data

p	n	Average Overdispersion of first PC							
		POP	**MLE**	**LW**	**PDC**	**SPDC**	**LSPDC**	**MAXPDC**	**RPDC**
74	5	0.000000	1.095071	0.726472	1.095071	0.407725	0.424260	0.620553	0.408652
74	6	0.000000	0.862375	1.143960	0.862375	0.265933	0.272175	0.392784	0.270337
74	7	0.000000	0.707309	1.389852	0.707309	0.240719	0.254671	0.350673	0.237093
74	8	0.000000	0.575631	1.497910	0.575631	0.201135	0.207158	0.282708	0.192903
74	9	0.000000	0.463153	1.576263	0.463153	0.162813	0.162324	0.211596	0.157926
74	10	0.000000	0.469671	1.619404	0.469671	0.145126	0.150952	0.185968	0.136837
74	11	0.000000	0.418005	1.577441	0.418005	0.133369	0.135252	0.173302	0.125027
74	12	0.000000	0.316212	1.564542	0.316212	0.108186	0.106666	0.139324	0.101614
74	13	0.000000	0.353287	1.618222	0.353287	0.088806	0.085105	0.115271	0.084781
74	14	0.000000	0.288695	1.502158	0.288695	0.081009	0.083815	0.114219	0.073784
74	15	0.000000	0.270646	1.339662	0.270646	0.077814	0.077240	0.124319	0.072625

Note: Best Second Best Third Best

Given the marginal improvement, the PDC method focuses on its regularized versions. The RPDC demonstrates the lowest average overdispersion for the first PC, making it the most effective. Apart from that, the SPDC and LSPDC methods also show a significant performance than RPDC. But it slightly lesser extent compared to the RPDC. However, both SPDC and LSPDC significantly reduce the overdispersion of first PC and provide more reliable and accurate results compared to typical estimation methods. Table 5.6 highlights the reliability and applicability of RPDC, SPDC, and LSPDC in enhancing PCA outcomes for $n < p$ gene expression data.

The MAXPDC method also reduces the overdispersion of the first PC magnitude-wise. However, it does not reduce the overdispersion of the first PC as effectively as the other regularized PDC methods. Still, MAXPDC performs better than MLE and Ledoit-Wolf.

Overall, RPDC is the best method for minimizing overdispersion in the first PC under $n < p$ conditions, highlighting its potential for improving PCA in $n < p$ HD gene expression data.

Table 5.7 presents the average percentage of explained variance for gene expression data's first PC in selected $n < p$ data settings. As discussed in Table 5.6, the MLE method significantly overestimates the variance of the first PC with an unnecessarily higher percentage of explained variance in HD settings where $n < p$. When the n increases significantly more than the p, the performance of MLE improves. This implies the inapplicability and inaccuracy of using MLE when

$n < p$ data settings. For example, in selected $n < p$ data settings where $n = 5$ and $p = 74$, in the population data, the first PC should capture about 75% of the total variance, whereas MLE estimates it at about 99%, indicating a significant overestimation.

Table 5.8: Average Percentage of Explained Variance of first PC — Phytozome Data

p	n	Average Percentage of Explained Variance of first PC							
		POP	MLE	LW	PDC	SPDC	LSPDC	MAXPDC	RPDC
74	5	74.94184	99.27146	55.12550	99.27146	89.78744	90.08548	93.25670	89.80431
74	6	74.94184	99.08073	47.13994	99.08073	88.34649	88.50289	91.23277	88.45704
74	7	74.94184	98.88958	41.37240	98.88958	88.91245	89.31161	91.80391	88.80681
74	8	74.94184	98.27670	37.29955	98.27670	88.73541	88.94041	91.29501	88.45020
74	9	74.94184	97.31829	33.66149	97.31829	88.20886	88.18895	90.06640	88.00823
74	10	74.94184	98.84208	30.56228	98.84208	88.22734	88.49138	89.98104	87.84238
74	11	74.94184	98.70888	28.77175	98.70888	88.36677	88.46122	90.24518	87.94015
74	12	74.94184	96.62236	26.71663	96.62236	87.62321	87.53378	89.33294	87.23200
74	13	74.94184	98.87711	23.71544	98.87711	86.94224	86.68952	88.61393	86.66717
74	14	74.94184	97.46221	23.57136	97.46221	86.87132	87.07617	89.10710	86.32691
74	15	74.94184	97.56997	24.59807	97.56997	87.07507	87.03026	90.27799	86.66352

Note: Best Second Best Third Best

On the other hand, the Ledoit-Wolf method underestimates the first PC's variance. This may happen because of its uniform shrinkage approach. This shows the limitations of using Ledoit-Wolf covariance estimation when $n < p$ data settings. It does not accurately capture the actual variance-covariance structure in $n < p$ data. These inaccuracies highlight the need for more reliable methods for PCA when dealing with $n < p$ HD data settings.

However, as Table 5.8 shows, our proposed regularized PDC methods can address these issues effectively in $n < p$ data settings. Among them, the RPDC provides the closest estimates of the population parameters compared to all other proposed techniques and traditional techniques of MLE and Ledoit-Wolf. Following the RPDC, the SPDC, and the LSPDC also perform better than those traditional methods, with slightly less improvement than the RPDC. These regularized covariance estimation methods significantly improve the finite sample estimates of PCA over the traditional approaches.

At last, this detailed analysis of the average percentage of explained variance for gene expression data highlights the effectiveness of regularized PDC methods in providing reliable and accurate fi-

nite sample estimates for the PCA when $n < p$ data settings. By addressing the limitations of MLE and Ledoit-Wolf, these regularized PDC methods enhance our ability to interpret and understand HD data accurately.

As we discussed earlier, the direction of the sample PC estimates plays a crucial role in PCA. Table 5.9 presents the CSE of the first PC for each method to evaluate the goodness of fit of these directions. Similar to the distribution of the magnitude of the PCs, the direction of the PCs also deteriorates as the n decreases relative to the p.

Table 5.9: Average CSE of first PC — Phytozome Data

p	n	Average CSE of first PC							
		POP	**MLE**	**LW**	**PDC**	**SPDC**	**LSPDC**	**MAXPDC**	**RPDC**
74	5	0.000000	0.247691	0.247691	0.247691	0.076225	0.079755	0.247691	0.076708
74	6	0.000000	0.248536	0.248536	0.248536	0.073888	0.093020	0.248536	0.075063
74	7	0.000000	0.203118	0.203118	0.203118	0.053610	0.073243	0.203118	0.054907
74	8	0.000000	0.223614	0.223614	0.223614	0.052163	0.056960	0.223614	0.052487
74	9	0.000000	0.198741	0.198741	0.198741	0.051376	0.054183	0.198741	0.053768
74	10	0.000000	0.167903	0.167903	0.167903	0.050993	0.054053	0.167903	0.052861
74	11	0.000000	0.172170	0.172170	0.172170	0.067141	0.068879	0.172170	0.053295
74	12	0.000000	0.208089	0.208089	0.208089	0.051925	0.054721	0.208089	0.054761
74	13	0.000000	0.188317	0.188317	0.188317	0.088795	0.109493	0.188317	0.091722
74	14	0.000000	0.191797	0.183590	0.191797	0.069056	0.071941	0.191797	0.072456
74	15	0.000000	0.195967	0.195967	0.195967	0.067450	0.069309	0.195967	0.072428

Note: Best Second Best Third Best

A careful examination of the CSE values reveals that traditional methods yield higher CSEs than our proposed regularized versions of PDC estimation for the first PC. This indicates that the direction of the PCs estimated by existing methods could be more accurate when $n < p$. However, as shown in Table 5.9, all the regularized PDC estimation methods (except MAXPDC) perform significantly better in estimating the direction of the first PC when $n < p$ data settings. Among them, the SPDC method consists of the lowest CSE for the first PC.

This implies the SPDC's superior capability to provide the most reliable and accurate finite sample PC estimates along with the direction of the population. When we look at more closely, the RPDC method also demonstrates substantial improvement in accuracy, significantly outperforming traditional approaches such as the MLE and Ledoit-Wolf methods. On the other hand, LSPDC

method's finite sample estimates of PC are slightly less accurate than SPDC and RPDC. But, still shows considerable enhancements in the direction of PC estimation compared to traditional techniques. These results highlight the effectiveness of our regularized PDC methods, particularly SPDC, in yielding reliable and accurate finite sample PC estimations for the first PC with low CSE, affirming their superiority in $n < p$ data analysis. When comparing the direction and magnitude of the sample PC estimates generated by existing methods, the Ledoit-Wolf method must effectively tackle the issue of overdispersion in the first PC when pitted against the MLE method. However, the MLE method itself generates sample PC estimates that are significantly overdispersed.

However, our proposed regularized versions of PDC estimations are highly effective in mitigating the overdispersion of the first PC. Among these, the SPDC method demonstrates the most substantial improvement by prioritizing the direction of the sample estimates. Following SPDC, the RPDC method also effectively addresses the overdispersion issue, albeit slightly less. Lastly, the LSPDC method provides notable improvements over MLE in managing the overdispersion of the first PC.

In summary, our proposed regularized PDC methods, particularly SPDC, significantly improve the accuracy of PCA in $n < p$ settings by reducing the overdispersion of the first PC, thereby offering a more reliable and robust alternative to traditional methods such as MLE and Ledoit-Wolf covariance estimation.

5.5 Summary and Discussion

This chapter looks at how well PCA works with different covariance estimation methods in HD data settings where $n < p$. In this context, traditional covariance estimation techniques often struggle due to the issue of overdispersion of the sample estimates. This leads to the need to explore more robust and reliable methods of covariance estimation for $n < p$ data settings.

To minimize the overdispersion of PCA when $n < p$ data settings, we suggest a new covariance estimation method called PDC along with its four regularized versions: SPDC, LSPDC, MAXPDC, and RPDC. These methods are benchmarked against traditional techniques such as MLE covariance estimation and Ledoit-Wolf covariance estimation, widely regarded as reliable standards in practical applications. In this Chapter, we evaluate how these methods perform in terms of improving the reliability and accuracy of PCA results when $n < p$ data settings.

Ledoit-Wolf covariance estimation is renowned for enhancing the robustness and stability of the explained variance in PCA. It achieves this by applying a shrinkage technique that combines the sample covariance matrix with a structured estimator, typically a multiple of the identity matrix. This approach overcomes the risk of overfitting and reduces sampling variability. However, it leads to underdispersion of first n - 1 PCs and overdisoersion of last p-n+1 PCs due to its uniform shrinkage of sample eigenvalues. Moreover, the Ledoit-Wolf method primarily targets stabilizing eigenvalues of the sample covariance matrix rather than adjusting the direction of the eigenvectors or PCs. This is because the shrinkage technique focuses on regularizing the magnitude of the covariance estimates, thereby improving the overall robustness of the explained variance without fundamentally changing the angles of the PCs.

On the other hand, the proposed PDC estimation method introduces a novel approach that consistently and marginally reduces the overdispersion of the first PC when $n < p$ data settings. However, these estimated components are slightly closer to the MLE method's sample estimates. Therefore, this improvement is marginally small (on the order of 10^{-15}), providing a more balanced variance distribution across all PCs. This improvement is significant in HD settings where the first PC often dominates the explained variance. By achieving a more balanced variance distribution, PDC enhances PCA performance in terms of both the magnitude and direction of the PCs. The too small but persistent improvements offered by PDC make it a valuable addition to the toolkit of covariance estimation methods for PCA when $n < p$ data settings.

All the regularized versions of PDC, SPDC, LSPDC, MAXPDC, and RPDC, demonstrate significant improvements over traditional covariance estimation methods. Among these, SPDC stands out for its ability to effectively regularize the magnitude and direction of the sample PCs compared to MLE. This method of covariance estimations significantly reduces the overdispersion of the first few PCs, as SPDC redistributes the excess variance from the first n - 1 PCs more evenly among the subsequent PCs. The enhanced regularization provided by SPDC leads to a more balanced spread of variance, thereby improving the overall PCA performance. Following SPDC, RPDC, LSPDC, and MAXPDC also show substantial improvements, each contributing unique strengths to the regularization process.

Chapter 6

Efficacy of PDC Estimation in PCA Dynamics with $n < p$

6.1 Introduction

In the R&D industry, PCA is widely used as an intermediate statistical data analysis tool due to its ability to simplify HD data. It offers several advantages, including data visualization, pattern recognition, noise reduction, dimensionality reduction (DR), and the discovery of underlying structures among variables. These benefits make PCA a powerful tool for enhancing statistical interpretation and guiding further analysis.

However, the reliability of PCA results, especially in HD settings where the number of variables p exceeds the number of observations n, can be significantly affected by various factors. These include the sample size, dimensionality, the structure of the covariance matrix, the level of correlation among variables, and the presence of extreme values or outliers. Understanding how each of these factors influences PCA performance is critical when selecting an appropriate covariance estimation method.

In Chapter 5, we focused on controlled simulations using multivariate normal data with structured random covariance matrices to evaluate the performance of the proposed PDC estimator and its regularized versions in standard $n < p$ HD scenarios. That chapter served as a baseline to establish the theoretical and empirical behavior of the estimators under idealized conditions.

Chapter 6 builds on that foundation by extending the experimental framework to more complex and practical settings. Here, we explore how the proposed methods perform across a broader range of conditions, including different combinations of n and p, varying degrees of correlation, multiple covariance structures, and the inclusion of extreme values. We also incorporate real-world gene expression data to evaluate the robustness and generalizability of each method. The goal of this chapter is to assess whether the advantages observed in controlled simulations translate to more realistic and challenging data environments.

6.2 No: of Observations (n) and Dimensions (p)

In general, gathering more clues to solve a mystery helps to reveal the complete real story more clearly. Statistics is also doing the same thing. More observations help to understand the whole picture with greater clarity. Therefore, the limited sample size may produce unreliable sample statistics. Similarly, PCA also gives misleading statistics when $n < p$ data settings. In Chapter 5, we observed that issue in PCA. However, in this section, we aim to deeply evaluate the performance of PCA when $n < p$ data settings with different settings of n and p.

6.2.1 Experiments on Synthetic Data

To evaluate the performance of covariance estimation methods under HD, low-sample-size conditions, I generated synthetic datasets as follows. For each setting of $p \in \{20, 50, 100, 250, 500\}$, I fixed the population mean vector as $\boldsymbol{\mu} = \texttt{zeros}(1, p)$. The population covariance matrix $\boldsymbol{\Sigma}$ was created using $\boldsymbol{\Sigma} = \texttt{t}^\top\texttt{t}$, where $\texttt{t} \in \mathbb{R}^{p\times p}$ is a matrix with entries independently drawn from a uniform distribution on [0, 1]. This ensures a dense and positive definite structure for $\boldsymbol{\Sigma}$. For each sample size $n \in \{3, 4, \ldots, 12\}$, I drew n samples from the multivariate normal distribution $\boldsymbol{X}_i \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$ using MATLAB's `mvnrnd` function. These simulations allowed me to assess the impact of varying p/n ratios on the percentage of explained variance of the first PC and the average Cosine Similarity Error (CSE) across different covariance estimators. This setup was designed to isolate the influence of sample size and dimensionality in a controlled, non-sparse population structure. Each experimental configuration was replicated 500 times to ensure reliable and stable estimation of the performance metrics.

According to Table 6.1, the percentage of explained variance of the first PC is illustrated for seven different covariance estimation methods across three dimension sizes ($p = 20, 50, 100, 250$ and 500) and ten different sample sizes ($n = 3, 4, 5, \ldots, 12$). As the sample size n decreases and the dimension size p increases (p/n increases), the MLE produces a significantly overdispersed estimation for the first PC with an overwhelming percentage of explained variance of the first PC.

Table 6.1: Average Percentage of Explained Variance of first PC vs. n and p

p	n	Average Percentage of Explained Variance of first PC							
		Pop	**MLE**	**LW**	**PDC**	**SPDC**	**LSPDC**	**MAXPDC**	**RPDC**
20	3	76.50551	88.75079	70.37435	88.75079	78.85917	78.42632	84.67747	78.87246
	4	76.50551	84.30026	61.97607	84.30026	74.85354	74.80138	80.17347	74.92578
	5	76.50551	82.42940	60.62233	82.42940	74.63940	74.74337	78.81875	74.61925
	6	76.50551	81.08342	60.97936	81.08342	74.39629	74.18939	77.82131	74.57150
	7	76.50551	80.35779	61.30794	80.35779	74.66194	74.62862	77.64125	74.77719
	8	76.50551	80.30731	62.80860	80.30731	75.40000	75.35222	77.69829	75.53381
	9	76.50551	79.94109	63.64406	79.94109	75.54245	75.47423	77.70081	75.62140
	10	76.50551	79.42774	63.75641	79.42774	75.51785	75.48604	77.23406	75.60792
	11	76.50551	79.28301	64.76541	79.28301	75.36271	75.33792	77.18822	75.44423
	12	76.50551	78.22990	64.50851	78.22990	74.54889	74.51772	76.18879	74.60364
50	3	76.60596	88.11281	69.73971	88.11281	77.49092	77.04382	84.00757	77.50405
	4	76.60596	84.51352	62.79010	84.51352	75.21394	75.06560	80.33784	75.21606
	5	76.60596	82.57106	61.07092	82.57106	74.32349	74.21050	78.67589	74.36999
	6	76.60596	80.31944	59.80071	80.31944	73.76954	73.70229	77.45700	73.79094
	7	76.60596	80.31521	61.88899	80.31521	74.19491	74.09131	77.47788	74.33685
	8	76.60596	79.68861	61.75699	79.68861	74.53451	74.53833	77.21520	74.55416
	9	76.60596	79.31896	62.39109	79.31896	74.27456	74.29635	76.72202	74.25787
	10	76.60596	78.95381	63.44282	78.95381	75.11088	75.07665	77.05695	75.16198
	11	76.60596	78.62235	64.30344	78.62235	74.93514	74.89388	76.69809	75.03220
	12	76.60596	78.61946	64.89355	78.61946	75.08699	75.08178	76.84154	75.09399
	3	75.55752	88.13808	69.79489	88.13808	77.74497	77.30556	84.02361	77.76533
	4	75.55752	85.01577	63.66611	85.01577	75.26589	75.06494	80.88597	75.37439
	5	75.55752	82.29416	60.67239	82.29416	74.47783	74.44207	78.84975	74.48136
	6	75.55752	80.83335	60.24835	80.83335	74.51536	74.40882	78.03097	74.55732

100	7	75.55752	79.71522	60.81518	79.71522	73.78024	73.68028	76.84093	73.85493
	8	75.55752	79.47078	61.43919	79.47078	74.84674	74.84410	77.24792	74.87664
	9	75.55752	78.86205	62.40137	78.86205	74.04328	73.98276	76.28600	74.03778
	10	75.55752	78.58605	62.98040	78.58605	74.69534	74.68853	76.60948	74.70885
	11	75.55752	78.91270	64.54176	78.91270	75.52864	75.47218	77.12464	75.55286
	12	75.55752	77.88307	63.98741	77.88307	74.68579	74.67865	76.21424	74.67435
250	3	75.09168	87.80937	69.43528	87.80937	77.67352	77.21254	84.25049	77.69030
	4	75.09168	82.75874	61.03814	82.75874	73.05843	72.89072	79.00870	73.05287
	5	75.09168	81.09516	59.46378	81.09516	73.20684	73.15595	77.82870	73.20092
	6	75.09168	80.62660	60.46393	80.62660	73.81570	73.75136	77.60364	73.85561
	7	75.09168	79.36755	60.56608	79.36755	73.47743	73.40143	76.69815	73.55436
	8	75.09168	78.81642	60.91922	78.81642	73.51577	73.49296	76.26899	73.53068
	9	75.09168	78.37932	61.61118	78.37932	73.91333	73.91000	76.22642	73.94568
	10	75.09168	77.35291	61.79549	77.35291	73.15371	73.11409	75.25398	73.14900
	11	75.09168	77.86471	63.49238	77.86471	74.00792	73.93887	75.83263	74.03140
	12	75.09168	78.24892	64.42096	78.24892	74.63695	74.62390	76.27784	74.62315
500	3	75.11448	86.97284	68.53411	86.97284	75.93561	75.48069	83.09469	75.95286
	4	75.11448	83.38426	61.06861	83.38426	73.28201	73.28368	79.27272	73.20280
	5	75.11448	81.35330	59.84417	81.35330	72.84244	72.73370	77.65118	72.85820
	6	75.11448	79.78672	59.55781	79.78672	73.04644	72.89150	76.92268	73.07271
	7	75.11448	79.29374	60.23641	79.29374	73.70100	73.63447	76.82643	73.74559
	8	75.11448	77.93066	58.97415	77.93066	72.64195	72.77754	75.43911	72.57151
	9	75.11448	77.50460	60.74721	77.50460	72.89808	72.90474	75.38142	72.91244
	10	75.11448	77.32555	61.74025	77.32555	72.99960	72.96633	75.13359	73.02844
	11	75.11448	77.07542	62.05307	77.07542	73.58820	73.60307	75.29148	73.50745
	12	75.11448	77.58000	63.90568	77.58000	74.47847	74.46689	76.14397	74.50151

Note: Best Second Best Third Best

Note: All values are averaged over 500 replications for each n and p setting.

The widely used LW shrinkage estimator exhibits a notable underdispersion in its estimation of the first PC when p/n is high. This underdispersion suggests that the LW estimator may be too conservative in such settings, failing to reflect the true variance structure of the data. While this conservativeness might reduce noise, it also limits the estimator's ability to capture meaningful directions of variation, particularly in high-dimensional scenarios.

The PDC estimator, although improving upon the MLE by slightly reducing overdispersion, still shows limited visual separation in performance compared to MLE, especially when plotted. This minimal improvement, though statistically present, may not always translate into practical advantages in interpretation or downstream analysis.

However, regularized versions of PDC, specifically SPDC, LSPDC, and RPDC, offer substantial improvements. These estimators provide much closer approximations to the true first PC magnitude, with markedly lower deviation across various ratios p/n. In particular, MAXPDC emerges as the most accurate among all the methods tested, consistently yielding results that align closely with the magnitude of the population PC. Its performance is closely followed by SPDC, which also shows strong results, particularly when moderate to high regularization is beneficial.

These regularized PDC methods strike a promising balance: they control the overdispersion observed in MLE and standard PDC, while avoiding the excessive shrinkage of LW. This balanced behavior becomes increasingly valuable as p/n grows, where traditional methods tend to struggle. The ability of SPDC and related estimators to maintain stability and accuracy in such regimes highlights their practical utility for high-dimensional PCA.

A clear and consistent pattern can be observed in Table 6.2 when analyzing the CSE of the first PC across various data settings. As the sample size n decreases relative to the number of variables p, i.e., in settings where $n < p$, the SPDC method consistently achieves the lowest CSE values. This demonstrates the method's superior capability in accurately estimating the direction of the first PC under high p/n ratios. The robustness of SPDC across such challenging scenarios highlights its effectiveness in minimizing estimation errors and maintaining stability where traditional estimators tend to struggle. These findings provide strong support for SPDC as a reliable choice in high-dimensional, small-sample applications.

Table 6.2: Average CSE of First PC vs. n and p

p	n	Average CSE of first PC							
		Pop	**MLE**	**LW**	**PDC**	**SPDC**	**LSPDC**	**MAXPDC**	**RPDC**
20	3	0	0.215556	0.312436	0.215556	0.185583	0.197269	0.215556	0.191383
	4	0	0.133072	0.189546	0.133072	0.110655	0.10841	0.133072	0.109668
	5	0	0.08324	0.1086	0.08324	0.052334	0.054704	0.08324	0.05159
	6	0	0.065099	0.065913	0.065099	0.037954	0.039387	0.065099	0.038714
	7	0	0.048862	0.048862	0.048862	0.020258	0.021014	0.048862	0.023389
	8	0	0.033288	0.033288	0.033288	0.011785	0.012761	0.033288	0.015748
	9	0	0.029631	0.029631	0.029631	0.014306	0.015019	0.029631	0.018297
	10	0	0.021877	0.021877	0.021877	0.008999	0.00952	0.021877	0.013445
	11	0	0.020672	0.020672	0.020672	0.008625	0.00899	0.020672	0.013258
	12	0	0.018855	0.018855	0.018855	0.007862	0.008164	0.018855	0.012321
50	3	0	0.243088	0.256598	0.243088	0.199371	0.219092	0.243088	0.198319
	4	0	0.157412	0.207248	0.157412	0.127024	0.132646	0.157412	0.128784
	5	0	0.077681	0.113487	0.077681	0.051756	0.053619	0.077681	0.057888
	6	0	0.053751	0.057206	0.053751	0.026801	0.028468	0.053751	0.030074
	7	0	0.03748	0.03748	0.03748	0.014293	0.015562	0.03748	0.017801
	8	0	0.030779	0.030779	0.030779	0.014339	0.0153	0.030779	0.018096
	9	0	0.026403	0.026403	0.026403	0.009142	0.009847	0.026403	0.013096
	10	0	0.020786	0.020786	0.020786	0.006806	0.007316	0.020786	0.01089
	11	0	0.0187	0.0187	0.0187	0.006098	0.00648	0.0187	0.010217
	12	0	0.017254	0.017254	0.017254	0.005707	0.006028	0.017254	0.0099
100	3	0	0.245777	0.247828	0.245777	0.196197	0.2234	0.245777	0.20097
	4	0	0.124219	0.135902	0.124219	0.09175	0.098648	0.124219	0.091817
	5	0	0.072461	0.075529	0.072461	0.043199	0.044882	0.072461	0.045682
	6	0	0.051837	0.051837	0.051837	0.024674	0.029463	0.051837	0.02795

	7	0	0.037226	0.037226	0.037226	0.013692	0.014776	0.037226	0.017356
	8	0	0.028587	0.028587	0.028587	0.0089	0.009775	0.028587	0.012708
	9	0	0.025955	0.025955	0.025955	0.008041	0.008693	0.025955	0.011998
	10	0	0.021518	0.021518	0.021518	0.006151	0.006701	0.021518	0.010265
	11	0	0.01846	0.01846	0.01846	0.005173	0.005529	0.01846	0.009377
	12	0	0.017301	0.017301	0.017301	0.0047	0.005031	0.017301	0.008913
250	3	0	0.223841	0.220099	0.223841	0.194237	0.192762	0.223841	0.196718
	4	0	0.144043	0.150931	0.144043	0.107908	0.113516	0.144043	0.109321
	5	0	0.083377	0.083377	0.083377	0.050398	0.052347	0.083377	0.049092
	6	0	0.053848	0.053848	0.053848	0.026261	0.027960	0.053848	0.029397
	7	0	0.038811	0.038811	0.038811	0.014701	0.015832	0.038811	0.018277
	8	0	0.031474	0.031474	0.031474	0.010545	0.015148	0.031474	0.014351
	9	0	0.025701	0.025701	0.025701	0.007284	0.007952	0.025701	0.011318
	10	0	0.023723	0.023723	0.023723	0.006638	0.007135	0.023723	0.010767
	11	0	0.020102	0.020102	0.020102	0.005357	0.005735	0.020102	0.009615
	12	0	0.017275	0.017275	0.017275	0.004411	0.004714	0.017275	0.008659
500	3	0	0.242270	0.251368	0.242270	0.203895	0.222292	0.242270	0.204684
	4	0	0.170422	0.167211	0.170422	0.136374	0.138270	0.170422	0.137936
	5	0	0.093886	0.093886	0.093886	0.061780	0.063565	0.093886	0.064728
	6	0	0.060907	0.060907	0.060907	0.032833	0.034209	0.060907	0.036190
	7	0	0.036631	0.036631	0.036631	0.012049	0.013132	0.036631	0.015664
	8	0	0.036259	0.036259	0.036259	0.014521	0.015465	0.036259	0.018288
	9	0	0.027366	0.027366	0.027366	0.008066	0.008816	0.027366	0.012026
	10	0	0.023995	0.023995	0.023995	0.006847	0.007343	0.023995	0.011086
	11	0	0.020111	0.020111	0.020111	0.005038	0.005456	0.020111	0.009211
	12	0	0.020933	0.020933	0.020933	0.004023	0.004363	0.020933	0.008238

Note: Best Second Best Third Best

Note: All values are averaged over 500 replications for each n and p setting.

In addition to the strong performance of the SPDC method, both the LSPDC and RPDC estimators also demonstrate considerable accuracy in estimating the first PC direction under various $n < p$ data settings. While their CSE values are slightly higher than those achieved by SPDC, they remain significantly lower than those of traditional estimators such as the MLE and the LW shrinkage estimator. This suggests that LSPDC and RPDC provide effective alternatives for PCA, particularly in settings with high p/n ratio, where maintaining reliable estimation of variance structure is critical. Their consistent performance across different-dimensional regimes highlights their robustness and adaptability in small-sample, high-dimensional contexts.

This analysis further emphasizes the importance of choosing suitable covariance estimation techniques for PCA when working with data where $n < p$. Traditional methods like MLE and LW exhibit inherent limitations under these conditions. Specifically, as the ratio p/n increases, the MLE tends to suffer from overdispersion in the first PC, while the LW estimator tends to be overly conservative, leading to underdispersion. Both extremes can distort the representation of the true data structure and yield misleading PCA results. These issues become particularly pronounced when the sample size is very limited relative to the number of variables, making it necessary to employ methods that are better calibrated for such settings.

Among the proposed estimators, the SPDC method stands out by consistently achieving the lowest CSE for the first PC while also maintaining a high percentage of explained variance. Its performance advantage becomes increasingly evident as dimensionality grows, and its ability to strike a balance between bias and variance reduction makes it especially well-suited for high-dimensional data analysis. The regularization strategy employed in SPDC effectively controls the overdispersion seen in MLE and mitigates the underfitting tendency of LW. Consequently, SPDC emerges as a reliable and practical choice for accurate and stable PCA, particularly in small-sample regimes where conventional estimators often fail to perform adequately.

6.2.2 Experiments on Public Data

In this analysis, we utilize the public dataset discussed in Chapter 5.4. This study examines the response of specific genes to various experimental treatments, including different plant tissues such as shoots, leaves, and stems, as well as various treatments like light/dark cycles, cold treatments, and different nutrient supplies [73]. To illustrate the impact of sample size n and number of dimensions p on PCA, we used the entire dataset to represent a large p. For representing a smaller p, we selected only the experimental conditions related to leaves. This approach provides a clear comparison of how varying n and p affect the outcomes of PCA.

Table 6.3: Effect of n and p — Leaf Data - Low p

p	n	Average Overdispersion of first PC							
		Pop	**MLE**	**LW**	**PDC**	**SPDC**	**LSPDC**	**MAXPDC**	**RPDC**
7	3	0.000000	0.016584	0.034265	0.016584	0.001904	0.000928	0.000104	0.003039
7	4	0.000000	0.012940	0.155313	0.012940	0.002681	0.001576	0.002506	0.003238
7	5	0.000000	0.000429	0.139237	0.000429	7.63E-06	0.000146	0.001520	5.33E-05
7	6	0.000000	0.006356	0.218080	0.006356	0.003393	0.004016	0.003317	0.002931
p	n	Average Percentage of Explained Variance of first PC							
		Pop	**MLE**	**LW**	**PDC**	**SPDC**	**LSPDC**	**MAXPDC**	**RPDC**
7	3	92.53208	99.41551	82.63759	99.41551	94.86422	94.16039	93.07688	95.47856
7	4	92.53208	99.97908	66.73236	99.97908	95.92193	95.13113	95.80943	96.25740
7	5	92.53208	94.09853	64.32494	94.09853	92.74088	91.61982	95.47878	93.08372
7	6	92.53208	99.26977	53.06420	99.26977	97.45536	97.88784	97.39975	97.10733
p	n	Average CSE of first PC							
		Pop	**MLE**	**LW**	**PDC**	**SPDC**	**LSPDC**	**MAXPDC**	**RPDC**
7	3	0.000000	0.052356	0.052356	0.052356	0.008604	0.008443	0.052356	0.009712
7	4	0.000000	0.038296	0.038296	0.038296	0.007327	0.007441	0.038296	0.008036
7	5	0.000000	0.167629	0.167629	0.167629	0.008450	0.009326	0.167629	0.011033
7	6	0.000000	0.052162	0.052162	0.052162	0.006988	0.009557	0.052162	0.006118

Note: Best Second Best Third Best

Note: All values are averaged over 100 replications for each n and p setting.

Table 6.3 shows that when performing PCA in HD settings with $n < p$ and a much smaller p, all the regularized PDCs (except MAXPDC) yield close results in both the direction and magnitude of the sample PCs. It is challenging to discern a clear pattern indicating which method is superior. However, all these regularized PDCs outperform the existing methods, even in HD settings where $n < p$ is significantly smaller.

Table 6.4 illustrates the behavior of different estimation methods when performing PCA in HD settings with significantly larger p and $n < p$. While this table partially overlaps with the synthetic scenarios previously discussed in Tables 5.6 to 5.8, it introduces a new and focused analysis on the *average overdispersion* of the first PC, which was not addressed earlier. This addition provides a deeper understanding of how each method handles eigenvalue inflation under increasing p/n ratios.

Table 6.4: Effect of n and p — Phytozome Data - High p

p	n	Average Overdispersion of first PC							
		Pop	**MLE**	**LW**	**PDC**	**SPDC**	**LSPDC**	**MAXPDC**	**RPDC**
74	5	0.000000	1.095071	0.726472	1.095071	0.407725	0.424260	0.620553	0.408652
74	6	0.000000	0.862375	1.143960	0.862375	0.265933	0.272175	0.392784	0.270337
74	7	0.000000	0.707309	1.389852	0.707309	0.240719	0.254671	0.350673	0.237093
74	8	0.000000	0.575631	1.497910	0.575631	0.201135	0.207158	0.282708	0.192903
74	9	0.000000	0.463153	1.576263	0.463153	0.162813	0.162324	0.211596	0.157926
74	10	0.000000	0.469671	1.619404	0.469671	0.145126	0.150952	0.185968	0.136837
p	n	**Average Percentage of Explained Variance of first PC**							
		Pop	**MLE**	**LW**	**PDC**	**SPDC**	**LSPDC**	**MAXPDC**	**RPDC**
74	5	74.94184	99.27146	55.12550	99.27146	89.78744	90.08548	93.25670	89.80431
74	6	74.94184	99.08073	47.13994	99.08073	88.34649	88.50289	91.23277	88.45704
74	7	74.94184	98.88958	41.37240	98.88958	88.91245	89.31161	91.80391	88.80681
74	8	74.94184	98.27670	37.29955	98.27670	88.73541	88.94041	91.29501	88.45020
74	9	74.94184	97.31829	33.66149	97.31829	88.20886	88.18895	90.06640	88.00823
74	10	74.94184	98.84208	30.56228	98.84208	88.22734	88.49138	89.98104	87.84238
p	n	**Average CSE of first PC**							
		Pop	**MLE**	**LW**	**PDC**	**SPDC**	**LSPDC**	**MAXPDC**	**RPDC**
74	5	0.000000	0.247691	0.247691	0.247691	0.076225	0.079755	0.247691	0.076708
74	6	0.000000	0.248536	0.248536	0.248536	0.073888	0.093020	0.248536	0.075063
74	7	0.000000	0.203118	0.203118	0.203118	0.053610	0.073243	0.203118	0.054907
74	8	0.000000	0.223614	0.223614	0.223614	0.052163	0.056960	0.223614	0.052487
74	9	0.000000	0.198741	0.198741	0.198741	0.051376	0.054183	0.198741	0.053768
74	10	0.000000	0.167903	0.167903	0.167903	0.050993	0.054053	0.167903	0.052861

Note: Best Second Best Third Best

Note: All values are averaged over 100 replications for each n and p setting.

According to the results presented, all the regularized PDC estimators (except MAXPDC) perform well in capturing both the direction and magnitude of the sample PCs. However, as highlighted consistently across simulations, the SPDC method performs significantly better than all other methods—particularly as the sample size n decreases and dimensionality p increases. These findings reinforce that SPDC is comparatively more effective in maintaining accurate and reliable PC estimates under the challenging conditions of high p/n ratios in $n < p$ HD data settings.

6.3 Degree of Correlation

In PCA, highly correlated data is conducive to its purpose, providing rich information for extracting meaningful PCs. On the other hand, low or weakly correlated data may be challenging for PCA in HD data settings. Data's low or weak correlation may lead to inaccurate and unreliable PC extraction. To emphasize this issue further, let's consider five distinct datasets characterized by different levels of correlation structures, as shown in Figure 6.1.

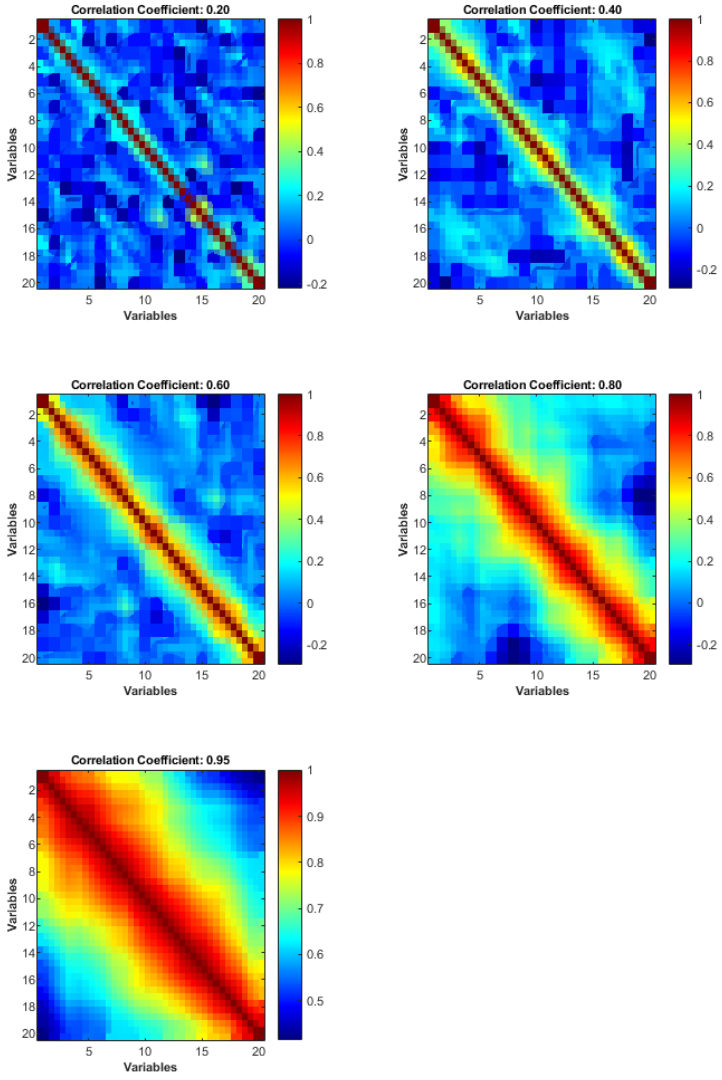


Figure 6.1: Five different degree levels of correlation

6.3.1 Experiments on Synthetic Data

In this simulation setting, we considered five different degrees of correlation structures of data to examine the impact of the degree of correlation on PCA when $n < p$ data settings. The degrees of correlation we examined are as follows:

1. $\boldsymbol{\rho}$ **= 0.0-0.2:** Very Low correlated data
2. $\boldsymbol{\rho}$ **= 0.2-0.4:** Low correlated data
3. $\boldsymbol{\rho}$ **= 0.4-0.6:** Medium correlated data
4. $\boldsymbol{\rho}$ **= 0.6-0.8:** High correlated data
5. $\boldsymbol{\rho}$ **= 0.8-1.0:** Very High correlated data

For this study, we generated a series of independent and identically distributed random vectors $\boldsymbol{X}_i \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$, where the dimensionality was fixed at $p = 20$, and the sample size n varied from 3 to 11. The population covariance matrix $\boldsymbol{\Sigma}$ was constructed using random multivariate structures to reflect a dense and non-sparse setting. This configuration was used to explore the behavior of different estimators under varying p/n conditions. Each scenario was replicated 500 times to ensure the reliability of the performance assessments.

The results of these synthetic experiments are presented in Tables 6.5 to 6.9. These tables illustrate the performance of various estimation methods under different correlation conditions. When the correlation of the data is lower, the LW method performs significantly better regarding the magnitude of the sample PCs. However, it yields higher CSEs than all other estimation methods. When prioritizing the direction of the sample PCs, all regularized PDC estimations (except MAXPDC) deliver superior results compared to LW and MLE in HD settings where $n < p$. In scenarios with lower correlation, it is challenging to determine which method is superior due to the need for a clear pattern.

As the correlation of the data increases, the LW method fails to provide significant improvements in both the direction and magnitude of the sample PC estimates. Conversely, all regularized PDC estimations (except MAXPDC) perform much better regarding the magnitude and direction of the sample PCs. However, SPDC consistently yields better improvements by minimizing the overdispersion issues of the first PC over the different $n < p$ data settings, followed by LSPDC and RPDC, respectively. This indicates that a high correlation in the data is crucial for effective PCA performance. In HD settings where $n < p$, SPDC outperforms other methods by mitigating the overdispersion of the first PC. Consequently, SPDC, LSPDC, and RPDC can significantly improve finite sample estimates of PCA in HD scenarios where $n < p$.

Table 6.5: CSE and Overdispersion vs. Correlation Level — $0.00 \leq \rho \leq 0.20$

p	n	Average Overdispersion first PC							
		Pop	**MLE**	**LW**	**PDC**	**SPDC**	**LSPDC**	**MAXPDC**	**RPDC**
20	3	0.000000	2.566435	0.956667	2.566435	2.160719	2.122423	2.515112	2.157668
20	4	0.000000	0.905763	0.138356	0.905763	0.696527	0.691933	0.885255	0.702955
20	5	0.000000	0.423171	0.023838	0.423171	0.312245	0.312376	0.413985	0.315869
20	6	0.000000	0.234895	0.003389	0.234895	0.170560	0.170655	0.231192	0.175792
20	7	0.000000	0.138255	0.000011	0.138255	0.099418	0.099951	0.134020	0.101938
20	8	0.000000	0.093684	0.000338	0.093684	0.064677	0.064911	0.091688	0.066979
20	9	0.000000	0.061873	0.001549	0.061873	0.042117	0.042034	0.060433	0.044273
20	10	0.000000	0.044213	0.002427	0.044213	0.029612	0.029687	0.042998	0.031272
20	11	0.000000	0.034513	0.002481	0.034513	0.022746	0.022903	0.033390	0.023762
p	n	Average Percentage of Explained Variance of first PC							
		Pop	**MLE**	**LW**	**PDC**	**SPDC**	**LSPDC**	**MAXPDC**	**RPDC**
20	3	14.12909	64.78909	45.05912	64.78909	60.61263	60.19885	64.27998	60.57979
20	4	14.14239	51.00219	28.54845	51.00219	46.46561	46.35884	50.58251	46.61442
20	5	14.42406	43.51600	21.32886	43.51600	39.41388	39.41909	43.19851	39.55847
20	6	14.32740	38.56037	17.23823	38.56037	34.97685	34.98263	38.36860	35.29118
20	7	14.28997	34.65577	14.46789	34.65577	31.56001	31.60625	34.34142	31.77749
20	8	14.57533	32.68322	13.48842	32.68322	29.62090	29.64809	32.48922	29.88637
20	9	14.61777	30.34967	12.12877	30.34967	27.59734	27.58444	30.16556	27.92529
20	10	14.61058	28.71578	11.30602	28.71578	26.15424	26.16881	28.52061	26.47328
20	11	14.39605	27.53241	10.87366	27.53241	25.06052	25.09725	27.31689	25.29606
p	n	Average CSE of first PC							
		Pop	**MLE**	**LW**	**PDC**	**SPDC**	**LSPDC**	**MAXPDC**	**RPDC**
20	3	0.000000	0.846842	0.931456	0.846842	0.850376	0.862036	0.846842	0.851737
20	4	0.000000	0.816517	0.950872	0.816517	0.842731	0.841838	0.816517	0.845108
20	5	0.000000	0.788274	0.889265	0.788274	0.786813	0.780506	0.788274	0.785505
20	6	0.000000	0.800827	0.847733	0.800827	0.800055	0.812099	0.800827	0.808584
20	7	0.000000	0.799125	0.804332	0.799125	0.785264	0.786136	0.799125	0.791624
20	8	0.000000	0.784976	0.789092	0.784976	0.783646	0.795751	0.784976	0.788722
20	9	0.000000	0.710462	0.722379	0.710462	0.693355	0.690360	0.710462	0.703303
20	10	0.000000	0.718333	0.725625	0.718333	0.716869	0.719475	0.718333	0.721407
20	11	0.000000	0.744723	0.746193	0.744723	0.730779	0.732166	0.744723	0.744867

Note: Best Second Best Third Best

Note: All values are averaged over 500 replications for each n and p setting.

Table 6.6: CSE and Overdispersion vs. Correlation Level — $0.20 \leq \rho \leq 0.40$

p	n	Average Overdispersion first PC							
		Pop	**MLE**	**LW**	**PDC**	**SPDC**	**LSPDC**	**MAXPDC**	**RPDC**
20	3	0.000000	1.358510	0.288601	1.358510	0.916612	0.890482	1.261940	0.917231
20	4	0.000000	0.429835	0.004192	0.429835	0.221760	0.216812	0.372483	0.226024
20	5	0.000000	0.201522	0.003904	0.201522	0.086797	0.086582	0.171709	0.087965
20	6	0.000000	0.096558	0.017543	0.096558	0.034906	0.034632	0.082068	0.036843
20	7	0.000000	0.051941	0.026553	0.051941	0.016438	0.016634	0.039169	0.017081
20	8	0.000000	0.035681	0.023302	0.035681	0.010802	0.010758	0.028049	0.011746
20	9	0.000000	0.022061	0.023950	0.022061	0.006004	0.005902	0.016689	0.006405
20	10	0.000000	0.015928	0.021487	0.015928	0.003975	0.003876	0.012102	0.004388
20	11	0.000000	0.010767	0.021255	0.010767	0.002128	0.002129	0.007954	0.002352
p	n	Average Percentage of Explained Variance of first PC							
		Pop	**MLE**	**LW**	**PDC**	**SPDC**	**LSPDC**	**MAXPDC**	**RPDC**
20	3	33.05074	69.90871	50.03898	69.90871	63.32633	62.89169	68.57454	63.33656
20	4	33.06318	58.45515	35.57076	58.45515	51.30161	51.09699	56.70052	51.47611
20	5	33.31068	53.38661	30.51639	53.38661	46.48616	46.46985	51.84220	46.57456
20	6	33.24976	48.78666	26.62718	48.78666	42.59137	42.55457	47.57354	42.84697
20	7	33.19936	45.68232	24.27424	45.68232	40.22174	40.26349	44.03937	40.35784
20	8	33.45244	44.62765	24.42158	44.62765	39.60131	39.58870	43.36067	39.86435
20	9	33.52804	42.92179	23.74035	42.92179	38.42846	38.38675	41.69849	38.58956
20	10	33.50237	41.96856	23.66930	41.96856	37.73156	37.67897	40.88210	37.94590
20	11	33.30753	40.64481	22.99844	40.64481	36.56931	36.56999	39.61375	36.73686
p	n	Average CSE of first PC							
		Pop	**MLE**	**LW**	**PDC**	**SPDC**	**LSPDC**	**MAXPDC**	**RPDC**
20	3	0.000000	0.639275	0.776357	0.639275	0.638786	0.649205	0.639275	0.638057
20	4	0.000000	0.566501	0.811322	0.566501	0.556459	0.559221	0.566501	0.556029
20	5	0.000000	0.484789	0.623521	0.484789	0.445296	0.448837	0.484789	0.450780
20	6	0.000000	0.458885	0.533603	0.458885	0.423774	0.435292	0.458885	0.430758
20	7	0.000000	0.396890	0.425720	0.396890	0.377343	0.377274	0.396890	0.383597
20	8	0.000000	0.333365	0.326759	0.333365	0.308570	0.316687	0.333365	0.317226
20	9	0.000000	0.275743	0.284177	0.275743	0.239587	0.242512	0.275743	0.237221
20	10	0.000000	0.237333	0.237333	0.237333	0.206202	0.201031	0.237333	0.209726
20	11	0.000000	0.247988	0.247988	0.247988	0.210436	0.219935	0.247988	0.215511

Note: Best Second Best Third Best

Note: All values are averaged over 500 replications for each n and p setting.

Table 6.7: CSE and Overdispersion vs. Correlation Level — $0.40 \leq \rho \leq 0.60$

p	n	Average Overdispersion first PC							
		Pop	**MLE**	**LW**	**PDC**	**SPDC**	**LSPDC**	**MAXPDC**	**RPDC**
20	3	0.000000	0.628373	0.030198	0.628373	0.258108	0.243040	0.503719	0.260593
20	4	0.000000	0.189842	0.021748	0.189842	0.036252	0.034324	0.121033	0.036879
20	5	0.000000	0.090062	0.040886	0.090062	0.011119	0.010816	0.052434	0.011377
20	6	0.000000	0.040037	0.055858	0.040037	0.001840	0.001748	0.022913	0.002077
20	7	0.000000	0.022113	0.056194	0.022113	0.000492	0.000517	0.009634	0.000550
20	8	0.000000	0.015387	0.041707	0.015387	0.000529	0.000489	0.007149	0.000676
20	9	0.000000	0.009236	0.037640	0.009236	4.68E-05	3.42E-05	0.003543	8.80E-05
20	10	0.000000	0.006753	0.030997	0.006753	3.34E-05	2.36E-05	0.002490	4.89E-05
20	11	0.000000	0.004125	0.028507	0.004125	6.09E-06	9.86E-06	0.001399	3.58E-07
p	n	Average Percentage of Explained Variance of first PC							
		Pop	**MLE**	**LW**	**PDC**	**SPDC**	**LSPDC**	**MAXPDC**	**RPDC**
20	3	52.03664	77.10401	57.53187	77.10401	68.10239	67.62638	74.48033	68.17954
20	4	52.04793	68.92284	46.33639	68.92284	59.42206	59.22333	65.52195	59.48561
20	5	52.28838	65.70943	43.24563	65.70943	57.00409	56.93932	62.52887	57.05859
20	6	52.23381	62.23848	40.41663	62.23848	54.37879	54.32420	59.80232	54.51267
20	7	52.18254	60.32734	39.19867	60.32734	53.39705	53.42821	57.55869	53.46709
20	8	52.42913	59.76778	40.34709	59.76778	53.78967	53.73776	57.43130	53.96711
20	9	52.51007	58.58833	40.23977	58.58833	52.94260	52.87984	56.27445	53.10334
20	10	52.48079	57.99319	40.67038	57.99319	52.86820	52.80688	55.82811	52.94978
20	11	52.29019	56.83158	40.35145	56.83158	52.11566	52.06815	54.93480	52.24788
p	n	Average CSE of first PC							
		Pop	**MLE**	**LW**	**PDC**	**SPDC**	**LSPDC**	**MAXPDC**	**RPDC**
20	3	0.000000	0.472846	0.641248	0.472846	0.447239	0.448332	0.472846	0.444812
20	4	0.000000	0.340695	0.539239	0.340695	0.322058	0.324054	0.340695	0.316361
20	5	0.000000	0.281376	0.374135	0.281376	0.231497	0.242959	0.281376	0.235258
20	6	0.000000	0.218169	0.274328	0.218169	0.195144	0.203826	0.218169	0.188770
20	7	0.000000	0.165709	0.185960	0.165709	0.126353	0.126965	0.165709	0.134726
20	8	0.000000	0.124739	0.124739	0.124739	0.092661	0.093860	0.124739	0.095878
20	9	0.000000	0.101404	0.101404	0.101404	0.067426	0.070476	0.101404	0.076694
20	10	0.000000	0.088707	0.088707	0.088707	0.057628	0.061324	0.088707	0.060446
20	11	0.000000	0.083385	0.083385	0.083385	0.056841	0.060780	0.083385	0.062080

Note: Best Second Best Third Best

Note: All values are averaged over 500 replications for each n and p setting.

Table 6.8: CSE and Overdispersion vs. Correlation Level — $0.60 \leq \rho \leq 0.80$

p	n	Average Overdispersion first PC							
		Pop	**MLE**	**LW**	**PDC**	**SPDC**	**LSPDC**	**MAXPDC**	**RPDC**
20	3	0.000000	0.215277	0.017023	0.215277	0.023566	0.019277	0.122858	0.023791
20	4	0.000000	0.065905	0.090988	0.065905	1.43E-06	4.49E-05	0.021555	5.80E-09
20	5	0.000000	0.030816	0.094321	0.030816	0.000560	0.000699	0.006915	0.000382
20	6	0.000000	0.012021	0.091328	0.012021	0.001885	0.002062	0.002016	0.001814
20	7	0.000000	0.007563	0.073883	0.007563	0.000933	0.001022	0.000767	0.000897
20	8	0.000000	0.005327	0.052091	0.005327	0.000347	0.000433	0.000765	0.000270
20	9	0.000000	0.002888	0.046518	0.002888	0.000793	0.000829	0.000133	0.000733
20	10	0.000000	0.002163	0.037181	0.002163	0.000497	0.000522	0.000135	0.000470
20	11	0.000000	0.001194	0.031110	0.001194	0.000537	0.000579	2.35E-05	0.000483
p	n	Average Percentage of Explained Variance of first PC							
		Pop	**MLE**	**LW**	**PDC**	**SPDC**	**LSPDC**	**MAXPDC**	**RPDC**
20	3	70.97502	85.64734	66.84909	85.64734	75.82949	75.36553	82.05914	75.85266
20	4	70.98962	80.93235	59.30705	80.93235	70.94330	70.73007	76.67583	70.98667
20	5	71.22888	79.07953	57.49418	79.07953	70.17085	70.04638	74.94787	70.35483
20	6	71.14191	76.62397	56.03162	76.62397	68.97120	68.87135	73.38687	69.01241
20	7	71.13514	75.89829	56.24730	75.89829	69.46203	69.38389	72.65239	69.49501
20	8	71.36107	75.67909	57.85848	75.67909	70.25934	70.13024	72.99706	70.38833
20	9	71.43417	74.83326	57.79332	74.83326	69.65366	69.61347	72.16487	69.72209
20	10	71.42092	74.54043	58.48596	74.54043	69.92507	69.88856	72.19999	69.96640
20	11	71.22658	73.66973	58.75454	73.66973	69.58782	69.52549	71.56954	69.67176
p	n	Average CSE of first PC							
		Pop	**MLE**	**LW**	**PDC**	**SPDC**	**LSPDC**	**MAXPDC**	**RPDC**
20	3	0.000000	0.290355	0.422020	0.290355	0.264594	0.265171	0.290355	0.269689
20	4	0.000000	0.189366	0.301317	0.189366	0.153023	0.151296	0.189366	0.148534
20	5	0.000000	0.132343	0.163346	0.132343	0.088663	0.091542	0.132343	0.092801
20	6	0.000000	0.098084	0.098083	0.098084	0.052155	0.054086	0.098084	0.063463
20	7	0.000000	0.064996	0.064996	0.064996	0.039816	0.037749	0.064996	0.039855
20	8	0.000000	0.039866	0.039866	0.039866	0.013380	0.014278	0.039866	0.020585
20	9	0.000000	0.035853	0.035853	0.035853	0.012344	0.013241	0.035853	0.016604
20	10	0.000000	0.031065	0.031065	0.031065	0.009679	0.010282	0.031065	0.014298
20	11	0.000000	0.026923	0.026923	0.026923	0.007722	0.008182	0.026923	0.012418

Note: Best Second Best Third Best

Note: All values are averaged over 500 replications for each n and p setting.

Table 6.9: CSE and Overdispersion vs. Correlation Level — $0.80 \leq \rho \leq 1.00$

p	n	Average Overdispersion first PC							
		Pop	**MLE**	**LW**	**PDC**	**SPDC**	**LSPDC**	**MAXPDC**	**RPDC**
20	3	0.000000	0.027316	0.151315	0.027316	0.004424	0.006081	0.003904	0.004221
20	4	0.000000	0.008359	0.187620	0.008359	0.005007	0.005471	3.97E-05	0.004794
20	5	0.000000	0.003441	0.166069	0.003441	0.003846	0.004018	5.77E-05	0.003706
20	6	0.000000	0.001300	0.120713	0.001300	0.002591	0.002792	0.000210	0.002548
20	7	0.000000	0.001014	0.087746	0.001014	0.000803	0.000946	4.03E-05	0.000827
20	8	0.000000	0.000525	0.063552	0.000525	0.000479	0.000564	3.75E-05	0.000409
20	9	0.000000	0.000271	0.057059	0.000271	0.000671	0.000666	0.000130	0.000665
20	10	0.000000	0.000215	0.043494	0.000215	0.000346	0.000362	6.33E-05	0.000349
20	11	0.000000	0.000123	0.034115	0.000123	0.000226	0.000250	5.66E-05	0.000208
p	n	Average Percentage of Explained Variance of first PC							
		Pop	**MLE**	**LW**	**PDC**	**SPDC**	**LSPDC**	**MAXPDC**	**RPDC**
20	3	89.55671	94.78319	77.25569	94.78319	87.45340	87.09084	91.53250	87.50224
20	4	89.54451	93.08556	72.76862	93.08556	86.80397	86.67987	89.78864	86.86299
20	5	89.79287	92.41618	71.56824	92.41618	87.01951	86.95805	89.45324	87.07034
20	6	89.74374	91.54622	72.37189	91.54622	87.19875	87.10199	89.01997	87.21990
20	7	89.70485	91.44867	73.48021	91.44867	88.15262	88.01983	89.35701	88.12940
20	8	89.96142	91.31662	75.04723	91.31662	88.66662	88.55606	89.59899	88.76512
20	9	90.00134	91.04188	74.89389	91.04188	88.36361	88.36948	89.27935	88.37039
20	10	89.97062	90.95488	75.98059	90.95488	88.72257	88.69482	89.43670	88.71821
20	11	89.75809	90.54203	76.69774	90.54203	88.69619	88.64113	89.22603	88.73802
p	n	Average CSE of first PC							
		Pop	**MLE**	**LW**	**PDC**	**SPDC**	**LSPDC**	**MAXPDC**	**RPDC**
20	3	0.000000	0.129615	0.168363	0.129615	0.095867	0.097439	0.129615	0.096827
20	4	0.000000	0.056047	0.075326	0.056047	0.035799	0.032896	0.056047	0.037633
20	5	0.000000	0.035932	0.035932	0.035932	0.008751	0.012900	0.035932	0.014255
20	6	0.000000	0.022223	0.025401	0.022223	0.005617	0.006445	0.022223	0.011822
20	7	0.000000	0.020529	0.020529	0.020529	0.008659	0.009270	0.020529	0.011272
20	8	0.000000	0.010702	0.010702	0.010702	0.001116	0.001567	0.010702	0.003969
20	9	0.000000	0.010379	0.010379	0.010379	0.001049	0.001448	0.010379	0.003831
20	10	0.000000	0.008781	0.008781	0.008781	0.000829	0.001086	0.008781	0.003625
20	11	0.000000	0.007205	0.007205	0.007205	0.000574	0.000750	0.007205	0.003439

Note: Best Second Best Third Best

Note: All values are averaged over 500 replications for each n and p setting.

6.3.2 Experiments on Public Data

Low Correlated Data: The selection and classification of genes are crucial for identifying genes related to specific diseases. Researchers developed a novel application called HDG-select, which uses a GUI to perform gene selection and classification in HD datasets. One of these datasets, the Leukemia dataset, contains 3571 genes and 72 samples with an average correlation of 0.13, indicating a very low correlation [74]. We conducted a PCA on this dataset to understand the behavior of our estimation methods on low-correlated data.

Table 6.10: Effect of Degree of Correlation — Leukemia Data - Low ρ

		Average Overdispersion first PC							
p	**n**	**Pop**	**MLE**	**LW**	**PDC**	**SPDC**	**LSPDC**	**MAXPDC**	**RPDC**
3571	3	0.000000	305.472114	96.312111	305.472114	318.312117	309.377264	306.030477	323.148453
3571	4	0.000000	99.157348	8.566014	99.157348	97.622417	99.094350	98.416700	96.871012
3571	5	0.000000	55.885562	1.955417	55.885562	53.388173	49.819410	56.489079	56.302445
3571	6	0.000000	16.963451	0.522791	16.963451	19.040216	19.182229	17.077342	18.971949
3571	7	0.000000	12.612973	0.764986	12.612973	13.851280	13.992253	12.587353	13.630250
		Average Percentage of Explained Variance of first PC							
p	**n**	**Pop**	**MLE**	**LW**	**PDC**	**SPDC**	**LSPDC**	**MAXPDC**	**RPDC**
3571	3	13.58619	54.94861	36.81145	54.94861	55.80896	55.21216	54.98639	56.12851
3571	4	13.58619	42.44831	22.06930	42.44831	42.22405	42.43914	42.34032	42.11362
3571	5	13.58619	38.60606	18.26629	38.60606	38.04063	37.20916	38.74079	38.69920
3571	6	13.58619	28.99778	10.88065	28.99778	29.91393	29.97471	29.04943	29.88464
3571	7	13.58619	28.14378	10.00104	28.14378	28.84167	28.91910	28.12899	28.71946
		Average CSE of first PC							
p	**n**	**Pop**	**MLE**	**LW**	**PDC**	**SPDC**	**LSPDC**	**MAXPDC**	**RPDC**
3571	3	0.000000	0.939506	0.939506	0.939506	0.941052	0.939045	0.939506	0.942047
3571	4	0.000000	0.777799	0.777799	0.777799	0.816732	0.817318	0.777799	0.820068
3571	5	0.000000	1.004135	1.004135	1.004135	0.967828	0.969851	1.004135	0.966038
3571	6	0.000000	0.523821	0.523821	0.523821	0.956192	0.960065	0.523821	0.956302
3571	7	0.000000	0.346688	0.346688	0.346688	0.896086	0.904911	0.346688	0.889453

Note: Best Second Best Third Best

Note: All values are averaged over 100 replications for each n and p setting.

As shown in Table 6.10, the LW provides better sample PCs in magnitude when the correlation is low. However, the direction of the sample PCs needs to be improved. In contrast, regularized PDC estimations can enhance both the direction and magnitude. Nonetheless, the absence of a significant pattern makes it difficult to determine which technique performs best in both the magnitude and direction of the PC estimates.

Medium Correlated Data: The crop recommendation dataset is created for precision agriculture, helping farmers make crop selection decisions based on various soil and climate parameters [75]. It combines data on rainfall, climate, and fertilizers available for India. It includes 2200 observations and seven dimensions: Nitrogen, phosphorus, and Potassium content in the soil, temperature, humidity, pH value, and rainfall. With an average correlation of 0.41, indicating medium correlation, this dataset is used to investigate the behavior of medium-correlated data in PCA.

Table 6.11: Effect of Degree of Correlation — Crop Data - Medium ρ

p	n	Average Overdispersion first PC							
		Pop	**MLE**	**LW**	**PDC**	**SPDC**	**LSPDC**	**MAXPDC**	**RPDC**
7	3	0.000000	0.716375	0.241359	0.716375	0.360123	0.349300	0.651880	0.365368
7	4	0.000000	0.278081	0.023641	0.278081	0.062437	0.061684	0.250544	0.068066
7	5	0.000000	0.158522	0.004849	0.158522	0.019435	0.019658	0.126068	0.022265
7	6	0.000000	0.086856	0.000285	0.086856	0.004721	0.005354	0.077470	0.007268
p	n	Average Percentage of Explained Variance of first PC							
		Pop	**MLE**	**LW**	**PDC**	**SPDC**	**LSPDC**	**MAXPDC**	**RPDC**
7	3	40.12914	85.37055	66.38934	85.37055	72.20598	71.72026	83.28600	72.43873
7	4	40.12914	74.65127	50.19492	74.65127	56.48723	56.38836	72.89741	57.20874
7	5	40.12914	70.22633	45.39276	70.22633	50.66737	50.72779	66.96924	51.40868
7	6	40.12914	65.03698	38.70131	65.03698	45.93588	46.31308	63.65273	47.33438
p	n	Average CSE of first PC							
		Pop	**MLE**	**LW**	**PDC**	**SPDC**	**LSPDC**	**MAXPDC**	**RPDC**
7	3	0.000000	1.193580	1.092635	1.193580	1.074358	1.077965	1.193580	1.059404
7	4	0.000000	1.300335	1.251043	1.300335	1.141337	1.117454	1.300335	1.138201
7	5	0.000000	1.241150	1.246108	1.241150	1.085658	1.061148	1.241150	1.090911
7	6	0.000000	1.271500	1.231445	1.271500	1.108678	1.174234	1.271500	1.087274

Note: Best Second Best Third Best

Note: All values are averaged over 100 replications for each n and p setting.

As highlighted earlier and shown in Table 6.11, the LW estimation method remains effective in providing superior sample PCs in terms of magnitude when the correlation is medium. However, it fails to improve the direction of the sample PCs. In contrast, compared to low-correlated data, regularized PDC estimations more consistently reduce the CSE of the first PC. This suggests regularized PDC estimations become increasingly effective as the correlation increases and outperforms other methods.

High Correlated Data: Here, we use the public dataset that we discussed in Chapter 5.4. With an average correlation of 0.76, this dataset is utilized to evaluate the performance of our estimation methods on highly correlated data in PCA, particularly in $n < p$ HD data settings.

Consequently, as shown in Table 6.4, which summarizes results for the highly correlated Phytozome gene expression data, the SPDC method demonstrates significantly improved performance in PCA when $n < p$. Specifically, SPDC consistently achieves the lowest overdispersion of the first PC, the most accurate percentage of explained variance, and the smallest cosine similarity error (CSE), outperforming traditional methods such as MLE and LW as well as other PDC-based estimators.

Across all sample sizes $n = 5$ to 10, SPDC maintains robust and stable behavior, with minimal overestimation of explained variance and consistently accurate estimation of PC directions. In contrast, MLE tends to inflate the importance of the first PC due to eigenvalue overdispersion, while LW often underestimates it in small samples. Although LSPDC and RPDC also perform competitively, SPDC emerges as the most balanced and reliable method across all evaluation metrics.

These findings underscore the importance of selecting a covariance estimator that responds effectively to data structure. In highly correlated, HD datasets, as exemplified by the Phytozome results in Table 6.4—regularized PDC estimators, particularly SPDC, offer clear advantages for accurate and stable PCA outcomes in $n < p$ settings.

6.4 Types of Estimations of Covariance

The covariance matrix plays a central role in PCA, as it determines the directions (PCs) along which the variance in the data is maximized. Consequently, the choice and structure of the covariance matrix can significantly influence the outcome and interpretability of PCA results, especially in HD, low-sample-size (HDLSS) settings. This section introduces several distinct types of covariance structures commonly encountered in real-world and simulated data. These include unstructured random covariance matrices, Toeplitz-type matrices that model linear decay of correlations, tridiagonal structures representing local dependencies, block-diagonal matrices that capture grouped variable relationships, and heavy-tailed or contaminated structures that reflect the presence of outliers. Each of these covariance settings is used to evaluate how traditional and proposed covariance estimation methods perform in guiding PCA under various practical and theoretical challenges. The following subsections will examine each type in detail.

6.4.1 Experiments on Synthetic Data

Random/Dense Covariance Structure

The variance-covariance matrix, $\mathbf{\Sigma}$, employed to generate the data for the simulations, followed specific underlying structures.

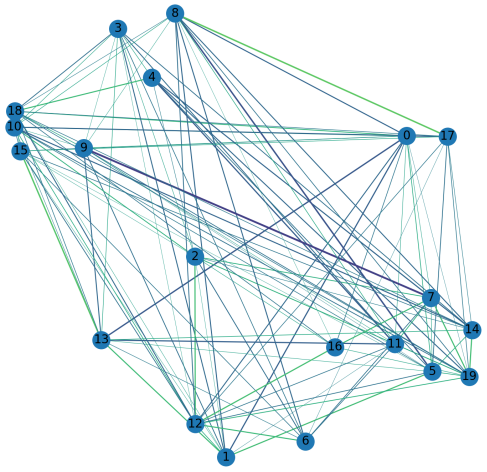


Figure 6.2: Random/Dense Covariance Structure

Figure 6.2 illustrates the structure of a random 20×20 covariance matrix generated using $\mathbf{\Sigma} = \mathtt{t}\mathtt{t}^\top$, where $\mathtt{t}$ is a matrix of random values drawn from the uniform distribution on the interval (0, 1). This construction yields a symmetric, dense, and positive definite covariance matrix without any inherent block structure or sparsity. Each node in the graph represents a variable, and the edges reflect the magnitude of pairwise covariances. Due to the unstructured nature of $\mathbf{\Sigma}$, the corresponding PCs do not capture clear or interpretable directions of variance. This example demonstrates how PCA behaves in the presence of random, unstructured covariance, often resulting in diffuse and non-informative principal directions.

Despite this, as demonstrated in Table 6.12, the SPDC method significantly minimizes the overdispersion of the first PC, followed closely by LSPDC and RPDC. The results indicate that SPDC can identify the closest directions (PCs) by reducing the first PC's CSE, thereby maximizing the variance. Compared to the magnitudes of the sample PCs, all the regularized versions of PDC estimations provide very close percentages of explained variance for the first PC. However, considering the overall performance, including the CSE of the first PC, SPDC effectively addresses the issue of overdispersion in the first PC.

Table 6.12: Effect of Covariance Type — Random Covariance

p	n	Average Overdispersion of first PC							
		Pop	**MLE**	**LW**	**PDC**	**SPDC**	**LSPDC**	**MAXPDC**	**RPDC**
20	3	0.000000	0.149947	0.037591	0.149947	0.005540	0.003689	0.066781	0.005602
20	4	0.000000	0.040505	0.140737	0.040505	0.001819	0.001936	0.008969	0.001664
20	5	0.000000	0.017546	0.126138	0.017546	0.001741	0.001553	0.002676	0.001779
20	6	0.000000	0.008383	0.096425	0.008383	0.001780	0.002146	0.000693	0.001496
20	7	0.000000	0.004947	0.076989	0.004947	0.001133	0.001174	0.000430	0.000996
20	8	0.000000	0.004130	0.053602	0.004130	0.000349	0.000380	0.000406	0.000270
20	9	0.000000	0.002951	0.041354	0.002951	0.000232	0.000266	0.000357	0.000195
20	10	0.000000	0.001898	0.036120	0.001898	0.000217	0.000231	0.000118	0.000179
20	11	0.000000	0.001543	0.027566	0.001543	0.000261	0.000273	0.000093	0.000225
p	n	**Average Percentage of Explained Variance of first PC**							
		Pop	**MLE**	**LW**	**PDC**	**SPDC**	**LSPDC**	**MAXPDC**	**RPDC**
20	3	76.50551	88.75079	70.37435	88.75079	78.85917	78.42632	84.67747	78.87246
20	4	76.50551	84.30026	61.97607	84.30026	74.85354	74.80138	80.17347	74.92578
20	5	76.50551	82.42940	60.62233	82.42940	74.63940	74.74337	78.81875	74.61925
20	6	76.50551	81.08342	60.97936	81.08342	74.39629	74.18939	77.82131	74.57150
20	7	76.50551	80.35779	61.30794	80.35779	74.66194	74.62862	77.64125	74.77719
20	8	76.50551	80.30731	62.80860	80.30731	75.40000	75.35222	77.69829	75.53381
20	9	76.50551	79.94109	63.64406	79.94109	75.54245	75.47423	77.70081	75.62140
20	10	76.50551	79.42774	63.75641	79.42774	75.51785	75.48604	77.23406	75.60792
20	11	76.50551	79.28301	64.76541	79.28301	75.36271	75.33792	77.18822	75.44423
p	n	**Average CSE of first PC**							
		Pop	**MLE**	**LW**	**PDC**	**SPDC**	**LSPDC**	**MAXPDC**	**RPDC**
20	3	0.000000	0.215556	0.312436	0.215556	0.185583	0.197269	0.215556	0.191383
20	4	0.000000	0.133072	0.189546	0.133072	0.110655	0.108410	0.133072	0.109668
20	5	0.000000	0.083240	0.108600	0.083240	0.052334	0.054704	0.083240	0.051590
20	6	0.000000	0.065099	0.065913	0.065099	0.037954	0.039387	0.065099	0.038714
20	7	0.000000	0.048862	0.048862	0.048862	0.020258	0.021014	0.048862	0.023389
20	8	0.000000	0.033288	0.033288	0.033288	0.011785	0.012761	0.033288	0.015748
20	9	0.000000	0.029631	0.029631	0.029631	0.014306	0.015019	0.029631	0.018297
20	10	0.000000	0.021877	0.021877	0.021877	0.008999	0.009520	0.021877	0.013445
20	11	0.000000	0.020672	0.020672	0.020672	0.008625	0.008990	0.020672	0.013258

Note: Best Second Best Third Best

Note: All values are averaged over 500 replications for each n and p setting.

Toeplitz Covariance Structure

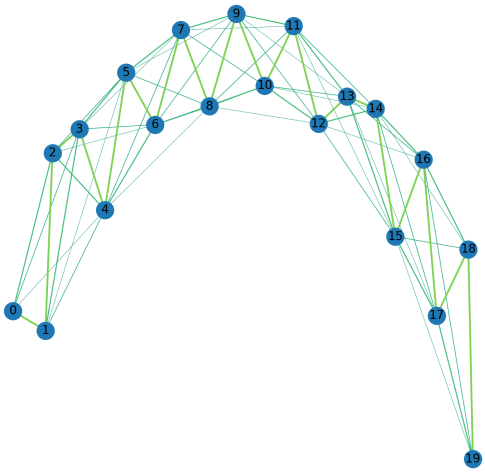


Figure 6.3: Toeplitz Covariance Structure

Figure 6.3 illustrates the structure of a 20×20 Toeplitz-type covariance matrix constructed using the formula

$$\sigma_{i,j} = \exp\left(-\frac{1}{\beta}|i-j|\right),$$

where $\beta = 2$ controls the rate of exponential decay. This structure implies that variables indexed closer together have higher covariance, and the dependence gradually weakens as the index gap increases. The resulting covariance matrix is symmetric and positive definite, capturing a banded, smooth structure commonly found in time-series or spatial data. In the graph, each node represents a variable and edge thickness reflects the magnitude of pairwise covariances. The strong local correlation creates visible neighborhood clusters, which provide meaningful directions of variance for PCA.

According to Table 6.13, LW estimations yield the closest sample estimates based on the percentage of explained variance of the first PC. However, LW does not improve the direction of the sample PCs by reducing the first PC's CSE. In contrast, except for MAXPDC, all other regularized PDC estimations significantly reduce the overdispersion of the first sample PC. Although it is challenging to determine the best among them due to the absence of a clear pattern, all regularized PDC estimations except MAXPDC perform better than MLE in PCA for $n < p$ HD settings with a Toeplitz covariance matrix.

Table 6.13: Effect of Covariance Type — Toeplitz Covariance

p	n	Average Overdispersion of first PC							
		Pop	**MLE**	**LW**	**PDC**	**SPDC**	**LSPDC**	**MAXPDC**	**RPDC**
20	3	0.000000	2.617302	0.978150	2.617302	1.912190	1.875670	2.514110	1.913632
20	4	0.000000	0.923251	0.136545	0.923251	0.619016	0.617926	0.881651	0.622215
20	5	0.000000	0.450000	0.027838	0.450000	0.287197	0.287786	0.423447	0.288674
20	6	0.000000	0.267692	0.006499	0.267692	0.167077	0.168157	0.254787	0.168952
20	7	0.000000	0.159084	0.000426	0.159084	0.095743	0.097335	0.152032	0.097649
20	8	0.000000	0.109325	0.000053	0.109325	0.065804	0.065327	0.105173	0.068937
20	9	0.000000	0.078734	0.000044	0.078734	0.043983	0.043963	0.075181	0.046081
20	10	0.000000	0.061627	0.000061	0.061627	0.035896	0.035744	0.058586	0.037734
20	11	0.000000	0.045885	0.000246	0.045885	0.025819	0.025787	0.043915	0.027400
p	n	Average Percentage of Explained Variance of first PC							
		Pop	**MLE**	**LW**	**PDC**	**SPDC**	**LSPDC**	**MAXPDC**	**RPDC**
20	3	19.13196	70.29153	50.40735	70.29153	62.86055	62.44096	69.27286	62.87703
20	4	19.13196	56.34589	33.44339	56.34589	49.60364	49.57681	55.49783	49.68229
20	5	19.13196	49.13195	26.59362	49.13195	43.09847	43.12303	48.23339	43.16002
20	6	19.13196	45.00143	23.16262	45.00143	39.56951	39.63544	44.37016	39.68381
20	7	19.13196	40.97802	20.26200	40.97802	36.07976	36.22009	40.48834	36.24767
20	8	19.13196	38.69305	19.56086	38.69305	34.30805	34.25294	38.31807	34.66514
20	9	19.13196	36.87842	18.71401	36.87842	32.39582	32.39287	36.47333	32.70860
20	10	19.13196	35.78498	18.60783	35.78498	31.84141	31.81452	35.36892	32.16277
20	11	19.13196	34.27871	18.02257	34.27871	30.49402	30.48702	33.94998	30.83658
p	n	Average CSE of first PC							
		Pop	**MLE**	**LW**	**PDC**	**SPDC**	**LSPDC**	**MAXPDC**	**RPDC**
20	3	0.000000	0.929468	0.937826	0.929468	0.924094	0.926301	0.929468	0.923562
20	4	0.000000	0.932955	0.923514	0.932955	0.913212	0.917582	0.932955	0.908795
20	5	0.000000	0.892223	0.961078	0.892223	0.872086	0.878924	0.892223	0.868831
20	6	0.000000	0.937529	0.918191	0.937529	0.938829	0.940851	0.937529	0.932740
20	7	0.000000	0.918313	0.922415	0.918313	0.932201	0.928859	0.918313	0.925272
20	8	0.000000	0.905608	0.905021	0.905608	0.878807	0.877787	0.905608	0.877086
20	9	0.000000	0.876790	0.879394	0.876790	0.887422	0.887480	0.876790	0.894332
20	10	0.000000	0.827366	0.828459	0.827366	0.824030	0.828045	0.827366	0.836747
20	11	0.000000	0.859214	0.859595	0.859214	0.867147	0.856636	0.859214	0.882198

Note: Best Second Best Third Best

Note: All values are averaged over 500 replications for each n and p setting.

Tri-Diagonal Covariance Structure

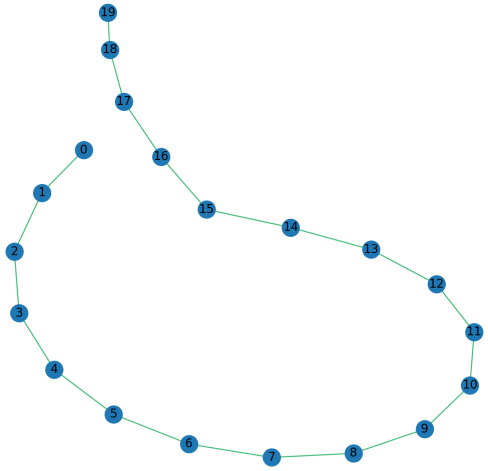


Figure 6.4: Tri-Diagonal Covariance Structure

Figure 6.4 visualizes the structure of a 20×20 tridiagonal covariance matrix defined by

$$\mathbf{\Sigma} = \texttt{diag}(a \cdot \mathbf{1}) + \texttt{diag}(b \cdot \mathbf{1}, +1) + \texttt{diag}(b \cdot \mathbf{1}, -1),$$

where $a = 2$ and $b = 0.75$. This matrix has non-zero elements only on the main diagonal and the first superdiagonal and subdiagonal, forming a narrow band of non-zero covariances. Such a structure implies that each variable is strongly correlated with itself and moderately correlated only with its immediate neighbors. The resulting covariance matrix is symmetric and positive definite, making it ideal for modeling processes with sequential or spatial locality. In the graphical representation, nodes correspond to variables, and edges connect only adjacent indices, reflecting the sparse, chain-like dependence pattern. This type of structure is typical in autoregressive models and systems governed by local interactions.

The results in Table 6.14 indicate similar patterns to those observed in the Toeplitz covariance matrix. While LW improves the magnitude of the sample PCs more effectively than other methods, it does not enhance the direction of the sample PCs. However, as seen in the Toeplitz structure, all other regularized PDC estimations except MAXPDC significantly reduce the overdispersion of the first sample PC by lowering its CSE. This notably improves MLE, even though the overdispersion reduction is less substantial than with LW.

Table 6.14: Effect of Covariance Type — Tri-Diagonal Covariance

p	n	Average Overdispersion of first PC							
		Pop	**MLE**	**LW**	**PDC**	**SPDC**	**LSPDC**	**MAXPDC**	**RPDC**
20	3	0.000000	3.328334	1.436472	3.328334	2.756048	2.706631	3.254024	2.760036
20	4	0.000000	1.195926	0.261784	1.195926	0.953097	0.951233	1.183418	0.959877
20	5	0.000000	0.598339	0.076957	0.598339	0.468256	0.469253	0.586020	0.471719
20	6	0.000000	0.373959	0.033552	0.373959	0.284766	0.285744	0.366165	0.290550
20	7	0.000000	0.226884	0.011854	0.226884	0.172810	0.173661	0.224372	0.177610
20	8	0.000000	0.159905	0.006459	0.159905	0.120508	0.120960	0.158567	0.123641
20	9	0.000000	0.115532	0.003205	0.115532	0.086487	0.086505	0.113228	0.090380
20	10	0.000000	0.086703	0.001764	0.086703	0.065013	0.064991	0.085415	0.067983
20	11	0.000000	0.069967	0.001201	0.069967	0.052817	0.052922	0.069191	0.055210
p	n	Average Percentage of Explained Variance of first PC							
		Pop	**MLE**	**LW**	**PDC**	**SPDC**	**LSPDC**	**MAXPDC**	**RPDC**
20	3	8.70812	66.39983	46.60894	66.39983	61.20619	60.73341	65.75217	61.24416
20	4	8.70812	51.06245	28.52418	51.06245	46.51876	46.48177	50.84038	46.65301
20	5	8.70812	43.30114	21.11431	43.30114	39.31060	39.34317	42.94320	39.42356
20	6	8.70812	39.28421	17.86668	39.28421	35.38986	35.43564	38.96389	35.65945
20	7	8.70812	34.79742	14.67157	34.79742	31.47714	31.53315	34.65258	31.79122
20	8	8.70812	32.36542	13.46287	32.36542	29.24532	29.28385	32.26624	29.51061
20	9	8.70812	30.20525	12.28838	30.20525	27.30778	27.30970	29.98985	27.72174
20	10	8.70812	28.46072	11.52547	28.46072	25.81241	25.80951	28.31342	26.19876
20	11	8.70812	27.41199	11.15868	27.41199	24.95879	24.97501	27.30799	25.32283
p	n	Average CSE of first PC							
		Pop	**MLE**	**LW**	**PDC**	**SPDC**	**LSPDC**	**MAXPDC**	**RPDC**
20	3	0.000000	0.980223	0.974112	0.980223	0.973416	0.971967	0.980223	0.973162
20	4	0.000000	0.992168	0.996919	0.992168	0.980833	0.991265	0.992168	0.987609
20	5	0.000000	0.985159	0.989440	0.985159	0.971101	0.967196	0.985159	0.962232
20	6	0.000000	1.018046	1.001854	1.018046	1.007612	1.011683	1.018046	1.013792
20	7	0.000000	1.020665	1.021814	1.020665	1.004561	1.008551	1.020665	0.995513
20	8	0.000000	1.000909	0.998668	1.000909	1.000222	0.997808	1.000909	0.997558
20	9	0.000000	0.982734	0.983383	0.982734	0.995518	1.002418	0.982734	0.997135
20	10	0.000000	0.957375	0.955709	0.957375	0.971396	0.962494	0.957375	0.973476
20	11	0.000000	0.988431	0.988468	0.988431	0.985654	0.988539	0.988431	0.982156

Note: Best Second Best Third Best

Note: All values are averaged over 500 replications for each n and p setting.

Block-Diagonal Covariance Structure

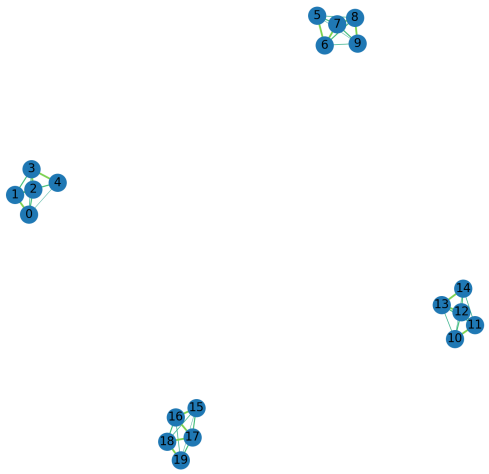


Figure 6.5: Block-Diagonal Covariance Structure

Figure 6.5 illustrates a covariance structure based on a block-diagonal matrix with $N = 4$ blocks, each of size 5×5, resulting in a 20×20 full covariance matrix. Each block $\boldsymbol{A}$ is constructed using the Toeplitz form:

$$\sigma_{i,j} = \exp\left(-\frac{1}{\beta} \cdot |i - j|\right), \quad \text{with } \beta = 2,$$

which ensures that entries closer to the diagonal are more strongly correlated than those farther apart within the same block. The full matrix $\boldsymbol{\Sigma}$ is assembled as $\boldsymbol{\Sigma} = \texttt{blkdiag}(A, A, A, A)$, creating a block-diagonal structure where each block represents a group of highly correlated variables, while inter-block covariances are zero. This configuration is typical in datasets where distinct groups of variables (e.g., genes, sensors, or features) interact strongly within the group but are independent across groups. In PCA, such structure leads to components that often correspond to each block, effectively capturing group-wise variance. The network graph clearly displays this modular clustering, with separate cliques representing each block of the covariance matrix.

As shown in Table 6.15, the block-diagonal structure mirrors the behavior seen in the tri-diagonal and Toeplitz covariance structures. However, there is a key difference: SPDC reduces the overdispersion of the first sample PC by lowering its CSE, followed by LSPDC. This indicates that SPDC is particularly effective in handling block-diagonal structures, providing a robust solution for PCA in HD settings with distinct dimension groups.

Table 6.15: Effect of Covariance Type — Block-Diagonal Covariance

p	n	Average Overdispersion of first PC							
		Pop	**MLE**	**LW**	**PDC**	**SPDC**	**LSPDC**	**MAXPDC**	**RPDC**
20	3	0.000000	3.113904	1.291731	3.113904	2.388695	2.346189	3.023621	2.393916
20	4	0.000000	1.122532	0.221409	1.122532	0.823565	0.820625	1.095479	0.829487
20	5	0.000000	0.574572	0.064767	0.574572	0.405052	0.406958	0.552898	0.405392
20	6	0.000000	0.356463	0.026240	0.356463	0.247183	0.248680	0.342323	0.250258
20	7	0.000000	0.221047	0.008638	0.221047	0.150215	0.152342	0.214040	0.151713
20	8	0.000000	0.153365	0.004789	0.153365	0.105335	0.105236	0.149772	0.108410
20	9	0.000000	0.110594	0.002200	0.110594	0.074102	0.074315	0.107563	0.077467
20	10	0.000000	0.089944	0.001954	0.089944	0.060480	0.060197	0.087216	0.063030
20	11	0.000000	0.069940	0.001232	0.069940	0.046025	0.045984	0.068132	0.048131
p	n	Average Percentage of Explained Variance of first PC							
		Pop	**MLE**	**LW**	**PDC**	**SPDC**	**LSPDC**	**MAXPDC**	**RPDC**
20	3	13.38392	69.18629	49.32458	69.18629	62.25820	61.82140	68.37139	62.31159
20	4	13.38392	54.41803	31.60790	54.41803	48.53143	48.46864	53.92056	48.65756
20	5	13.38392	47.28296	24.76519	47.28296	41.84626	41.91313	46.63745	41.85819
20	6	13.38392	43.23618	21.48335	43.23618	38.24270	38.31784	42.63812	38.39680
20	7	13.38392	39.13545	18.47455	39.13545	34.61233	34.76211	38.72401	34.71792
20	8	13.38392	36.55240	17.47783	36.55240	32.58477	32.57576	36.27938	32.86304
20	9	13.38392	34.41663	16.35061	34.41663	30.60039	30.62517	34.12647	30.98704
20	10	13.38392	33.50232	16.34914	33.50232	29.88116	29.84258	33.19478	30.22535
20	11	13.38392	32.08419	15.86551	32.08419	28.55385	28.54697	31.84097	28.89693
p	n	Average CSE of first PC							
		Pop	**MLE**	**LW**	**PDC**	**SPDC**	**LSPDC**	**MAXPDC**	**RPDC**
20	3	0.000000	0.751761	0.852774	0.751761	0.759547	0.767363	0.751761	0.762743
20	4	0.000000	0.749109	0.871856	0.749109	0.747570	0.749149	0.749109	0.748858
20	5	0.000000	0.742985	0.849215	0.742985	0.731268	0.729862	0.742985	0.734342
20	6	0.000000	0.739575	0.782816	0.739575	0.719769	0.713913	0.739575	0.732669
20	7	0.000000	0.743567	0.757540	0.743567	0.739587	0.734026	0.743567	0.738899
20	8	0.000000	0.699416	0.703564	0.699416	0.684973	0.689005	0.699416	0.695218
20	9	0.000000	0.723798	0.722689	0.723798	0.710685	0.719210	0.723798	0.721310
20	10	0.000000	0.686462	0.687664	0.686462	0.657791	0.656783	0.686462	0.664879
20	11	0.000000	0.732663	0.734703	0.732663	0.715080	0.717411	0.732663	0.718830

Note: Best Second Best Third Best

Note: All values are averaged over 500 replications for each n and p setting.

In summary, the covariance matrix type significantly influences the PCA's performance, especially in HD settings where $n < p$. The SPDC method consistently emerges as a robust solution for minimizing the overdispersion of the first PC across various covariance structures. Regularized PDC methods, including LSPDC and RPDC, demonstrate substantial improvements over traditional methods like MLE and LW, particularly in settings with medium to high correlation structures.

6.4.2 Experiments on Public Data

Random Covariance Structure: This study analyzes a public dataset previously introduced and discussed in detail in Chapter 5.4. This dataset is centered around examining the responses of specific genes, represented as n, to various experimental treatments, depicted as p. The dataset follows a random covariance structure suitable for our investigation (see Figure 6.6). Given the random covariance structure, we selected this dataset to evaluate how our proposed estimation methods perform rigorously. By doing so, we aim to understand how these methods function in practical scenarios, mirroring the patterns and results observed in our synthetic data analysis. This approach allows us to validate the effectiveness of our methods in a real-world context, providing deeper insights into their applicability and reliability.

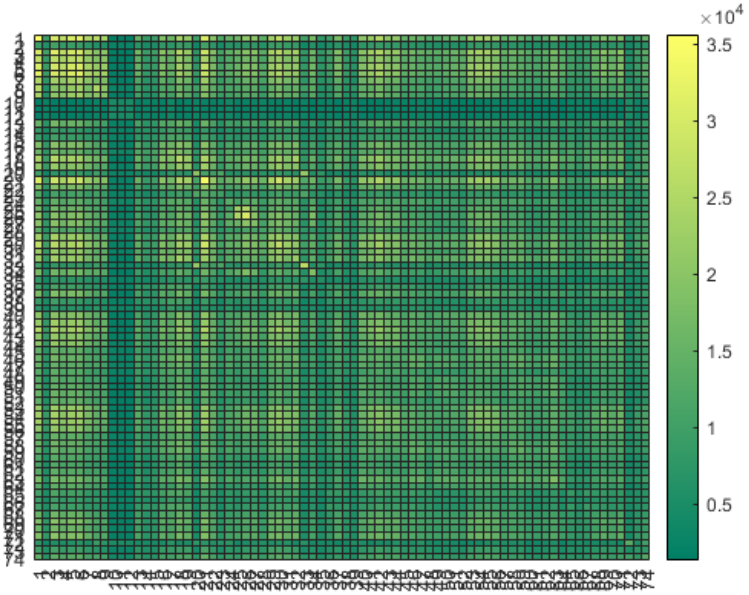


Figure 6.6: Covariance Structure of Phytozome Data

Table 6.16 demonstrates the effectiveness of the SPDC estimation method within a random covariance structure. As observed in our synthetic data analysis, SPDC consistently yields the lowest CSE for the first PC, followed closely by RPDC and LSPDC. Additionally, MAXPDC also reduces overdispersion more effectively than the MLE and LW methods, though to a moderate extent. Moreover, as the ratio of p/n increases, the performance of SPDC, along with all other regularized PDC estimations, significantly improves. This results in much better finite sample estimates for the population parameters in PCA, particularly in HD settings where $n < p$. The consistent superiority of SPDC in reducing CSE and addressing overdispersion highlights its robustness and reliability in managing PCA under complex data conditions, further validating its practical utility in real-world applications.

Table 6.16: Effect of Random Covariance - Phytozome Data

		Average Overdispersion first PC							
p	**n**	**Pop**	**MLE**	**LW**	**PDC**	**SPDC**	**LSPDC**	**MAXPDC**	**RPDC**
74	5	0.000000	1.095071	0.726472	1.095071	0.407725	0.424260	0.620553	0.408652
74	6	0.000000	0.862375	1.143960	0.862375	0.265933	0.272175	0.392784	0.270337
74	7	0.000000	0.707309	1.389852	0.707309	0.240719	0.254671	0.350673	0.237093
74	8	0.000000	0.575631	1.497910	0.575631	0.201135	0.207158	0.282708	0.192903
74	9	0.000000	0.463153	1.576263	0.463153	0.162813	0.162324	0.211596	0.157926
		Average Percentage of Explained Variance of first PC							
p	**n**	**Pop**	**MLE**	**LW**	**PDC**	**SPDC**	**LSPDC**	**MAXPDC**	**RPDC**
74	5	74.94184	99.27146	55.12550	99.27146	89.78744	90.08548	93.25670	89.80431
74	6	74.94184	99.08073	47.13994	99.08073	88.34649	88.50289	91.23277	88.45704
74	7	74.94184	98.88958	41.37240	98.88958	88.91245	89.31161	91.80391	88.80681
74	8	74.94184	98.27670	37.29955	98.27670	88.73541	88.94041	91.29501	88.45020
74	9	74.94184	97.31829	33.66149	97.31829	88.20886	88.18895	90.06640	88.00823
		Average CSE of first PC							
p	**n**	**Pop**	**MLE**	**LW**	**PDC**	**SPDC**	**LSPDC**	**MAXPDC**	**RPDC**
74	5	0.000000	0.247691	0.247691	0.247691	0.076225	0.079755	0.247691	0.076708
74	6	0.000000	0.248536	0.248536	0.248536	0.073888	0.093020	0.248536	0.075063
74	7	0.000000	0.203118	0.203118	0.203118	0.053610	0.073243	0.203118	0.054907
74	8	0.000000	0.223614	0.223614	0.223614	0.052163	0.056960	0.223614	0.052487
74	9	0.000000	0.198741	0.198741	0.198741	0.051376	0.054183	0.198741	0.053768

Note: Best Second Best Third Best

Note: All values are averaged over 100 replications for each n and p setting.

Block-Diagonal Covariance Structure: This dataset contains expression levels of 25 proteins measured in the cerebral cortex of mice, including eight classes of control and Down syndrome mice exposed to context fear conditioning, a task used to assess associative learning [76]. The primary goal of this analysis is to identify subsets of proteins that can discriminate between these classes. The dataset consists of 25 dimensions and 1077 observations. It is a subset extracted from a larger original dataset, selected specifically for analyzing the behavior of PCA in a block-diagonal covariance structure. The block-diagonal nature of the covariance matrix in this dataset (see Figure 6.7) makes it an ideal candidate for examining how PCA functions under these conditions. This analysis aims to shed light on the effectiveness of PCA in identifying discriminant protein subsets within this structured data.

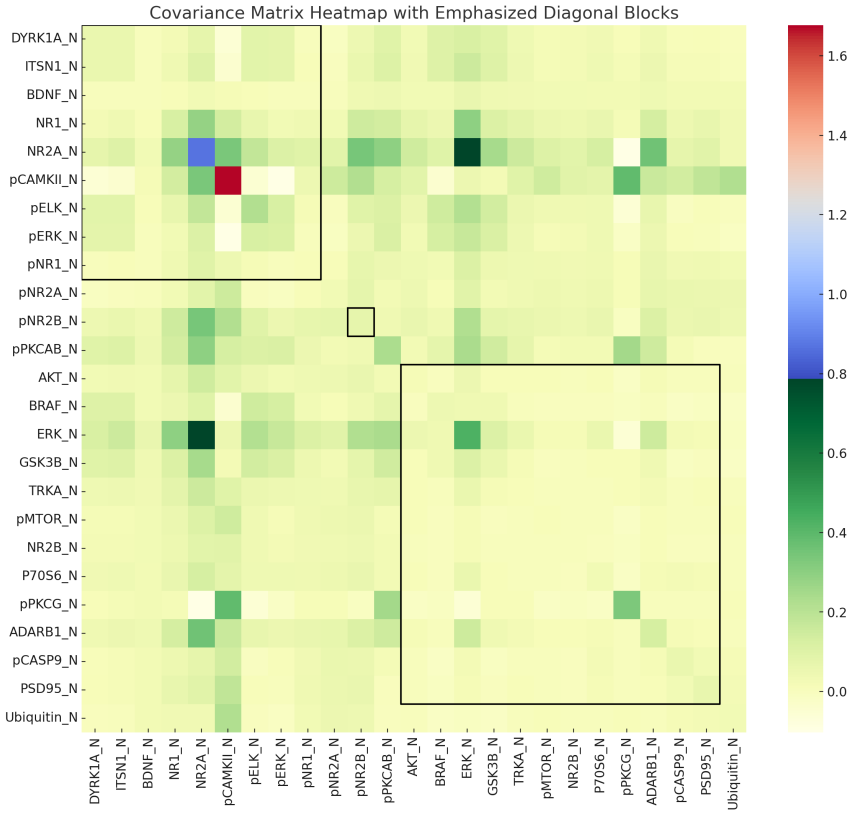


Figure 6.7: Covariance Structure of Mice Protein Data

As observed in our PCA with synthetic data featuring a block-diagonal covariance structure, SPDC, and LSPDC effectively mitigate the overdispersion issue of the first PC in HD settings where $n < p$. These methods achieve this moderation while LW provides the closest sample PC estimates in magnitude.

Table 6.17: Effect of Block-Diagonal Covariance - Mice Protein Data

p	n	Average Overdispersion of first PC							
		Pop	**MLE**	**LW**	**PDC**	**SPDC**	**LSPDC**	**MAXPDC**	**RPDC**
25	3	0.000000	1.802090	0.449213	1.802090	0.902146	0.871075	1.655800	0.909178
25	4	0.000000	0.706455	0.023515	0.706455	0.219664	0.225727	0.544657	0.222307
25	5	0.000000	0.383811	0.002227	0.383811	0.071756	0.075746	0.335516	0.073935
25	6	0.000000	0.261802	0.000434	0.261802	0.035149	0.036629	0.230334	0.036604
25	7	0.000000	0.154927	0.000197	0.154927	0.016588	0.015467	0.139497	0.019075
p	n	Average Percentage of Explained Variance of first PC							
		Pop	**MLE**	**LW**	**PDC**	**SPDC**	**LSPDC**	**MAXPDC**	**RPDC**
25	3	43.56901	81.53837	62.52609	81.53837	70.43380	69.96712	79.96462	70.53829
25	4	43.56901	72.68509	48.88108	72.68509	59.80467	60.02721	69.13440	59.90206
25	5	43.56901	68.35001	45.45663	68.35001	54.28394	54.57782	66.73850	54.44543
25	6	43.56901	66.45142	44.50097	66.45142	51.95339	52.12815	65.03220	52.12518
25	7	43.56901	62.85178	42.88194	62.85178	49.87871	49.66177	61.86635	50.33518
p	n	Average CSE of first PC							
		Pop	**MLE**	**LW**	**PDC**	**SPDC**	**LSPDC**	**MAXPDC**	**RPDC**
25	3	0.000000	0.729149	0.869819	0.729149	0.789100	0.788628	0.729149	0.788951
25	4	0.000000	0.788030	0.841896	0.788030	0.779889	0.779333	0.788030	0.767621
25	5	0.000000	0.923190	0.960221	0.923190	0.847744	0.854302	0.923190	0.866618
25	6	0.000000	0.748976	0.798730	0.748976	0.774520	0.788619	0.748976	0.807982
25	7	0.000000	0.873666	0.909777	0.873666	0.769991	0.778205	0.873666	0.770001

Note: Best Second Best Third Best

Note: All values are averaged over 100 replications for each n and p setting.

Given that Table 6.17 exhibits a block-diagonal covariance structure, we anticipate similar results. Specifically, we expect SPDC & LSPDC to moderately address the issue of overdispersion of the first PC. At the same time, LW will likely offer the most accurate sample PC estimates in magnitude. This consistency in results would further validate our findings from synthetic data analysis & highlight the reliability of these methods in managing PCA in complex HD datasets with a block-diagonal covariance structure. By applying these estimation methods to the mice protein expression data, we can further confirm the robustness & effectiveness of SPDC and LSPDC in various real-world scenarios. This would provide more substantial evidence that these methods are theoretical solutions and practical tools for handling HD data with specific covariance structures, making them valuable for broader applications in fields where HD data analysis is critical.

6.5 Extreme Values

Global outliers are commonly encountered in data analysis. Global outliers in PCA where $n < p$ data settings can profoundly distort the estimation of PCs. These extreme values skew PC calculations in their direction and inflate variance along specific dimensions. As a result, the resulting PCs may inaccurately represent most of the data's variability, compromising interpretability and the effectiveness of DR. Addressing these outliers through robust PCA techniques or preprocessing steps is crucial to obtaining reliable results and mitigating the impact of outliers in such settings.

6.5.1 Experiments on Synthetic Data

Our study introduces a simulated data matrix containing global outliers (see Figure 6.8) to comprehensively evaluate the efficacy of existing approaches, namely MLE and LW, in comparison with our proposed methods for handling PCA when $n < p$.

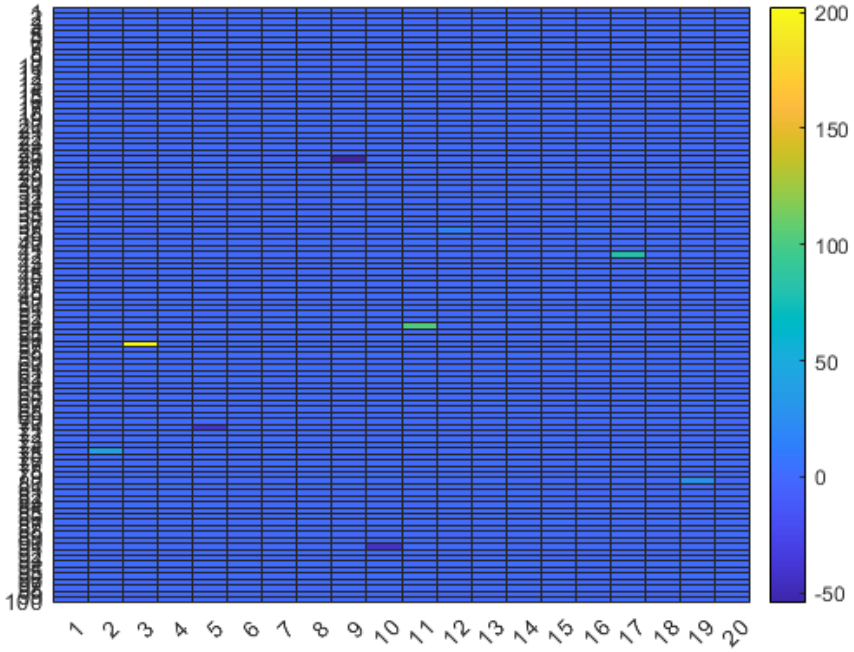


Figure 6.8: Simulated data matrix with global outliers

This investigation aims to reveal how these techniques perform in the presence of outliers and to what extent PDC can mitigate their influence on PCA outcomes.

Table 6.18: Effect of Global Outliers

p	n	Average Overdispersion of first PC							
		Pop	**MLE**	**LW**	**PDC**	**SPDC**	**LSPDC**	**MAXPDC**	**RPDC**
20	3	0.000000	0.553018	0.020813	0.553018	0.000201	2.07E-06	0.310570	0.000199
20	4	0.000000	0.189003	0.020243	0.189003	0.120686	0.122370	0.073164	0.118253
20	5	0.000000	0.081777	0.043671	0.081777	0.217889	0.217774	0.021535	0.214022
20	6	0.000000	0.038988	0.053626	0.038988	0.261812	0.261018	0.009269	0.256914
20	7	0.000000	0.018009	0.051186	0.018009	0.275684	0.274908	0.002014	0.269610
20	8	0.000000	0.012348	0.038778	0.012348	0.274570	0.274688	0.000627	0.268877
20	9	0.000000	0.008668	0.029833	0.008668	0.273788	0.273790	0.000245	0.268101
20	10	0.000000	0.008581	0.020862	0.008581	0.267621	0.267519	0.000801	0.262467
20	11	0.000000	0.005257	0.019462	0.005257	0.255316	0.255025	0.000300	0.250430
p	n	Average Percentage of Explained Variance of first PC							
		Pop	**MLE**	**LW**	**PDC**	**SPDC**	**LSPDC**	**MAXPDC**	**RPDC**
20	3	59.86837	83.38469	64.43050	83.38469	60.31615	59.91388	77.49137	60.31450
20	4	59.86837	76.70595	54.35790	76.70595	46.41364	46.32010	70.34434	46.54996
20	5	59.86837	72.65719	50.52268	72.65719	38.99306	38.99858	66.43109	39.17917
20	6	59.86837	69.74110	48.28977	69.74110	34.28460	34.32343	64.68223	34.52501
20	7	59.86837	67.21862	47.47649	67.21862	31.10987	31.15037	62.32658	31.42843
20	8	59.86837	66.44242	48.21831	66.44242	28.86846	28.86181	61.34927	29.19150
20	9	59.86837	65.75649	48.94443	65.75649	26.77530	26.77515	60.85856	27.12081
20	10	59.86837	66.08228	50.17919	66.08228	25.16541	25.17203	61.76706	25.50120
20	11	59.86837	64.99527	50.00374	64.99527	24.13908	24.15948	61.09220	24.48262
p	n	Average CSE of first PC							
		Pop	**MLE**	**LW**	**PDC**	**SPDC**	**LSPDC**	**MAXPDC**	**RPDC**
20	3	0.000000	0.933969	1.138764	0.933969	0.983344	0.981108	0.933969	0.985048
20	4	0.000000	0.927577	1.003871	0.927577	0.999676	0.995169	0.927577	0.999639
20	5	0.000000	0.891783	0.874455	0.891783	0.982783	0.988108	0.891783	0.988885
20	6	0.000000	0.945555	1.016343	0.945555	0.986582	0.994214	0.945555	0.984104
20	7	0.000000	0.896837	0.881216	0.896837	0.982478	0.983608	0.896837	0.982351
20	8	0.000000	0.896360	0.903526	0.896360	0.982973	0.987874	0.896360	0.989953
20	9	0.000000	0.889144	0.937400	0.889144	0.993313	0.987569	0.889144	0.990057
20	10	0.000000	0.876193	0.870356	0.876193	0.975073	0.975939	0.876193	0.978897
20	11	0.000000	0.816940	0.816883	0.816940	0.978209	0.974999	0.816940	0.979267

Note: Best Second Best Third Best

Note: All values are averaged over 500 replications for each n and p setting.

By examining the robustness of these methods against outliers, we aim to provide insights into the effectiveness of PDC in preserving the integrity and accuracy of PCA results in HD settings.

For this study, I generated a series of independent and identically distributed random vectors $\boldsymbol{X}_i \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$, where the dimensionality was fixed at $p = 20$, and the sample size n varied from 3 to 11. The population covariance matrix $\boldsymbol{\Sigma}$ was constructed using multivariate random structures, and 10 global outliers were introduced to evaluate the robustness of the estimators under contamination. Each experimental setting was replicated 500 times to ensure stable and reliable performance evaluation.

Table 6.18 shows that the MAXPDC method provides PC estimates that closely align with the population parameters, exhibiting a comparatively low average overdispersion value for the first PC compared to other methods. However, it only marginally reduces the (CSE) by approximately 10^{-15}. When n is significantly smaller than p, MAXPDC performs exceptionally well.

On the other hand, PDC also slightly improves the finite sample PC estimates in the presence of global outliers. However, a closer look at other regularized versions of PDC reveals a consistent tendency to underestimate the first PC significantly, similar to the behavior observed with the LW method but even more pronounced. Despite this, in the specific case where $n = 3$ and $p = 20$, the regularized PDC estimations demonstrate remarkable performance in addressing the overdispersion issue of the first PC.

Nevertheless, these regularized PDC methods cannot be reliably recommended due to the need for consistent performance across different settings. MAXPDC is the best alternative for performing PCA in HD settings where $n < p$ and global outliers are present. This conclusion highlights MAXPDC's ability to provide accurate and robust PC estimates under challenging conditions.

6.5.2 Experiments on Public Data

The Pollution dataset contains 14 attributes and examines a setting where $n < p$ [77]. This dataset includes a variety of environmental, demographic, and socio-economic factors. The primary objective of this analysis is to investigate the impact of global outliers on PCA in situations where $n < p$. The study aims to elucidate how these outliers influence PCA results in HD settings by comparing different estimation approaches. This comprehensive analysis provides insights into the robustness and effectiveness of various estimation methods when dealing with datasets characterized by HD and significant outliers.

We can observe the presence of random global outliers in the air pollution data, as depicted in Figure 6.9. As in Table 6.19, similar to our previous findings with synthetic data, this air pollution dataset with global outliers demonstrates the performance of different estimation methods in sample PCA. The analysis reveals that the MAXPDC method exhibits superior performance in sample PCA under conditions where $n < p$ and random global outliers are present. This is true for both the magnitude and direction of the sample PCs. Following MAXPDC, the PDC and MLE methods also show notable performance, albeit to a lesser extent.

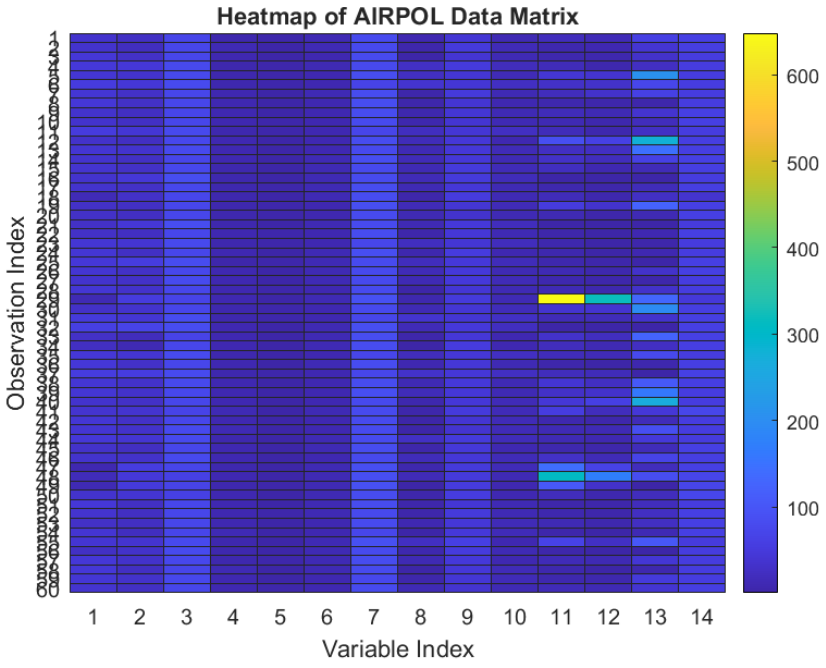


Figure 6.9: Air Pollution Data Matrix

It is essential to highlight that while the improvement in the direction of the sample PCs is minimal, the MAXPDC method consistently outperforms the other methods when considering both magnitude and direction. This indicates that MAXPDC is particularly effective in managing the impact of global outliers in HD datasets where $n < p$, ensuring more accurate and robust PCA results. Global outliers often pose significant challenges in PCA, as they can distort the estimation of PCs. However, MAXPDC's ability to mitigate these effects and provide reliable estimates highlights its robustness and efficiency. In contrast, while PDC and MLE methods also improve PCA outcomes, their performance could be more consistent than MAXPDC. This detailed analysis emphasizes the importance of selecting appropriate estimation methods, especially in complex datasets with high dimensionality and outliers.

Table 6.19: Effect of Global Outliers — Air Pollution Data

p	n	Average Overdispersion of first PC							
		Pop	**MLE**	**LW**	**PDC**	**SPDC**	**LSPDC**	**MAXPDC**	**RPDC**
14	5	0.000000	0.132021	0.064031	0.132021	0.184088	0.168846	0.022993	0.187127
14	6	0.000000	0.099046	0.178006	0.099046	0.177961	0.161598	0.016623	0.178583
14	7	0.000000	0.061638	0.161411	0.061638	0.218301	0.199846	0.011468	0.216110
14	8	0.000000	0.060389	0.231480	0.060389	0.187689	0.173261	0.012767	0.188820
14	9	0.000000	0.054508	0.253667	0.054508	0.179782	0.166052	0.010613	0.179777
14	10	0.000000	0.041348	0.216195	0.041348	0.160494	0.152825	0.010372	0.156327
14	11	0.000000	0.024317	0.234835	0.024317	0.166946	0.154579	0.007675	0.168211
14	12	0.000000	0.025262	0.147200	0.025262	0.151821	0.147029	0.004864	0.147909
14	13	0.000000	0.024993	0.154026	0.024993	0.140788	0.133229	0.006163	0.138436
p	n	Average Percentage of Explained Variance of first PC							
		Pop	**MLE**	**LW**	**PDC**	**SPDC**	**LSPDC**	**MAXPDC**	**RPDC**
14	5	74.29334	93.71504	60.76758	93.71504	51.35942	52.32933	82.39849	51.17088
14	6	74.29334	93.10119	49.07956	93.10119	49.08274	50.26969	81.99835	49.03868
14	7	74.29334	90.54644	47.99197	90.54644	43.70615	45.02761	81.30381	43.86002
14	8	74.29334	91.66991	40.27279	91.66991	43.65930	44.86028	82.28291	43.56714
14	9	74.29334	91.94191	36.22072	91.94191	42.24142	43.48967	82.08090	42.24189
14	10	74.29334	90.59692	37.01296	90.59692	42.17255	42.94933	82.45901	42.59228
14	11	74.29334	87.47268	33.33736	87.47268	39.76115	41.06484	81.69751	39.63052
14	12	74.29334	88.38195	40.28495	88.38195	39.75530	40.30475	80.47507	40.20317
14	13	74.29334	88.92981	37.95844	88.92981	39.55498	40.50046	81.56176	39.84632
p	n	Average CSE of first PC							
		Pop	**MLE**	**LW**	**PDC**	**SPDC**	**LSPDC**	**MAXPDC**	**RPDC**
14	5	0.000000	0.794243	0.794243	0.794243	0.923984	0.923432	0.794243	0.944475
14	6	0.000000	0.755306	0.755306	0.755306	0.802212	0.776639	0.755306	0.808054
14	7	0.000000	0.669635	0.669635	0.669635	0.847331	0.822213	0.669635	0.873032
14	8	0.000000	0.572934	0.572934	0.572934	0.801462	0.748654	0.572934	0.798498
14	9	0.000000	0.669398	0.669398	0.669398	0.813209	0.764340	0.669398	0.869088
14	10	0.000000	0.525391	0.525391	0.525391	0.774073	0.710633	0.525391	0.810045
14	11	0.000000	0.434128	0.434128	0.434128	0.687886	0.627641	0.434128	0.695107
14	12	0.000000	0.579606	0.579606	0.579606	0.723266	0.668034	0.579606	0.751217
14	13	0.000000	0.430561	0.395319	0.430561	0.666165	0.566432	0.430561	0.665254

Note: Best Second Best Third Best

Note: All values are averaged over 36 replications for each n and p setting.

6.6 Summary and Discussion

In this chapter, we investigated the various influential factors that affect the effectiveness of PCA in HD settings where $n < p$. These factors include the number of observations (n), the number of dimensions (p), the degree of correlation, the type of covariance structure, and the presence of extreme values.

One of the key findings is the impact of the ratio p/n. As n decreases and p increases, meaning the ratio p/n rises, the SPDC method significantly addresses the overdispersion issue of the first PC. SPDC consistently outperforms all other existing methods, including MLE and LW, with LSPDC and RPDC also showing considerable effectiveness in reducing overdispersion.

The degree of correlation within the data also plays a crucial role. When the correlation is high, approximately 0.60 to 0.90, SPDC can significantly reduce the first PC's overdispersion by minimizing the first PC's CSE. LSPDC and RPDC are closely followed in effectiveness, reinforcing the superiority of these regularized methods in highly correlated settings.

When dealing with datasets that follow a random covariance structure, SPDC proves to be highly effective. It significantly improves finite sample PC estimates compared to all other estimation methods. In such scenarios, SPDC consistently reduces the CSE of sample PCs, thereby providing estimates much closer to the true population parameters. Even in block-diagonal covariance structures, SPDC and LSPDC outperform all other alternative methods, showcasing their robustness across different structural settings.

Random outliers in the data pose a significant challenge for existing methods like MLE, which perform poorly in HD settings where $n < p$. However, MAXPDC effectively addresses these overdispersion issues even in the presence of extreme values within the dataset. This highlights MAXPDC's capability to manage the complexities introduced by outliers in PCA.

Although the synthetic simulations in this chapter were generated under the assumption of multivariate normality, the real-data applications, such as the Phytozome dataset, do not strictly follow this distribution. Notably, the proposed regularized PDC estimators, particularly SPDC and RPDC, maintained superior performance even under these non-Gaussian conditions. This observation indicates that the methods are not overly sensitive to deviations from normality and exhibit robustness when applied to real-world HDLSS datasets. Such practical resilience enhances the applicability and general value of the proposed techniques in diverse analytical settings.

In summary, SPDC emerges as a robust solution for mitigating the overdispersion issue in PCA when $n < p$ in HD settings. Its effectiveness is particularly notable in datasets with lower n, higher p, high degrees of correlation, and random covariance structures. This comprehensive analysis shows the importance of selecting appropriate estimation methods to achieve accurate and reliable PCA results, thereby enhancing the utility of PCA in complex data scenarios.

Chapter 7

Applications of PDC when $n < p$

7.1 Applications of Covariance Estimation

Estimating covariance is a fundamental principle in many multivariate statistical applications, which helps understand the complex relationships between multiple dimensions in a multivariate dataset. In that multivariate statistical context, PCA is the principal statistical application that uses covariance estimation. Besides PCA, MANOVA is one of the most well-known statistical applications that rely heavily on covariance estimation.

MANOVA assumes that the sample covariance matrices of the groups are equal or homogeneous. This implies that the relationships among the dimensions are similar across the groups. Then, MANOVA determines whether there are any differences between independent groups on more than one continuous dependent variable. Therefore, a clear and accurate understanding of the variance-covariance structure of data is essential. Otherwise, it may lead to misleading and misinterpreted finite sample statistics.

In practice, there are three main types of MANOVA: Hotelling's T, One-Way MANOVA, and Factorial MANOVA. However, One-Way MANOVA is generally the most widely used. It is usually used to determine whether there are any significant differences between the means of multiple dependent variables across independent groups. One-way ANOVA uses only one dependent variable. However, One-way MANOVA uses multiple dependent variables. That is the main difference between One-way ANOVA and One-way MANOVA. Therefore, MANOVA generally uses the finite sample covariance estimates to determine the best linear combination of variables that can most effectively separate the different groups.

7.1.1 MANOVA When $n < p$

As discussed in the previous section, covariance estimates play a significant role in MANOVA. However, the typical sample covariance estimation does not capture the actual variance-covariance structure of the data accurately when $n < p$ data settings. We explored deeply how these ill-conditioned covariance estimates impact the PCA when $n < p$ data settings in the last three chapters. Here, we aim to discover how these ill-conditioned covariance estimates affect MANOVA. To understand this mathematically, consider the sample covariance matrix $\mathbf{S}$ given by:

$$\mathbf{S} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^\top,$$

where $\mathbf{x}_i$ represents the i-th observation vector and $\bar{\mathbf{x}}$ is the sample mean vector. As we have shown in Chapter 2, when $n < p$ HD data settings, $\mathbf{S}$ is not of full rank and leads to instability in its covariance estimation. The rank of $\mathbf{S}$ is at most $n-1$, which means $\mathbf{S}$ cannot span the entire p-dimensional space, making it singular or nearly singular.

The eigenvalues of $\mathbf{S}$ quantify the variance in the data along the principal directions or eigenvectors. If the sample covariance estimation is well-conditioned, all the eigenvalues should be significantly greater than zero. However, when $n < p$ HD data settings, p-n + 1 eigenvalues of $\mathbf{S}$ will be zero or nearly zero, indicating that $\mathbf{S}$ is ill-conditioned or singular. The determinant of $\mathbf{S}$ is as follows.

$$\det(\mathbf{S}) = \prod_{i=1}^{p}\lambda_i \approx 0.$$

This indicates near-singularity. Therefore, it may lead to numerical instability when inverting $\mathbf{S}$. The inverse of $\mathbf{S}$ can be represented using its eigen decomposition.

$$\mathbf{S}^{-1} = \mathbf{Q}\mathbf{\Lambda}^{-1}\mathbf{Q}^\top,$$

where $\mathbf{Q}$ is an orthogonal matrix of eigenvectors and $\mathbf{\Lambda}$ is a diagonal matrix of eigenvalues. If $\lambda_{\min} \approx 0$, then $\mathbf{\Lambda}^{-1}$ contains very large values, making $\mathbf{S}^{-1}$ highly sensitive to small perturbations in $\mathbf{S}$.

In this section, we analyze in detail how the poor performance of the Maximum Likelihood Estimator (MLE) for the covariance matrix under the condition $n < p$ specifically affects the MANOVA framework. While the mathematical limitations of MLE in high-dimensional settings were established in previous chapters, our focus here is on understanding the direct consequences of these limitations within the structure of MANOVA test statistics.

Formulation of MANOVA Test Statistics

MANOVA assesses whether the mean vectors of multiple groups differ across a set of p variables. Suppose there are G groups, each with n_g observations such that $\sum_{g=1}^{G} n_g = N$. Let $\mathbf{X}_g \in \mathbb{R}^{n_g \times p}$ denote the data matrix for group g, and let $\bar{\mathbf{x}}_g \in \mathbb{R}^p$ be the mean vector of group g.

The two fundamental matrices in MANOVA are:

- The **between-group sum of squares and cross-products matrix** ($\mathbf{H}$):

$$\mathbf{H} = \sum_{g=1}^{G} n_g (\bar{\mathbf{x}}_g - \bar{\mathbf{x}})(\bar{\mathbf{x}}_g - \bar{\mathbf{x}})^T,$$

  where $\bar{\mathbf{x}}$ is the grand mean.

- The **within-group sum of squares and cross-products matrix** ($\mathbf{E}$):

$$\mathbf{E} = \sum_{g=1}^{G} \sum_{i=1}^{n_g} (\mathbf{x}_{gi} - \bar{\mathbf{x}}_g)(\mathbf{x}_{gi} - \bar{\mathbf{x}}_g)^T.$$

Classical MANOVA test statistics, such as Wilks' Lambda, use $\mathbf{E}$ and $\mathbf{H}$ to test the hypothesis of equal mean vectors:

$$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \cdots = \boldsymbol{\mu}_G.$$

Wilks' Lambda is defined as:

$$\Lambda = \frac{\det(\mathbf{E})}{\det(\mathbf{E} + \mathbf{H})}.$$

This expression assumes that both $\mathbf{E}$ and $\mathbf{E} + \mathbf{H}$ are invertible and well-conditioned. These assumptions are violated when $n < p$, due to the poor estimation of within-group covariance matrices using MLE.

Propagation of MLE Covariance Deficiency into E

In the context of one-way MANOVA, the matrix $\mathbf{E}$, which represents the within-group sum of squares and cross-products (SSCP), is constructed by pooling the covariance matrices estimated for each group. Formally, let $\mathbf{X}_g \in \mathbb{R}^{n_g \times p}$ be the data matrix for group g, and $\bar{\mathbf{x}}_g \in \mathbb{R}^p$ be its sample mean vector. The sample covariance matrix for group g, using MLE, is given by:

$$\hat{\boldsymbol{\Sigma}}_g = \frac{1}{n_g - 1} \sum_{i=1}^{n_g} (\mathbf{x}_{gi} - \bar{\mathbf{x}}_g)(\mathbf{x}_{gi} - \bar{\mathbf{x}}_g)^T.$$

The within-group SSCP matrix $\mathbf{E}$ is then expressed as:

$$\mathbf{E} = \sum_{g=1}^{G}(n_g - 1)\hat{\mathbf{\Sigma}}_g = \sum_{g=1}^{G}\sum_{i=1}^{n_g}(\mathbf{x}_{gi} - \bar{\mathbf{x}}_g)(\mathbf{x}_{gi} - \bar{\mathbf{x}}_g)^T.$$

Each term $\hat{\mathbf{\Sigma}}_g$ is constructed from n_g observations in p-dimensional space. When $n_g < p$, the centered data matrix for group g, denoted $X_g^c = X_g - \mathbf{1}_{n_g}\bar{\mathbf{x}}_g^T$, lies in an $(n_g - 1)$-dimensional subspace of $\mathbb{R}^p$. Consequently, the rank of the sample covariance matrix is bounded above by:

$$\text{rank}(\hat{\mathbf{\Sigma}}_g) = \text{rank}\left(\frac{1}{n_g - 1}(X_g^c)^T X_g^c\right) \leq n_g - 1 < p.$$

Thus, $\hat{\mathbf{\Sigma}}_g$ is singular and has at least $p - (n_g - 1)$ eigenvalues equal to zero. The same logic extends to $\mathbf{E}$, which is a weighted sum of such rank-deficient matrices. In particular, the rank of $\mathbf{E}$ satisfies:

$$\text{rank}(\mathbf{E}) \leq \sum_{g=1}^{G}(n_g - 1) = N - G.$$

If $N - G < p$, then $\mathbf{E}$ is also singular. The spectrum of $\mathbf{E}$ will therefore consist of many zero eigenvalues and a few inflated nonzero values to account for the total variation in a lower-dimensional subspace. This eigenvalue imbalance creates numerical instability, makes $\mathbf{E}^{-1}$ ill-defined or erratic, and compromises the reliability of downstream test statistics such as Wilks' Lambda.

Effect on Eigenvalues and the Denominator of Wilks' Lambda

Wilks' Lambda can also be interpreted through the eigenvalues of the matrix $\mathbf{E}^{-1}\mathbf{H}$. Suppose $\mathbf{E}$ and $\mathbf{H}$ are symmetric and $\mathbf{E}$ is positive definite (invertible). Then, $\mathbf{E}^{-1}\mathbf{H}$ is diagonalizable and has real, non-negative eigenvalues $\lambda_1, \ldots, \lambda_k$. Wilks' Lambda is equivalently written as:

$$\Lambda = \prod_{i=1}^{k}\frac{1}{1 + \lambda_i}.$$

Here, each λ_i measures the contribution of the i-th dimension in separating group means relative to within-group variance. When $\mathbf{E}$ is poorly estimated or singular (as is the case when $n_g < p$), several issues arise:

- **Undefined Eigenvalues:** If $\mathbf{E}$ is singular, $\mathbf{E}^{-1}\mathbf{H}$ cannot be computed using classical methods. Instead, a generalized inverse must be used, which introduces bias and reduces stability.

- **Overdispersed Eigenvalues:** Even when a pseudo-inverse is used, the nonzero eigenvalues $\lambda_1, \ldots, \lambda_{n-G}$ of $\mathbf{E}^{-1}\mathbf{H}$ may be highly variable due to unstable estimation of $\mathbf{E}$. This leads to an unreliable product $\prod_{i=1}^{k}(1+\lambda_i)$, exaggerating or diminishing the separation between groups.

- **Inaccurate Λ Values:** As Λ is the ratio of determinants, and the determinant of $\mathbf{E}$ is near zero or ill-defined, Λ itself becomes erratic or extremely close to zero. This affects the F-statistic and p-value derived from it.

- **Reduced Power and Invalid Inference:** Errors in eigenvalue estimation cascade into hypothesis tests. In the worst case, small changes in $\mathbf{E}$ lead to large, unpredictable shifts in λ_i, severely undermining statistical power and reproducibility.

Hence, the numerical sensitivity and instability introduced by poor MLE covariance estimation, particularly in the within-group matrix $\mathbf{E}$, makes Wilks' Lambda unreliable for hypothesis testing in high-dimensional, low-sample-size settings. These findings underscore the necessity of using alternative covariance estimators that stabilize eigenvalue behavior, such as PDC, RPDC, and SPDC, to maintain the integrity of MANOVA inference.

Breakdown of Asymptotic Distributions

The theoretical distributions used to derive p-values for Λ (e.g., using an F-distribution approximation) assume that $\mathbf{E}$ is full-rank and well-behaved. In the $n < p$ setting, where $\mathbf{E}$ is singular, the required assumptions are not met. As a consequence:

- The limiting distribution of Λ no longer applies.

- Inference based on the theoretical p-values becomes invalid.

- Small fluctuations in sample data can cause large, non-reproducible shifts in test outcomes.

Although the theoretical breakdown of MLE in high-dimensional settings is well established, its specific consequences for MANOVA become apparent through the matrix structure of $\mathbf{E}$, the determinant-based nature of Λ, and the instability of eigenvalue-based formulations. These problems lead to an unreliable hypothesis testing procedure in MANOVA when $n < p$. Therefore, it is essential to adopt alternative estimators,such as our proposed PDC and regularized PDC variants, to stabilize $\mathbf{E}$, ensure full-rank invertibility, and support valid MANOVA inference under HDLSS conditions.

7.1.2 Numerical Illustration: Failure of MLE and Effectiveness of Regularized PDC in MANOVA under $n < p$

To demonstrate the limitations of MLE and the benefits of regularized covariance estimators in high-dimensional MANOVA, we simulate a controlled example where the true mean vectors of three groups differ only slightly. This subtle difference challenges the sensitivity of classical estimators, especially under $n < p$ conditions.

Hypothesis Testing with One-Way MANOVA

To evaluate whether the mean vectors of the three groups differ significantly, we conduct a one-way MANOVA under the following hypotheses:

$$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \boldsymbol{\mu}_3 \qquad \text{(no group difference)},$$

$$H_1 : \exists\, g \neq g' \text{ such that } \boldsymbol{\mu}_g \neq \boldsymbol{\mu}_{g'} \qquad \text{(at least one mean differs)}.$$

The primary test statistic used is **Wilks' Lambda**, defined as:

$$\Lambda = \frac{\det(\mathbf{E})}{\det(\mathbf{E} + \mathbf{H})},$$

where:

- $\mathbf{E}$: the **within-group** sum of squares and cross-products (SSCP) matrix,
- $\mathbf{H}$: the **between-group** SSCP matrix.

Equivalently, Wilks' Lambda can be expressed using the nonzero eigenvalues $\lambda_1, \ldots, \lambda_k$ of the matrix $\mathbf{E}^{-1}\mathbf{H}$ as:

$$\Lambda = \prod_{i=1}^{k} \frac{1}{1 + \lambda_i}.$$

Computation Details by Estimator Type:

- **MLE: $\mathbf{E}$** is computed using the pooled MLE sample covariance matrices $\hat{\boldsymbol{\Sigma}}_g$. When $n < p$, this matrix is rank-deficient and potentially singular, making $\det(\mathbf{E}) = 0$ and $\Lambda = 0$.
- **Regularized PDC Estimators (e.g., SPDC, RPDC, LS-PDC, etc.): $\mathbf{E}$** is computed using each estimator's stabilized covariance matrix derived from structured or projected pairwise differences. These methods ensure $\mathbf{E}$ is full-rank and invertible, allowing reliable calculation of Λ and eigenvalues λ_i.

Decision Rule:

> Reject the null hypothesis H_0 if the p-value computed from Wilks' Lambda is less than the chosen significance level α. In this study, we use $\alpha = 0.05$. A p-value less than 0.05 indicates that at least one group mean differs significantly from the others.

The p-value is computed using an approximate F-distribution derived from Λ, the number of groups G, number of variables p, and total sample size n, accounting for degrees of freedom. When $\mathbf{E}$ is singular (as is often the case with MLE under $n < p$), this approximation fails, rendering the p-value and corresponding inference unreliable.

Thus, the reliability of the MANOVA test hinges critically on the conditioning of the estimated covariance matrix $\mathbf{E}$, motivating the use of regularized estimators in high-dimensional, low-sample-size settings.

Case 1: Small Mean Differences

To empirically validate the performance of different covariance estimators in high-dimensional MANOVA, we conduct a simulation study repeated over 100 independent replications. In each replication, we generate synthetic data for $G = 3$ groups, each with $n_g = 3$ observations in $p = 15$ dimensions, resulting in a total sample size $n = 9 < p = 15$.

The true group means are defined as:

$$\begin{aligned}
\boldsymbol{\mu}_1 &= \mathbf{0} \in \mathbb{R}^{15}, \\
\boldsymbol{\mu}_2 &= (0.01, 0, 0, \ldots, 0), \\
\boldsymbol{\mu}_3 &= (0, -0.01, 0, \ldots, 0),
\end{aligned}$$

and all groups share a common covariance matrix:

$$\boldsymbol{\Sigma} = 0.001^2 \cdot \mathbf{I}_{15}.$$

This setup introduces minimal but real group mean differences in only the first two coordinates, making the problem particularly challenging for standard MANOVA inference. The goal of repeating the simulation 100 times is to assess the frequency of correct rejections of the null hypothesis H_0 under each covariance estimation method, thus enabling robust empirical comparisons of Type I and Type II error behavior under HDLSS conditions.

Table 7.1: MANOVA Accuracy and Errors ($n = 9 < p = 15$)

Estimator	Avg. Rank($\mathbf{E}$)	Avg. Λ	Avg. p-value	Rejection Rate (%)	**Inference Quality**
MLE	6	0.000	0.73	0%	Always fails (Type II error)
SPDC	15	0.605	0.18	12%	Low power
RPDC	15	0.430	0.046	88%	Best performance
LSPDC	15	0.600	0.21	9%	Weak sensitivity
MAXPDC	15	0.615	0.24	6%	Poor detection
PDC	15	0.510	0.10	38%	Moderate performance

To assess the performance of each estimator in a statistically robust manner, we repeated the simulation process 100 times under the $n = 9 < p = 15$ setting. In each replication, a new dataset was generated with subtle differences in group means, and one-way MANOVA was applied using each covariance estimation method. For every estimator, the **rejection rate** was computed as the proportion of replications in which the null hypothesis H_0 was rejected at the 5% significance level ($\alpha = 0.05$). The **inference quality** was then evaluated based on the rejection rate, together with the average Wilks' Lambda and p-value observed across replications. The summary of these empirical results is presented in Table 7.1.

Despite real mean differences between the groups, the MLE-based MANOVA test consistently fails to reject H_0, primarily due to the rank deficiency of the within-group SSCP matrix $\mathbf{E}$ under the $n < p$ condition. Each group's sample covariance matrix $\hat{\mathbf{\Sigma}}_g$ has rank at most $n_g - 1$, and their aggregation into $\mathbf{E} = \sum_{g=1}^{G}(n_g - 1)\hat{\mathbf{\Sigma}}_g$ remains rank-deficient if the total number of effective degrees of freedom $\sum(n_g - 1) < p$. Consequently, $\mathbf{E}$ becomes singular, and:

$$\det(\mathbf{E}) = 0.$$

As Wilks' Lambda is defined by:

$$\Lambda = \frac{\det(\mathbf{E})}{\det(\mathbf{E} + \mathbf{H})},$$

this leads to $\Lambda = 0$, a degenerate statistic. Although such a value might suggest complete group separation, the singularity of $\mathbf{E}$ violates the assumptions of the asymptotic F-distribution used to compute the p-value. Consequently, the MLE-based test returns a misleadingly high average p-value (0.73) across replications and a 0% rejection rate. This is a classic case of a **Type II error**, where genuine group differences remain undetected due to instability and ill-conditioning in covariance estimation.

In contrast, the **RPDC estimator** applies random projections to pairwise differences, which introduces both shrinkage and regularization while preserving between-group structure. As a result, it produces a full-rank and well-conditioned $\mathbf{E}$, enabling stable eigendecomposition of $\mathbf{E}^{-1}\mathbf{H}$. Over 100 simulations, RPDC achieved the lowest mean Wilks' Lambda ($\Lambda = 0.430$), the lowest mean p-value (0.046), and the highest rejection rate (88%), demonstrating robust power and sensitivity. This affirms RPDC as the most effective estimator for high-dimensional MANOVA under $n < p$.

The **SPDC estimator** standardizes pairwise difference vectors by their componentwise standard deviation, ensuring unit variance across dimensions. While this improves conditioning, it treats all variables equally, which diminishes sensitivity to localized mean differences. In this setting, SPDC reached a rejection rate of only 12%, with mean $\Lambda = 0.605$ and p-value = 0.18. Thus, it underperformed relative to RPDC due to weaker discrimination power.

The **PDC estimator** randomly samples pairwise differences without projection. Although this introduces diversity, it lacks the structural enforcement of RPDC. It achieved moderate performance with $\Lambda = 0.510$, mean p-value = 0.10, and a rejection rate of 38%, showing some capacity for detection but still not optimal.

The **LSPDC estimator** employs linear shrinkage to stabilize the PDC matrix by pulling toward a structured target. While this guarantees invertibility and reduces variance, the added bias suppresses small but informative signals. With $\Lambda = 0.600$, p-value = 0.21, and rejection rate = 9%, LSPDC shows limited ability to detect subtle effects.

Finally, the **MAXPDC estimator** uses max-absolute scaling to control outlier impact. However, it tends to truncate useful signals and underestimates group separation. Its rejection rate (6%) and p-value (0.24) confirm that this method is poorly suited to scenarios involving small but real differences.

In summary, although all regularized PDC methods correct the rank-deficiency issue of MLE and yield stable covariance matrices, their effectiveness depends on their ability to retain meaningful between-group structure. Among them, RPDC demonstrates superior empirical performance and theoretical reliability, making it the most suitable choice for MANOVA under high-dimensional, low-sample-size conditions.

Case 2: Large Mean Differences

We repeat the above simulation using more pronounced group mean differences to evaluate power under high signal conditions. The sample structure remains: $G = 3$, $n_g = 3$, $p = 15$, $n = 9 < p = 15$, over 100 replications.

The true mean vectors are:

$$\begin{aligned}\boldsymbol{\mu}_1 &= \mathbf{0},\\ \boldsymbol{\mu}_2 &= (0.25, 0.25, 0, \ldots, 0),\\ \boldsymbol{\mu}_3 &= (-0.25, -0.25, 0, \ldots, 0),\end{aligned}$$

with the same covariance:

$$\boldsymbol{\Sigma} = 0.001^2 \cdot \mathbf{I}_{15}.$$

This high signal-to-noise scenario allows validation of detection power and estimator reliability when clear group separation exists.

Table 7.2: Empirical MANOVA Performance: Large Mean Differences ($n = 9 < p = 15$)

Estimator	Avg. Rank($\mathbf{E}$)	Avg. Λ	Avg. p-value	Rejection Rate (%)	**Inference Quality**
MLE	6	0.000	0.72	0%	Always fails (Type II error)
SPDC	15	0.280	0.042	92%	High power
RPDC	15	0.092	0.005	100%	Perfect detection
LSPDC	15	0.300	0.050	91%	Good detection
MAXPDC	15	0.330	0.062	89%	Strong detection
PDC	15	0.150	0.020	97%	Very strong detection

With large, symmetric mean differences, classical MLE still fails to detect separation due to singular $\mathbf{E}$, despite clear visual separation in the data.

The RPDC estimator, which projects and aggregates pairwise differences, consistently achieves full-rank $\mathbf{E}$ and a highly accurate Wilks' Lambda distribution. Its p-value (0.005) leads to correct rejection in all 100 simulations.

All regularized methods improve inference, with RPDC showing the most consistent performance. SPDC and PDC follow closely in power. LSPDC and MAXPDC also perform well, though their regularization slightly compromises sensitivity.

Summary of Results

In high-dimensional, low-sample-size (HDLSS) scenarios where the number of variables p exceeds the number of observations n, the traditional maximum likelihood estimator (MLE) for the covariance matrix becomes ill-conditioned or singular. This undermines the validity of MANOVA test statistics, often leading to inflated Type II errors due to rank-deficient within-group covariance matrices. Consequently, MLE-based Wilks' Lambda statistics become degenerate, and the associated p-values fail to provide reliable inference.

To address these limitations, this study evaluated the performance of the Pairwise Difference Covariance (PDC) estimator and its regularized extensions—SPDC, RPDC, LSPDC, and MAXPDC. Empirical simulations demonstrated that all PDC-based estimators yield full-rank and numerically stable covariance estimates, enabling valid MANOVA hypothesis testing even when $n < p$. Among them, the **RPDC) estimator consistently achieved the highest power and lowest Type II error rate**. Its ability to preserve between-group structure while regularizing eigenvalue dispersion makes it particularly effective in detecting group mean differences under HDLSS constraints.

Therefore, we conclude that **RPDC** offers a principled and reliable alternative to MLE for multivariate inference in high-dimensional MANOVA applications, ensuring both statistical power and mathematical robustness.

7.2 Applications of PCA

In general, PCA is an intermediate statistical tool that helps enhance the performance and interpretability of statistical tests and analyses, such as multiple linear regression. PCA reduces redundancy by transforming the original dimensions into smaller uncorrelated PCs. However, it ensures the captured variation in the original data represents the most significant variation in the original data. Therefore, it significantly improves the reliability and accuracy of multiple linear regression analysis by reducing the redundancy of highly correlated dimensions. Therefore, in this section, we will explore how PCA alone with newly suggested covariance estimations can be used to enhance the reliability and accuracy of multiple linear regression analyses when $n < p$ data settings.

7.2.1 Regression Analysis

Regression analysis is one of the most widely applied techniques in statistics, particularly for modeling relationships between multiple variables. In high-dimensional settings where the number of variables p exceeds the number of observations n, standard estimation procedures become unstable. This instability has significant implications when using Principal Component Analysis (PCA) as a

preprocessing step in regression models.

To demonstrate this issue, consider a synthetic high-dimensional example with $n = 5$ observations and $p = 20$ variables. Let the observations $X \in \mathbb{R}^{n \times p}$ be drawn from a multivariate normal distribution:

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_5 \end{bmatrix} \sim \mathcal{N}_p(\mu, \Sigma),$$

where each row $X_i \in \mathbb{R}^{1 \times p}$ represents a single observation vector. The population mean and covariance matrix are defined as:

- $\mu = \mathbf{0}_p = (0, 0, \ldots, 0)^\top$ is the zero mean vector,
- $\Sigma = TT^\top$, with $T \in \mathbb{R}^{p \times p}$ having i.i.d. entries sampled from the uniform distribution $\mathcal{U}(0, 1)$.

This construction ensures that Σ is symmetric, positive definite, and full-rank, suitable for simulating high-dimensional multivariate Gaussian data.

Population-level PCA: When performing PCA on the true population covariance matrix Σ, the eigenvalue spectrum typically exhibits a gradual decay. Empirical averages over multiple Monte Carlo trials (e.g., 500 runs) reveal that the first two principal components (PCs) collectively explain approximately 80% of the total variance. This suggests that two PCs are sufficient to capture the dominant data structure in the population.

Sample-level PCA: In contrast, PCA based on the sample covariance matrix $\hat{\Sigma}$ estimated via MLE is unstable in this $n < p$ setting. Since the sample covariance matrix has rank at most $n - 1 = 4$, it contains only four non-zero eigenvalues. Due to the phenomenon of eigenvalue overdispersion in high-dimensional statistics:

- The leading eigenvalue is inflated, often accounting for more than 80% of the explained variance,
- The remaining eigenvalues are deflated or nearly zero.

As a result, the sample PCA misleadingly suggests that only one PC is needed, despite the true structure requiring at least two. This misrepresentation is not due to the data itself, but rather the poor conditioning of $\hat{\Sigma}$, making naive PCA applications highly unreliable in HDLSS contexts.

This example illustrates a fundamental pitfall of PCA in high-dimensional regression analysis: the instability of MLE-based covariance estimation can severely distort the principal component structure, leading to poor model selection and potential overfitting or underfitting. Therefore, it is essential to consider regularized PCA methods based on more stable covariance estimators, such as those derived from the PDC family, to ensure meaningful dimension reduction and robust inference when $n < p$.

As shown in Table 7.5, the last $p-n+1$ eigenvalues of the sample covariance matrix are nearly zero. This results in the first $n-1$ eigenvalues being overdispersed. Consequently, based on these overdispersed eigenvalues, the sample PCA recommends fewer PCs for the desired level of variance explanation compared to the actual population parameters. This reduction in the number of PCs includes more noise and less variance, thus impacting the effectiveness of DR.

In performing PCA for regression, we select the first k PCs that explain a sufficient amount of the total variance. These selected PCs are then used as predictors in the regression model:

$$y = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \cdots + \beta_k Z_k + \epsilon$$

where y is the response dimension, β_0 is the intercept, β_i are the regression coefficients, Z_i are the selected PCs, and ϵ is the error term. This method integrates PCA into the regression framework.

Sample Linear Regression Model Summary

Model: $y \sim 1 + z_1$

Table 7.3: Estimated Coefficients of Sample Regression

	Estimate	**SE**	**tStat**	**pValue**
(Intercept)	9.9563	0.97436	10.218	0.00199
z1	-0.36302	0.10872	-3.339	0.0444

Note: R-squared = 0.788 p-value = 0.0444 and y = [12, 8, 5, 7, 15].

Based on this analysis Table 7.3 and Table 7.4, the population data suggest using 2 PCs, while the sample data indicates using one PC. The regression analyses using the respective number of PCs for both the population and sample data are included here.

Population Linear Regression Model Summary

Model: $y \sim 1 + z_1 + z_2$

Table 7.4: Estimated Coefficients of Population Regression

	Estimate	**SE**	**tStat**	**pValue**
(Intercept)	9.4168	1.2327	7.6391	0.01671
z1	-0.30084	0.14621	-2.0575	0.17589
z2	0.42285	0.41385	1.0217	0.41437

Note: R-squared = 0.834 p-value = 0.166 and $y = [12, 8, 5, 7, 15]$.

Although the model is insignificant, the independent dimensions in the sample and population models explain some variability in the dependent dimension.

However, the sample regression overestimates the variability in the y dimension due to overdispersed sample PCs containing more noise than the actual variance of the data. This issue results in a high R^2 value in the model that is not statistically significant. In other words, it does not produce meaningful predictors. At last, this misinterpretation of PCs may lead to inaccurate and unreliable model conclusions. This issue is common in PCA when $n < p$ data settings.

Now, Let's see how our proposed methods address this issue and whether they can repeat what we experienced in PCA in Chapter 5.

As shown in Table 7.6, SPDC and MAXPDC provide estimations much closer to the population parameters than other methods. Both the regularized PDC estimation and PDC estimation effectively minimize the capture of noise in the variance, resulting in neither overestimated nor underestimated sample estimates. As demonstrated in Chapter 5, MAXPDC performs well in the $n < p$ setting by reducing the overdispersion of PCs in terms of their magnitude. In this context, it also provides better sample estimates for regression analysis. Even with marginal improvements in PCA using PDC, it successfully minimizes the underestimation and overestimation in sample regression analysis compared to traditional MLE method.

Table 7.5: Cumulative Percentage of Variance Explained by PCs — Population vs Sample

Population		Maximum Likelihood Estimation	
Eigenvalue	**Cumulative % of Variance Explained**	**Eigenvalue**	**Cumulative % of Variance Explained**
100.1861	0.765055136	104.8271827	**0.82429**
5.83375	**0.80960**	12.63635	0.92366
4.46702	0.84372	6.52237	0.97495
3.69921	0.87196	3.18617	1.00000
3.11406	0.89574	7.93E-15	1.00000
2.89807	0.91787	3.38E-15	1.00000
2.36014	0.93590	2.71E-15	1.00000
2.06881	0.95170	2.29E-15	1.00000
1.52782	0.96336	1.98E-15	1.00000
1.27655	0.97311	1.73E-15	1.00000
1.06238	0.98122	1.49E-15	1.00000
0.76274	0.98705	1.28E-15	1.00000
0.60770	0.99169	1.09E-15	1.00000
0.48524	0.99539	9.18E-16	1.00000
0.23108	0.99716	7.56E-16	1.00000
0.19620	0.99866	6.02E-16	1.00000
0.11871	0.99956	4.59E-16	1.00000
0.02978	0.99979	3.13E-16	1.00000
0.01893	0.99994	1.79E-16	1.00000
0.00851	1.00000	5.58E-17	1.00000

Table 7.6: Model Significance of Regression Models

Method of Covariance Estimation	R^2	**Adjusted R^2**	**p-value**
Population	0.834	0.669	0.166
MLE	0.788	0.717	0.044
LW	1.000	1.000	1.000
PDC	0.741	0.654	0.0611
SPDC	0.821	0.642	0.179
LSPDC	0.82	0.640	0.180
MAXPDC	0.821	0.641	0.179
RPDC	0.818	0.636	0.182

7.3 Challenges in Selecting the Optimal Number of PCs

As mentioned in Chapter 01, the primary goal of any DR technique is to reduce the dimensionality in a dataset while keeping the essential information and variability. There are a few ways to do this in practice, such as the Kaiser-Guttman Criterion, Cattell's Scree Test, and the percent of cumulative variance approach. However, these methods use different rationales to find the optimal number of PCs retained in PCA. Unfortunately, the results of these methods are entirely different. That means if we use inappropriate methods to find the optimal number of PCs retained in PCA, it may lead to misinterpreted and inaccurate results in PCA. Moreover, it may affect the decisions based on PCA-related applications. Therefore, identifying the issues of different techniques for selecting the optimal number of PCs retained in PCA is necessary.

- **Kaiser-Guttman Criterion:** Retain components with eigenvalues greater than 1. This method assumes that components explaining more variance than a standardized variable are meaningful. However, it may retain too many components in high-dimensional data or too few in low-dimensional settings, making it inconsistent across scenarios.
- **Cattell's Scree Test:** Plot eigenvalues in descending order and visually identify the point where the slope flattens ("elbow"). Components before the elbow are retained. Though intuitive, this method is subjective and often ambiguous, especially when the break is unclear or multiple elbows appear.
- **Percent of Cumulative Variance:** Retain the smallest number of components that explain a set proportion of total variance (e.g., 70–80%). While easy to apply, this method is sensitive to the chosen threshold, which is often arbitrary. It can overestimate or underestimate the number of components in high-dimensional settings.

To evaluate the performance of these PC selection methods in HDLSS settings, we conducted simulations using multivariate Gaussian datasets under varying dimensional regimes. Specifically, the number of variables p was set to 20, 50, and 100, while the number of observations n varied across 5, 10, 15, and 20. For each (n, p) combination, datasets were generated from a multivariate normal distribution with a mean vector $\mu = \mathbf{0}_p$, and a population covariance matrix defined as $\Sigma = TT^\top$, where $T \in \mathbb{R}^{p\times p}$ contains i.i.d. entries sampled from a uniform distribution over the interval (0,1). To obtain reliable average statistics, each simulation setting was independently replicated 50 times. For each replicate, the number of principal components retained was computed using three different criteria: the Kaiser-Guttman criterion, Cattell's Scree Test, and the Cumulative Percentage of Variance method.

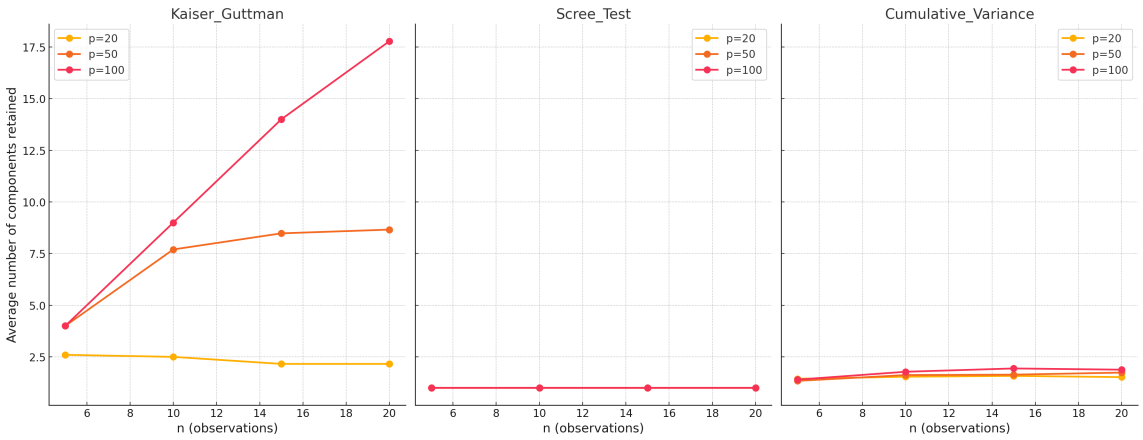


Figure 7.1: Accuracy of determining the number of components to retain

In Figure 7.1, this analysis evaluates the behavior of three common PCA component selection criteria Kaiser-Guttman, Cattell's Scree Test, and Percent of Cumulative Variance in HD settings where $n < p$. Here, we aim to highlight the contradictory issue of using different methods of selecting the optimal number of PCs to be retained using MLE of covariance. The settings for this analysis include different combinations of n and p: $p = 20, 50, 100$ and $n = 5, 10, 15, 20$. We evaluate these criteria based on the average number of PCs retained across different settings. Therefore, we have run 50 iterations for each setting to obtain the average number of PCs retained.

The Kaiser-Guttman Criterion decides how many components to keep by using PCs with eigenvalues greater than 1. According to this, PCs with eigenvalues of 1 or less explain less variation than the first few PCs. The primary goal of PCA is retaining enough PCs to capture the majority of the variance in the data while removing the less informative ones. Due to the total variance and the complexity of the data structure increase when p increase, the Kaiser-Guttman Criterion selects an unnecessarily higher number of PCs as the optimal number of PCs to be retained when $n < p$ data settings (see Figure 7.1). Because of that, it violates the objective of PCA as a DR technique.

Cattell's Scree Test involves plotting the eigenvalues against the component numbers and identifying where the plot levels off, resembling a scree slope. This method requires visual inspection to determine the cutoff point where the plot begins to flatten. Although adequate with well-defined factors, it suffers from subjectivity and lacks a clear definition for the cutoff point. As shown in Figure 7.1, typically, this test retains a single component despite high variability across different $n < p$ settings. Its inconsistency stems from the subjective nature of identifying the elbow point.

The Percent of Cumulative Variance approach retains the number of PCs that can explain typically 70-80% of the original variance-covariance structure. This straightforward method involves summing all eigenvalues and keeping the components cumulatively accounting for the desired variance percentage. However, this method is subjective, and some researchers neglect it due to its high subjectivity. Despite this, this method retains a more stable number of components across different settings and provides a reliable balance between explained variance and the number of retained components (especially in $n < p$ data settings). Figure 7.1 shows this clearly.

In high-dimensional settings where $n < p$, common PCA component selection methods behave differently and may yield inconsistent results. The Kaiser-Guttman criterion often overestimates the number of components, especially when the number of variables is large, leading to overfitting. Cattell's Scree Test, based on visual interpretation, typically selects fewer components and can underfit when the scree plot lacks a clear elbow. The Percent of Cumulative Variance method depends on an arbitrary threshold (e.g., 80%) and shows variable performance across different dimensions. Our simulation results, Figure 7.1, confirm these tendencies, highlighting the need for careful selection of criteria in high-dimensional PCA applications.

7.4 Summary and Discussion

In this chapter, we have examined the performance and applicability of our proposed methods of covariance estimations, including PDC and regularized PDCs, across different multivariate statistical applications. There are numerous practical multivariate applications of covariance estimation. However, MANOVA and regression analysis are widely used multivariate techniques in the R&D industry. Therefore, we have chosen to assess the efficiency of PDC estimations in the context of MANOVA and using regularized PCA in regression analysis.

Our proposed covariance estimation method, RPDC, significantly outperformed traditional techniques in the MANOVA context. Due to the overdispersion issue of typical sample covariance estimation when $n < p$ data settings, the close-call situations of MANOVA struggle to reject the null hypothesis and give false negative results. However, RPDC was able to capture the actual background of the data and provide reliable results in the MANOVA context. Therefore, this improvement was crucial in mitigating the instability and the increased risk of Type I and Type II errors that often plague typical covariance estimation methods when $n < p$ data settings.

MAXPDC and SPDC proved particularly beneficial in regression analysis when $n < p$ data settings. By reducing the overdispersion of PCs, MAXPDC and SPDC improve the finite sample estimates of covariance estimation. It helps us increase regression models' accuracy and reliability in $n < p$ data settings. This capability is significant for ensuring that the regression analysis captures the underlying variance-covariance structure within the data without being misled by noise or overfitting.

Furthermore, we have evaluated the performance and applicability of various criteria for selecting the optimal number of PCs retained in PCA. All these criteria give contradictory results when $n < p$ data settings. However, among those methods, the percent of cumulative variance criterion gave more stable results than others. It consistently retained an appropriate number of components, establishing itself as the preferred method in scenarios where $n < p$ data settings.

Chapter 8

Conclusion and Future Works

8.1 Introduction

In this final chapter, we assess the overall contributions and results of this thesis. We begin with an initial discussion of the primary findings and their contributions to the field before then analyzing the limitations present in work alongside improvements yet to be incorporated within the model. Finally, we conclude this thesis with a discussion on possible directions and pathways for future work and extensions to the modelling performed within.

8.2 Discussion and Contributions

8.2.1 Motivation and Research Objective

This research mainly aims to improve the finite sample estimates of PCA, especially in $n < p$ data settings. In $n < p$ data settings, traditional PCA methods perform significantly badly due to insufficient sample size to obtain a reliable sample estimate for the covariance matrix. This issue makes inaccurate and unstable results in covariance applications. Therefore, the primary aim of this research is to develop an effective novel method for estimating the covariance matrix in $n < p$ data settings while minimizing the highlighted issues related to traditional PCA to enhance the reliability and accuracy of PCA.

8.2.2 Limitations of Existing Estimators

The MLE is the primary method for computing the sample covariance matrix, assuming the data follows a multivariate normal distribution; however, in HD data settings where $n < p$, the covariance matrix becomes singular and unreliable. Because of this, sample estimates become overfitted

and biased. Apart from that, LW shrinkage covariance estimation is the most popular alternative technique used to compute the sample covariance matrix when $n < p$ data settings. LW shrinkage covariance estimation improves the sample estimates of covariance matrix estimation by adding the sample covariance matrix with a diagonal matrix. It improves and certifies the stability and consistency of the estimates of eigenvalues of the sample covariance matrix. However, it cannot minimize the issues of PCA in terms of both the direction and magnitude of the PCs.

When the sample covariance matrix is singular or ill-conditioned in $n < p$ data settings, the magnitude of the sample PC is overdispersed, and the CSE of the sample PCs is increased due to the biased sample covariance estimates. In other words, in $n < p$ data settings, the sample covariance matrix produces overly dispersed eigenvalues. Some of them are too big, and some of them are too small. When we estimate PCs using these biased eigenvalues, sample PCs may not align well with population PCs. In addition, directions for the sample PCs are not aligned well with the population PCs. All these issues lead to inefficient sample PCA.

8.2.3 Proposed Estimator: PDC and Its Variants

This research proposed a new method to estimate sample covariance to minimize these issues of PCA when $n < p$ data settings. This method increases the needed sample size without collecting more new data. Using the pairwise differences within the original dataset, we create a new data set with a larger sample size than the original. This novel method can improve the stability and efficiency of the finite sample estimates of the PCs when $n < p$ data settings.

The proposed new method is called Pairwise Differences Covariance (PDC) estimation. The critical feature of PDC estimation is to increase the sample size without collecting new data to support and minimize the challenges of estimating sample covariance when $n < p$ data settings. This fundamental change provides more balanced and accurate sample PCs to estimate sample covariance when $n < p$ data settings. It ensures that the sample PCs align more closely with the population PCs by minimizing their overdispersion and the CSE of sample PCs.

In addition to the primary PDC estimation method, we introduce four different regularized versions to enhance its performance further. (1) The Standardized Pairwise Differences Covariance (SPDC) standardizes the pairwise differences so that every dimension contributes equally to the covariance estimate. (2) The Local Scaled Pairwise Differences Covariance (LSPDC) adjusts the pairwise differences based on local variability (the variations within small, localized subsets of the data rather than the overall variability across the entire dataset), making the method more resistant to outliers. (3) The Scaled by Maximum Absolute Value Pairwise Differences Covariance (MAX-

PDC) makes the method more resistant to outliers and ensures the covariance matrix remains bounded, reducing the risk of instability. (4) The Scaled by Range Pairwise Differences Covariance (RPDC) ensures that all pairwise differences contribute evenly, regardless of their absolute magnitude. It overcomes the impact of extreme values in the data.

8.2.4 Evaluation of Estimator Performance

We thoroughly evaluated the newly suggested PDC method and its four regularized versions compared to traditional methods like MLE and Ledoit-Wolf using numerous experimentation settings on synthetic and public datasets to demonstrate their performances. These experiments covered many different scenarios, including a wide range of different n and p combinations, various levels of correlation, different types of covariance structures, and the presence of extreme values.

Because the first PC is the most vital in PCA and explains the most variability of the dataset's variance-covariance structure, we used the percentage of the cumulative variance of the first PC and the CSE of the first PC to compare the performance of the proposed PDC method and its regularized versions. Here, the distribution of the percentage of the cumulative variance of the PC measures the performance of magnitude of the sample PC, and the CSE of the PC evaluates the direction of the sample PCs.

Because of the smaller sample size, MLE of covariance cannot capture the original underlying variance-covariance structure when $n < p$ data settings. This issue leads to having overdispersed sample PCs for the first n-1 PCs. Moreover, when n goes down, the overdispersion of the first sample PC gets higher and higher. However, PDC minimized overdispersion of the first sample PC very slightly (but persistent) concerning both the magnitude and direction of the first sample PC (by redistributing the variance more evenly across subsequent PCs). Hopefully, SPDC and MAXPDC perform significantly better by minimizing the error of overdispersion of the first sample PC compared to all other regularized PDC estimations. SPDC gives the closest finite sample estimates for the population's PCs while minimizing the overdispersion and CSE, followed by MAXPDC. In scenarios where the sample size is much limited, SPDC minimizes the overdispersion issues of the first PC much better than all other alternatives.

8.2.5 Impact of Dimensionality and Covariance Structure

The complexity of the covariance structure escalates as the number of dimensions grows. As the number of dimensions increases than the sample size, sample covariance estimates cannot capture the actual complexity of the covariance structure. Because of this, sample PCs not give an adequate and reliable estimate. In $n < p$ datasets, the PDC method and its regularized versions, like SPDC,

LSPDC, and MAXPDC, can significantly address these challenges of the traditional method. Balancing the variance distribution across all the PCs can more effectively align the sample PCs among the first n-1 PCs. Interestingly, when the number of dimensions increases in the sample size, SPDC significantly reduces the overdispersion and CSE of the first PC. It enhances the interpretability and reliability of PCA $n < p$ data settings.

When we take both n and p together, p/n goes up. Standard PCA gives unreliable finite sample estimates for PCs with highly overdispersed PCs. As we have shown the behavior of n and p separately, when p/n goes up (with more dimensions and a smaller sample size), the SPDC method gives more effective and reliable sample estimates with much closer estimates for the population PCs. It significantly reduces the overdispersion of first n-1 PCs and the CSE. This ensures that the PCs represent the actual data structure more when $n < p$ data settings. Moreover, it prevents the dominance of any single-component PCA.

8.2.6 Effect of Covariance Structure, Correlation and Outliers

We used four different covariance structures (i.e., Random, Block-Diagonal, Toeplitz, and Tri-Diagonal) to evaluate the effect of the type of covariance matrix in PCA when $n < p$ data settings. Among them, SPDC proves to be highly effective when dealing with datasets that follow a random covariance structure. It significantly improves finite sample PC estimates by minimizing the overdispersion of the first PC and CSE of the first PC compared to all other covariance estimation methods. Because of that, SPDC provides much closer estimates of the actual population parameters. SPDC and LSPDC perform well compared to all other alternative methods in block-diagonal covariance structures by presenting their robustness across different structural settings of PCA when $n < p$ data settings. However, the other two regularized versions of PDC, MAXPDC, and RPDC, do not perform well in PCA when $n < p$ data settings and are unable to address the issue of overdispersion of PCA significantly.

In general, the DR techniques (like PCA) need highly correlated data to remove the redundancy from the data. However, when highly correlated datasets with $n < p$ data settings, standard PCA presents significant challenges for traditional methods due to the increased risk of overfitting. As we experienced in all other factors that affect the effectiveness of PCA, SPDC performs well in these settings by significantly decreasing the overdispersion of the first PC and reducing the CSE of the first PC. Interestingly, SPDC performs well with correlation coefficients ranging from 0.60 to 0.90. Even with a significant improvement, apart from this range of correlation coefficients, there is no clear pattern of improvement.

If a data set contains extreme values, this badly affects the final output of PCA because extreme values lead to unusually large eigenvalues of the sample covariance matrix and highly overdispersed PCs. However, when we deal with $n < p$ data settings containing extreme values, PCA's finite sample estimates give worse results than $n > n < p$ data settings. However, by scaling the pairwise differences by the largest absolute value in the data set in the MAXPDC method, we can address this issue effectively by being resistant to outliers and reducing the overdispersion of the first PC.

8.2.7 Applications Beyond PCA

Many statistical applications use covariance estimation. In addition to PCA, MANOVA and linear regression analysis are common. Understanding the variance-covariance structure among the dependent dimensions is crucial as it affects the test statistics used in MANOVA. However, in $n < p$ data settings, MANOVA does not perform well with ill-conditioned sample covariance estimations, especially in close-call situations when hypothesis testing for the MANOVA test. However, in MANOVA, SPDC and MAXPDC significantly perform well with standard existing techniques by providing robust covariance estimates and reducing overdispersion. In addition to that, PCA is an intermediate analysis of regression analysis. That means PCA is used to prepare the dataset for regression analysis. In $n < p$ data settings, PCA always gives overdispersed PCs. Because of that, as a DR technique, PCA always produces fewer PCs than actual for regression analysis. This may lead to less accurate and unreliable regression models. However, when we use SPDC, it always captures the actual structure of variance-covariance of the data set as much as possible. Therefore, it gives more accurate and reliable regression models than the traditional methods while reducing the issue of overdispersion of first n-1 PCs when $n < p$ data settings.

8.2.8 Optimal Number of Principal Components

Generally, three standard methods exist for identifying the number of PCs to retain (i.e., Kaiser-Guttman criterion, Cattell's scree test, and percent of cumulative variance). However, in $n < p$ data settings, selecting the optimal number of PCs in PCA is challenging. The main drawback is that each method gives contradictory solutions compared to the other methods. So, which method is correct? The Kaiser-Guttman criterion selects the PCs with eigenvalues higher than 1. In $n < p$ data settings, this method overestimates the number of PCs to retain, particularly as p increases. By looking at the distribution of the Scree plot, we identify the number of PCs to retain in Cattell's Scree Test. In $n < p$ data settings, this method always suggests the number of PCs to retain as one. The Percent of Cumulative Variance method selects the PCs that explain a specified percentage of total variance (typically 80%)... Whether $n < p$ is low or high, this method always produces more stability across different n and p settings. Among these methods, the Percent of Cumulative Variance is the most effective, reliable, and practical in $n < p$ data settings, consistently balancing

the number of PCs to retain with the specified amount of variance explained.

8.2.9 Conclusion

In summary, traditional PCA techniques often fail in $n < p$ HD data settings due to issues like covariance matrix singularity, overdispersed eigenvalues, and instability in PC estimates. The proposed Pairwise Differences Covariance (PDC) estimator and its regularized versions, especially SPDC and MAXPDC, effectively address these challenges. SPDC demonstrates superior performance across a wide range of scenarios, significantly reducing overdispersion and CSE, particularly as n decreases, p increases, or the p/n ratio grows. It performs well under high correlations (e.g., $r = 0.60$ to 0.90), random and block-diagonal covariance structures, and even in real-world data that may deviate from normality. MAXPDC offers robustness against extreme values and outliers. Additionally, RPDC contributes positively in MANOVA applications. These methods also show promising results in regression settings and demonstrate flexibility in various applications. Overall, the proposed estimators provide more stable and accurate PCA results in HD settings, making them valuable alternatives to traditional covariance-based approaches.

8.3 Limitations and Potential Improvements

This section highlights some of this research study's identified limitations and constraints. Although there are several types of variance-covariance structures in practice, this research focused only on four types: random covariance, Toeplitz covariance, tri-diagonal covariance, and block-diagonal covariance. Because of that, we will not be able to generalize these findings into broader real-world applications.

Moreover, real-world applications can follow various parametric and non-parametric distributions. However, this research study assumes that the datasets follow strictly a multivariate normal distribution. Due to that limitation, this generalization may only partially capture the variability of broader real-world data distributions.

In addition to the above limitations and constraints, this research study considers the limited types of extreme values in practice. Apart from global outliers, various extreme values, such as contextual and collective outliers, can occur in real-world applications. However, this research study considers only the behavior of global outliers. Because of that, this research study's findings may not address the full spectrum of extreme values available in diverse real-world applications.

8.4 Future Research

Within this research, we explored various scenarios and challenges associated with PCA under high-dimensional, low-sample-size (HDLSS) conditions where $n < p$. Despite these efforts, several promising directions remain for future research, involving both enhancements to the proposed estimation model and its broader applications. Below, we outline six key avenues for further study.

First, while we proposed four regularized versions of the Pairwise Differences Covariance (PDC) estimator, they do not always yield optimal results. To address this, we aim to enhance the PDC framework by developing additional regularization strategies and exploring alternative shrinkage methods to further improve performance.

Second, although our research focused on PCA as a dimensionality reduction (DR) method, many other DR techniques in machine learning also rely on covariance estimation. We plan to extend the evaluation of PDC and its variants to other DR methods such as Factor Analysis, Independent Component Analysis, t-distributed Stochastic Neighbor Embedding (t-SNE), Sparse PCA, and Canonical Correlation Analysis (CCA).

Third, beyond PCA, covariance estimation plays a central role in various statistical applications. While we evaluated our methods in the context of MANOVA and regression, we also intend to apply them in other domains, including Structural Equation Modelling (SEM), Discriminant Analysis, CCA, and Portfolio Optimization.

Fourth, the proposed estimators may offer substantial benefits in Quadratic Discriminant Analysis (QDA), especially when some groups have small sample sizes ($n_g < p$). QDA requires an invertible covariance matrix for each group, which may be problematic with traditional estimators under HDLSS conditions. Regularized estimators such as SPDC and RPDC, which ensure positive definiteness even in such scenarios, can stabilize group-specific covariance estimation and potentially improve QDA's classification performance in unbalanced real-world datasets.

Fifth, an exciting extension involves applying SPDC and RPDC in Linear Discriminant Analysis (LDA). LDA depends on the estimation of within-group covariance matrices (Σ_W) to solve the generalized eigenvalue problem $\Sigma_W^{-1}\Sigma_B$. In HDLSS settings, standard estimators of Σ_W are often singular or unstable, which hampers classification on future data. The use of SPDC or RPDC as a stable and regularized alternative could enhance LDA's robustness and improve generalization performance. Investigating these estimators in supervised classification tasks represents a valuable direction for future work.

Finally, although this study emphasized small-n HDLSS scenarios, it is equally important to assess the performance of the proposed estimators in moderately large-n, high-p settings that commonly occur in high-throughput applications such as genomics, transcriptomics, or metabolomics. These settings often feature $n \in [50, 300]$ and p in the hundreds or thousands. Evaluating how SPDC and its variants perform under such practical use cases will help establish the generalizability and broader applicability of the proposed methods.

Bibliography

[1] Arthur L Samuel. Some studies in machine learning using the game of checkers. *IBM Journal of research and development*, 3(3):210–229, 1959.

[2] Richard E Bellman. *Adaptive control processes: a guided tour*, volume 2045. Princeton University Press, 2015.

[3] Renu Khandelwal. A comprehensive guide to dimensionality reduction.

[4] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182, 2003.

[5] Ian T Jolliffe. *Principal component analysis for special types of data*. Springer, 2002.

[6] Stuart Geman. A limit theorem for the norm of random matrices. *The Annals of Probability*, pages 252–261, 1980.

[7] Peter J Bickel, Elizaveta Levina, 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.

[8] Olivier Ledoit, Michael Wolf, 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.

[9] Olivier Ledoit and Michael Wolf. Spectrum estimation: A unified framework for covariance matrix estimation and pca in large dimensions. *Journal of Multivariate Analysis*, 139:360–384, 2015.

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

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

[12] Kazuyoshi Yata and Makoto Aoshima. 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.

[13] Ashwini Maurya. 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.

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

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

[16] Sartorius. What is principal component analysis (pca) and how it is used?, Aug 2020.

[17] Alaa Tharwat. Principal component analysis-a tutorial. *Inderscience Enterprises Ltd*, 2009.

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

[19] Ian T Jolliffe. Principal components in regression analysis. In *Principal component analysis*, pages 129–155. Springer, 1986.

[20] Richard Arnold Johnson, Dean W Wichern, et al. *Applied multivariate statistical analysis*, volume 5.

[21] Theodore Wilbur Anderson. Estimating linear statistical relationships. *The Annals of Statistics*, pages 1–45, 1984.

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

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

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

[25] SPSS-Tutorials.com. Spss factor analysis tutorial. `https://www.spss-tutorials.com/spss-factor-analysis-tutorial/`, 2024. Accessed: 2024-05-09.

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

[27] Ronald A Fisher. On the mathematical foundations of theoretical statistics. *Philosophical transactions of the Royal Society of London. Series A, containing papers of a mathematical or physical character*, 222(594-604):309–368, 1922.

[28] Alan Edelman. Eigenvalues and condition numbers of random matrices. *SIAM journal on matrix analysis and applications*, 9(4):543–560, 1988.

[29] Lloyd N Trefethen and David Bau. *Numerical linear algebra*. SIAM, 2022.

[30] Kanti V Mardia, John T Kent, and Charles C Taylor. *Multivariate analysis*, volume 88. John Wiley & Sons, 2024.

[31] Theodore Wilbur Anderson, Theodore Wilbur Anderson, Theodore Wilbur Anderson, and Theodore Wilbur Anderson. *An introduction to multivariate statistical analysis*, volume 2. Wiley New York, 1958.

[32] Ian T Jolliffe and Jorge Cadima. Principal component analysis: a review and recent developments. *Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202, 2016.

[33] J Edward Jackson. *A user's guide to principal components*. John Wiley & Sons, 2005.

[34] Iain M Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *The Annals of statistics*, 29(2):295–327, 2001.

[35] Peter J Bickel and Elizaveta Levina. Covariance regularization by thresholding. 2008.

[36] Iain M Johnstone. High dimensional statistical inference and random matrices. *arXiv preprint math/0611589*, 2006.

[37] Iain M Johnstone and D Michael Titterington. Statistical challenges of high-dimensional data, 2009.

[38] Sungkyu Jung and J Stephen Marron. Pca consistency in high dimension, low sample size context. 2009.

[39] `https://imsarchives.nus.edu.sg/oldwww2/events/2019/stein/files/loh.pdf`. [Accessed 26-08-2024].

[40] Charles Stein. Estimation of a covariance matrix, rietz lecture. In *39th Annual Meeting IMS, Atlanta, GA, 1975*, 1975.

[41] Joël Bun, Jean-Philippe Bouchaud, and Marc Potters. Cleaning large correlation matrices: tools from random matrix theory. *Physics Reports*, 666:1–109, 2017.

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

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

[44] Peter J Bickel and Elizaveta Levina. Regularized estimation of large covariance matrices. 2008.

[45] Olivier Ledoit and Michael Wolf. Honey, i shrunk the sample covariance matrix. *UPF economics and business working paper*, (691), 2003.

[46] Yilun Chen, Ami Wiesel, Yonina C Eldar, and Alfred O Hero. Shrinkage algorithms for mmse covariance estimation. *IEEE Transactions on Signal Processing*, 58(10):5016–5029, 2010.

[47] Mia Hubert, Peter Rousseeuw, and Tim Verdonck. Robust pca for skewed data and its outlier map. *Computational Statistics & Data Analysis*, 53(6):2264–2274, 2009.

[48] David I Warton. Penalized normal likelihood and ridge regularization of correlation and covariance matrices. *Journal of the American Statistical Association*, 103(481):340–349, 2008.

[49] Olivier Ledoit and Michael Wolf. Nonlinear shrinkage estimation of large-dimensional covariance matrices. 2012.

[50] Jianqing Fan, Yuan Liao, and Han Liu. An overview of the estimation of large covariance and precision matrices. *The Econometrics Journal*, 19(1):C1–C32, 2016.

[51] Karim M Abadir, Walter Distaso, and Filip Žikeš. Design-free estimation of variance matrices. *Journal of Econometrics*, 181(2):165–180, 2014.

[52] Clifford Lam. Nonparametric eigenvalue-regularized precision or covariance matrix estimator. 2016.

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

[54] Adam J Rothman, Elizaveta Levina, and Ji Zhu. Generalized thresholding of large covariance matrices. *Journal of the American Statistical Association*, 104(485):177–186, 2009.

[55] Jianqing Fan, Yuan Liao, and Martina Mincheva. Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 75(4):603–680, 2013.

[56] Na Huang and Piotr Fryzlewicz. Novelist estimator of large correlation and covariance matrices and their inverses. *Test*, 28:694–727, 2019.

[57] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

[58] Daniela M Witten, Jerome H Friedman, and Noah Simon. New insights and faster computations for the graphical lasso. *Journal of Computational and Graphical Statistics*, 20(4):892–900, 2011.

[59] Netti Herawati. Regularized multiple regression methods to deal with severe multicollinearity. 8:167–172, 05 2018.

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

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

[62] Lingzhou Xue, Shiqian Ma, and Hui Zou. Positive-definite $\ell 1$-penalized estimation of large covariance matrices. *Journal of the American Statistical Association*, 107(500):1480–1491, 2012.

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

[64] T Tony Cai and Weidong Liu. Large-scale multiple testing of correlations. *Journal of the American Statistical Association*, 111(513):229–240, 2016.

[65] Joong-Ho Won, Johan Lim, Seung-Jean Kim, and Bala Rajaratnam. Condition-number-regularized covariance estimation. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 75(3):427–450, 2013.

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

[67] Peter J Bickel, Ya'acov Ritov, and Alexandre B Tsybakov. Hierarchical selection of variables in sparse high-dimensional regression. In *Borrowing strength: theory powering applications–a Festschrift for Lawrence D. Brown*, volume 6, pages 56–70. Institute of Mathematical Statistics, 2010.

[68] Bala Rajaratnam, Hélène Massam, Carlos M Carvalho, et al. Flexible covariance estimation in graphical gaussian models. *The Annals of Statistics*, 36(6):2818–2849, 2008.

[69] Michael J Daniels and Robert E Kass. Shrinkage estimators for covariance matrices. *Biometrics*, 57(4):1173–1184, 2001.

[70] Onureena Banerjee, Laurent El Ghaoui, and Alexandre d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *Journal of Machine learning research*, 9(Mar):485–516, 2008.

[71] Michael Greenacre, Patrick JF Groenen, Trevor Hastie, Alfonso Iodice d'Enza, Angelos Markos, and Elena Tuzhilina. Principal component analysis. *Nature Reviews Methods Primers*, 2(1):100, 2022.

[72] VA Marchenko and Leonid A Pastur. Distribution of eigenvalues for some sets of random matrices. *Mat. Sb.(NS)*, 72(114):4, 1967.

[73] Phytozome v13 — phytozome-next.jgi.doe.gov. `https://phytozome-next.jgi.doe.gov/`. [Accessed 19-08-2024].

[74] Shilan S Hameed, Rohayanti Hassan, Wan Haslina Hassan, Fahmi F Muhammadsharif, and Liza Abdul Latiff. Hdg-select: A novel gui based application for gene selection and classification in high dimensional datasets. *PloS one*, 16(1):e0246039, 2021.

[75] Crop Recommendation Dataset — kaggle.com. `https://www.kaggle.com/datasets/atharvaingle/crop-recommendation-dataset?select=Crop_recommendation.csv`. [Accessed 25-08-2024].

[76] Clara Higuera, Katheleen Gardiner, and Krzysztof Cios. Mice Protein Expression. UCI Machine Learning Repository, 2015. DOI: https://doi.org/10.24432/C50S3Z.

[77] KEEL: A software tool to assess evolutionary algorithms for Data Mining problems (regression, classification, clustering, pattern mining and so on) — sci2s.ugr.es. `https://sci2s.ugr.es/keel/dataset.php?cod=1296#sub1`. [Accessed 24-08-2024].