AIMS | African Institute for Mathematical Sciences CAMEROON

Comparison of Principal Component Analysis (PCA), Sparse Principal Component Analysis (SPCA), and Exploratory Factor Analysis (EFA) in Social Science Applications

Elizabeth Mukusya Mwania (elizabeth.mwania@aims-cameroon.org)
African Institute for Mathematical Sciences (AIMS)
Cameroon

Supervised by: Professor Gabriel Wallin
Lancaster University, UK

28 May 2025

Submitted in Partial Fulfillment of a Coopeative Masters Degree at AIMS-Cameroon

Abstract

With the era of big data, dimension reduction methods are essential, especially in high-dimensional datasets where the number of variables largely exceeds the number of observations. Three of the most popular multivariate statistical techniques used are principal component analysis (PCA), sparse principal component analysis (SPCA), and exploratory factor analysis (EFA). Although these methods aim to represent a data matrix into a lower-dimension space, they differ in mathematical foundations, assumptions, and interpretability.

PCA, which is probably the most commonly used technique, reduces a dataset with numerous variables into a sequence of uncorrelated components. SPCA introduces sparsity constraints which produce components with many zero loadings, improving interpretability. EFA is predominantly used in social sciences to identify underlying relationships between observed variables and the latent constructs within data structure. Although the underlying theory behind these methods differ, there is lack of numerical investigation to show how these methods perform dimension reduction under different data conditions.

In this research, we conduct a simulation study to identify scenarios where each of the three methods perform well or poorly. This will help providing systematical guidelines to practitioners for selecting appropriate methods. An empirical application to social science data is also conducted to demonstrate a practical use of each method, with a particular focus on introducing SPCA which is relatively underused yet a valuable tool for social scientists. The guidelines provided in this study will help social scientists select suitable methods based on data characteristics and research objectives. Moreover, this work equips practitioners with the knowledge to make informed methodological choices, bridging the gap between theoretical innovation and applied research, advancing the use of modern dimension-reduction techniques in social science.

Keywords: high-dimensional data, principal component analysis, sparse principal component analysis, exploratory factor analysis.

Declaration

I, the undersigned, hereby declare that the work contained in this essay is my original work, and that any work done by others or by myself previously has been acknowledged and referenced accordingly.

Elizabeth Mukusya Mwania, 28 May 2025.

Contents

1. Introduction

In modern research, high-dimensional datasets are increasingly available across all fields to support data-driven inquiry. Dimensionality reduction methods play a critical role in enabling researchers to simplify these datasets and enhance interpretability while retaining essential information. In the social sciences, high-dimensional data arise from sources such as surveys, administrative records and digital platforms. The growing ease and decreasing cost of data collection is a positive development, but it also presents new challenges for analysis. This creates a pressing need for accessible and efficient methods to select appropriate dimensionality reduction techniques.

Principal component analysis (PCA), sparse principal component analysis (SPCA), and exploratory factor analysis (EFA) are commonly used methods for dimension reduction. However, the comparative evaluation of their effectiveness remains underexplored, particularly in applied social science research. While PCA and EFA are well-established in the literature [30, 20, 26], Sparse PCA – an extension of PCA that introduces sparsity constraints to improve interpretability – is less well known in the social sciences [40].

Theoretically, these methods aim to achieve a common goal–dimensionality reduction. However, each technique is grounded in a distinct mathematical formulation, with different underlying assumptions and interpretations. Since the choice of method can significantly influence research outcomes, selecting an appropriate technique is essential to ensure results align with theoretical expectations. Therefore, this study conducts a systematic simulation to compare and evaluate these methods, using both simulated study and an empirical application in the social sciences. The objective is to provide guidelines for the appropriate use of each technique.

1.1 Background Information

High-dimensional data are datasets whose number of features is large compared to the number of observations. This results in complex data whose underlying meaning requires an appropriate method to uncover. Consequently, we need to transform the numerous features into manageable and meaningful features that are easy to interpret. Therefore, to solve data complexity, dimension reduction methods such as PCA, SPCA, and EFA are commonly used.

PCA, introduced by Karl Pearson [31] in 1901, is one of the oldest and most used techniques in many fields including psychometrics, sociology and economics [20]. It transforms correlated features into fewer uncorrelated features called principal components (PCs). These PCs are linear combinations of the original variables and are ordered such that the first component captures the largest possible variance in the data [20, 1]. The principal components thus retain the most important information from the original dataset, facilitating data reduction. However, a key limitation of PCA is that each component is generally a linear combination of all original features, which can hinder interpretation and labeling, especially in high-dimensional settings [40, 1]. Despite this, PCA stands out for its computational simplicity in cases where factors are theoretically not grounded.

EFA estimates unobservable variables called latent factors from data with numerous observable (manifest) variables. In social science research, e.g., psychology and education, variables like intelligence, attitude, and socioeconomic status are targeted, yet they cannot be directly measured. EFA aims to uncover such variables by learning hidden concepts and the relationships underlying the observed variables.

Various methods have been developed to improve the interpretability of the extracted factors. These methods are categorized into two classes: rotation and regularized estimation, which will be discussed in Chapter 2. These methods help improve the interpretation of factors and reduce overfitting in complex datasets.

SPCA is a modern technique that addresses the interpretability limitations of traditional PCA by imposing sparsity constraints on PCs. While SPCA has seen widespread adoption in fields like genomics and finance [40], its utilization in applied social science remains limited despite its potential to enhance clarity in dimensionality reduction [30]. By constraining many loadings to zero, SPCA produces components that isolate key variables, improving readability and simplifying substantive interpretation. This contrasts with standard PCA, where components often mix numerous variables, complicating inference in social science applications like psychological surveys or socioeconomic studies [30].

SPCA is a modern technique that addresses the interpretability limitations of traditional PCA by imposing sparsity constraints on PCs. While SPCA has seen widespread adoption in fields like genomics and finance [40], its utilization in applied social science remains limited despite its potential to improve clarity in dimensionality reduction [30]. By constraining many loadings to zero, SPCA produces components that isolate key variables, improving readability and simplifying substantive interpretation. This contrasts with standard PCA, where components often mix numerous variables, complicating inference in social science applications like psychological surveys or socioeconomic studies [30]. Recent research has demonstrated that SPCA improves factor interpretability and reduces overfitting in fields like education and sociology, where high-dimensional datasets are common [11]. The main advantage of SPCA over PCA is interpretability due to the sparsity nature of the loadings. This is a potential benefit for social scientists. Originally, SPCA was motivated by problems in the physical and engineering sciences, and is thus less known to social scientists.

A few researchers have brought out some practical comparisons in social science. In [35], a simulation study is conducted to compare PCA and EFA under various conditions including sample size. However, SPCA was not included in the study. Similar research was done by [8]. The study focuses on evaluating the performance and applicability of PCA and EFA across diverse datasets – both simulated and real-word. However, it does not explore SPCA or the effects of sparsity and interpretability in the context of dimensionality reduction. Another researcher, Park [30], acknowledges SPCA's sparsity component. The research critically distinguishes between sparse weights and sparse loadings in SPCA, highlighting that these represent fundamentally different concepts – sparse weights are the coefficients forming principal components as linear combinations of the original variables, and sparse loadings are the correlations (or covariances) between variables and component scores. However, the literature still lacks direct methodological comparisons of how SPCA performs in terms of interpretability relative to EFA and PCA. This gap motivates the present research, which aims to directly compare the structure recovery of PCA, SPCA and EFA across varying data conditions.

1.2 Problem Statement

With the increasing availability of high-dimension data that social scientists often handle, attention has turned to the selection of effective dimension reduction methods. This challenge arises from applying these techniques without practical comparison of each method's strengths and limitations across different data conditions. Since each technique has unique assumptions, the application of each method varies depending on the characteristics of the data and the research goal. Consequently, poor selection of these methods may lead to misinterpretation of the data.

1.3 Research Objective

In this thesis, we mainly focus on providing knowledge to practitioners, helping them select most suitable methods for dimension reduction, by achieving the following objectives:

- Present mathematical formulations of how PCA, SPCA and EFA perform dimension reduction.
- Perform a simulation study under different scenarios to illustrate where each technique performs more or less effectively.
- Provide practitioners with evidence-based guidelines for selecting suitable dimension reduction techniques based on a given data structure and specific research objectives.
- Demonstrate a practical implementation of PCA, SPCA and EFA, using an empirical social science dataset. Particular attention is given to introducing SPCA, which is less commonly used in social science research.

1.4 Thesis Organization

This work is organized as follows: Chapter 1 provides an introduction and background information of PCA, SPCA and EFA. Chapter 2 presents the mathematical formulations of the three methods. A simulation study is described in Chapter 3, followed by an empirical application in Chapter 4. Finally, Chapter 5 presents final remarks of the study.

2. Methods: Principal Component Analysis (PCA), Sparse Principal Component Analysis (SPCA) and Exploratory Factor Analysis (EFA)

2.1 PCA Formulation

One of the earliest and most widely used methods for reducing high-dimensional datasets while retaining maximal variability is PCA [20, 11]. PCA transforms correlated variables into a smaller set of uncorrelated principal components, which are linear combinations of the original variables [11]. The coefficients defining these linear combinations are termed component weights, while loadings are weights scaled by the standard deviations of the original variables. In standardized PCA (where variables are scaled to unit variance), loadings correspond to correlations between variables and component scores. In unstandardized PCA, loadings represent covariances, making them dependent on variable scales [2, 11].

While PCA can be derived differently [20, 11], one way is to minimize reconstruction error. Let $\mathbf{X} \in \mathbb{R}^{n \times p}$ be the data matrix, where n is the number of observations and p is the number of variables. It's common practice to assume the data are centered and scaled. We are interested in finding $m < p$ components. The most commonly used data structure is defined as [38]:

$$\mathbf{X} = \mathbf{L}\mathbf{F}^\top + \mathbf{E}, \tag{2.1.1}$$

where $\mathbf{F} \in \mathbb{R}^{n \times m}$ is the matrix of component scores, $\mathbf{L} \in \mathbb{R}^{p \times m}$ is the component loadings matrix, and $\mathbf{E}$ is the error matrix of $n \times p$. Here, the aim is to minimize the squared Frobenius norm of the error matrix $\mathbf{E}$ (also known as the least-squares approach) such that data matrix $\mathbf{X}$ is approximated as $\mathbf{X} \approx \mathbf{L}\mathbf{F}^\top$. This error term is given by $\mathbf{E} = \mathbf{X} - \mathbf{L}\mathbf{F}^\top$ and its Frobenius norm, $\|\mathbf{E}\|_F$, is defined as:

$$\|\mathbf{E}\|_F = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{p} e_{ij}^2}, \tag{2.1.2}$$

where e_{ij} are the elements of the error matrix. To do so, we focus on maximizing the variance captured by $\mathbf{L}$ and $\mathbf{F}$ such that:

$$(\hat{\mathbf{F}}, \hat{\mathbf{L}}) = \underset{\mathbf{F},\mathbf{L}}{\arg\min} \|\mathbf{X} - \mathbf{L}\mathbf{F}^\top\|_F^2 \quad \text{subject to} \quad \mathbf{F}^\top\mathbf{F} = \mathbf{I}, \tag{2.1.3}$$

where the constraint, $\mathbf{F}^\top\mathbf{F} = \mathbf{I} \in \mathbb{R}^{m \times m}$, with $\mathbf{I}$ being the identity matrix ensures that the columns of the factor matrix $\mathbf{F}$ are orthogonal [11, 20] .

A solution for the problem 2.1.3 can be obtained from the truncated singular value decomposition (SVD) of $\mathbf{X}$. The SVD of $\mathbf{X}$ is given by:

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^\top, \tag{2.1.4}$$

where $\mathbf{U} \in \mathbb{R}^{n \times m}$ and $\mathbf{V} \in \mathbb{R}^{p \times m}$ are semi-orthogonal matrices such that $\mathbf{U}^\top\mathbf{U} = \mathbf{V}^\top\mathbf{V} = \mathbf{I} \in \mathbb{R}^{m \times m}$, and $\mathbf{D} \in \mathbb{R}^{m \times m}$ is a diagonal matrix.

The solution [7, 11] to 2.1.4 is given by:

$$\hat{\mathbf{F}} = \mathbf{U}, \quad \hat{\mathbf{L}} = \mathbf{V}\mathbf{D}.$$

PCA is widely used in dimension reduction because of its efficient computation. However, its main limitation to the interpretation of the loadings is that each component is usually a linear combination of all variables [20, 11, 18]. Therefore, we use varimax rotation in this thesis to ensure a fair comparison with the other methods. Varimax rotation is a post-estimation method that seeks an orthogonal matrix $\mathbf{R}$ such that $\mathbf{R}^\top\mathbf{R} = \mathbf{I}$. So we replace $\mathbf{L}$ with $\mathbf{L}_{rot} = \mathbf{LR}$. Since $\mathbf{R}$ is orthogonal, our components remain uncorrelated. The rotated components are forced to have larger loading while others have loadings close to zero which makes it easier to interpret.

2.2 SPCA Formulation

Several SPCA techniques have been proposed to address PCA's interpretability issue by producing principal components with sparse loadings, where each component depends on only a subset of variables. This sparsity simplifies interpretation and variable selection, particularly in high-dimensional settings e.g., psychometric data where there are many more variables than observations [18, 11]. Some of the existing techniques impose sparsity on loadings (the correlations between variables and components), while others impose sparsity on weights (the coefficients defining the linear combination) [11]. While traditional methods such as semidefinite programming [5], penalized least squares [34] and rotation-thresholding techniques [25, 20] help to achieve simple structures, modern techniques like Zou's SPCA [40, 36] excel as they directly apply an ℓ_1 penalty (lasso) to the weights, encouraging sparsity, while an ℓ_2 penalty (ridge) stabilizes the solution.

In this thesis we intend to use Zou's SPCA [40], which is both popular in the literature and readily available for use in statistical applications like R. In this method, PCA is formulated as a regression-type optimization problem that incorporates both ℓ_1 and ℓ_2 penalties. This approach modifies the traditional PCA framework by introducing regularization terms that encourage sparsity in the component weights, resulting in the following optimization problem:

$$(\hat{\mathbf{W}}, \hat{\mathbf{P}}) = \underset{\mathbf{W},\mathbf{P}}{\arg\min} \left\|\mathbf{X} - \mathbf{XWP}^\top\right\|_F^2 + \sum_{k=1}^{K} \lambda \left\|\mathbf{w}_k\right\|_2^2 + \sum_{k=1}^{K} \lambda_{1,k} \left\|\mathbf{w}_k\right\|_1 \quad \text{s.t.} \quad \mathbf{P}^\top\mathbf{P} = \mathbf{I}, \tag{2.2.1}$$

where $\mathbf{X} \in \mathbb{R}^{n\times p}$ is the scaled and centered data matrix,$\mathbf{W} \in \mathbb{R}^{p\times K}$ is the sparse component weight matrix, $\mathbf{P} \in \mathbb{R}^{K\times K}$ is the orthonormal loadings matrix (to be estimated), and $\mathbf{w}_k$ are the k-th column of $\mathbf{W}$.

The terms λ and $\lambda_{1,k}$ are regularization parameters. The ridge ℓ_2 penalty, $\sum_{k=1}^{K} \lambda \left\|\mathbf{w}_k\right\|_2^2$, shrinks weights toward zero but does not promote sparsity. In contrast, lasso ℓ_1 penalty, $\sum_{k=1}^{K} \lambda_{1,k} \left\|\mathbf{w}_k\right\|_1$, encourages many entries in $\mathbf{w}_k$ to be exactly zero, thus achieving sparsity. When the two penalties are combined, thus we have the term *elastic net* regularization, which utilizes each of the penalty's benefit [40]. The difference between the newly constructed data $\mathbf{XWP}^\top$ and the original data $\mathbf{X}$ is termed the reconstruction error. It is measured by $\left\|\mathbf{X} - \mathbf{XWP}^\top\right\|_F^2$, which is the squared Frobenius norm 2.1.2. The component scores formed by $\mathbf{L} = \mathbf{XW}$, links the formulation to the classical PCA 2.1.3. Additionally, the constraint $\mathbf{P}^\top\mathbf{P} = \mathbf{I}$ enforces columns of $\mathbf{P}$ to be orthonormal, to ensure that the components are uncorrelated.

The objective is to estimate both component weights matrix $\hat{\mathbf{W}}$ and the component loadings matrix $\hat{\mathbf{P}}$ to solve 2.2.1. The proposed alternating minimization algorithm iteratively updates $\mathbf{W}$ and $\mathbf{P}$ while

keeping the other fixed. When updating $\mathbf{W}$ with fixed $\mathbf{P}$, the subproblem becomes a standard elastic net regression, which can be efficiently solved using existing techniques [40]. Therefore, SPCA achieves sparsity in the component weights matrix $\mathbf{W}$ which enforces some entries to be zero while leaving a few non-zero loadings to enhance interpretability.

In practice, careful selection of the penalty parameters is crucial for optimizing equation 2.2.1. In our methodology, we use cross-validation, which is the most commonly recommended approach for tuning these parameters in Zou's SPCA since it adapts well to different data structures

2.3 EFA Formulation

EFA is an essential technique used by statistician to identify latent factors underlying high-dimensional data[28, 4]. The objective of EFA is to identify interpretable loading matrices that explain relationship between the observed variables and latent factors [15, 21]. Therefore, sparsity on the loading is imposed to maximize the number of entries close to zero, allowing a few manifest variables to associate with each factor [15, 26]. This improves the interpretation of the loading matrix.

EFA has two classes of sparsity-inducing methods: rotation and regularized estimation methods. In the rotation method, a two-step procedure is involved. The first step is to estimate the loading matrix often obtained using maximum likelihood estimator [21], least square [22] or weighted-least-square [4]. The second step is the rotation which applies sparsity by using various loss functions. These rotations are categorized depending whether factors are allowed to have correlation or not [22, 23]. When factors are allowed to have correlation, the rotation is termed as orthogonal, otherwise oblique. Each method has different loss functions. Various loss functions used in rotation include varimax [23], oblimin[17], geomin [39], simplimax[24] and component loss functions (CLFs) [15, 16]. Contrary to rotation methods, regularized methods, achieve sparsity by imposing penalties like lasso, ridge or both directly during the estimation of the loading matrix [36]. The induced sparsity penalty shrinks most of the loadings towards zero, hence sparsity is achieved during estimation.

Having mentioned various EFA methods, we focus only on varimax (orthogonal rotation) and oblimin (oblique rotation), applied after estimating factors using the maximum likelihood estimator [21]. We also consider regularized EFA specifically the lasso penalty, to induce sparsity in the loading matrix. Recall the initial data structure from equation 2.1.1. For EFA, we define $\mathbf{L} \in \mathbb{R}^{p\times m}$ as the factor loadings matrix (relationship between observed variables and factors), $\mathbf{F} \in \mathbb{R}^{n\times m}$ is the matrix of factor scores and $\mathbf{E}$ is $n \times p$ matrix of unique factors or errors, capturing the variance in $\mathbf{X}$ not explained by the factors. The model assumes that the factor scores and errors are uncorrelated, i.e.,

$$\mathrm{Cov}(\mathbf{F}, \mathbf{E}) = \mathbf{0},$$

and that the errors follow a multivariate normal distribution with zero mean and diagonal covariance matrix $\mathbf{\Psi}$, such that

$$\mathbf{E} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Psi}).$$

For rotation method, the first step is to maximize the factor loading matrix $\mathbf{L}$ with respect to unique variances $\mathbf{\Psi}$

$$(\hat{\mathbf{L}}, \hat{\mathbf{\Psi}}) = \arg\max_{\mathbf{L},\mathbf{\Psi}} \log \ell(\mathbf{L}, \mathbf{\Psi}),$$

where $\log \ell(\mathbf{L}, \mathbf{\Psi})$ is the log-likelihood function which measures how well $\mathbf{L}$ and $\mathbf{\Psi}$ explain the observed data $\mathbf{X}$

Second, we rotate $\hat{\mathbf{L}}$ using varimax rotation when factors are assumed to be uncorrelated or oblimin when factors are allowed to be correlated. In regularized EFA, a sparsity penalty is incorporated with the maximum likelihood estimator. Specifically, the lasso variant imposes sparsity on the loading matrix $\mathbf{L}$, such that the objective function is modified to:

$$\log \ell(\mathbf{L}, \boldsymbol{\Psi}) - \lambda \sum_{j=1}^{p} \sum_{k=1}^{m} |l_{jk}| \,,$$

where $\lambda \geq 0$ is the regularization parameter, and l_{jk} are elements of the loading matrix $\mathbf{L}$. In this thesis, λ is typically selected using cross-validation, where the data is split into training and validation sets and the value of λ that minimizes the prediction error on the validation set is chosen. This approach balances model fit and interpretability by simplifying the factor loadings without losing important information.

3. Simulation Study

In this chapter, we present the procedures used to conduct the simulation study. A systematic approach was developed to compare principal component analysis (PCA), sparse principal component analysis (SPCA) and exploratory factor analysis (EFA).

First, synthetic data were generated based on a predefined population model 3.1.1. Then, a Monte Carlo simulation was used to repeatedly draw samples from this model, estimate the parameters in each iteration and evaluate the methods' performance by averaging the results over multiple iterations. The simulation study was carried out in R, which offers a wide range of packages for statistical computation and visualization [33]. Second, we applied each method (PCA, SPCA and EFA) under varying data conditions, including different levels of sparsity, sample sizes and number of observed variables, to evaluate their performance. This process allowed us to investigate each method's ability to uncover latent structure within the simulated data. Third, we evaluated the performance of each method with different metrics and analyzed their results for effective comparison. Finally, based on these results, we provide guidelines to practitioners on selecting appropriate methods.

3.1 Data-Generation Process

We simulated data from the following model:

$$\mathbf{X} = \mathbf{L}\mathbf{F}^\top + \mathbf{E}, \tag{3.1.1}$$

where the term $\mathbf{X}$ is the data matrix with dimension $n \times p$ such that n is number of observations and p is the number of observed variables. The term $\mathbf{L}$ is a matrix of true loading matrix of size $p \times m$. The term $\mathbf{F}$ is an $n \times m$ matrix of standard normal factors such that m is the number of factors. The term $\mathbf{E}$ is a noise matrix of size $n \times p$.

To generate the data matrix $\mathbf{X}$, we begin by constructing each component of the factor model. Specifically, we define the loading matrix $\mathbf{L}$ to represent three distinct factor loading structures such that the first αp variables load only on the first factor; all other loadings in their respective rows are set to zero, the second αp variables load only on the second factor, with zero loadings elsewhere in those rows and the remaining $(1-2\alpha)p$ variables have non-zero loadings on all factors, forming a dense structure. Each factor $\mathbf{F}$ is generated from a multivariate normal distribution $\mathbf{F}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_m)$ independently. An eigenvalue decay γ is also generated to determine the factor contribution which is given by $j^{-\gamma}$ where $\gamma \in \{0, 1, 2\}$. Next, we generate the error matrix $\mathbf{E}$, where each entry is drawn from a normal distribution with mean zero and variance $\sigma^2 \in \{0.1, 1, 5\}$. This allows us to simulate varying levels of noise in the data. To incorporate correlation among error terms, we impose a weak correlation structure with $\rho = 0.3$. Finally, the factor matrix $\mathbf{F}$ is generated from a multivariate normal distribution with moderate correlation among the factors, to reflect realistic dependency structures often encountered in social science data.

To achieve variety of datasets from the simulation, we perform a parallel computation with a grid of different factors ($n, p, m, \alpha, \gamma,$ and σ^2) as presented in table 3.1. This yields a total of $2\times3\times2\times3\times3\times3 = 324$ scenarios, where each scenario is repeated 100 times, making a total of $324\times100 = 32,400$ runs. By doing so, we produced a dataset with wide range of varying conditions to ensure that results generalize across realistic scenarios. This approach aligns with recommendations from previous simulation research in factor analysis such as [32]. We have discussed these data conditions below.

Sparsity (α): Sparsity is a key condition in our analysis. It refers to the proportion of zero entries in the factor loading matrix, controlling how many manifest variables each factor loads on. A high sparsity level ($\alpha = 0.8$) indicates that most loadings are zero, meaning each factor is associated with only a small subset of variables. Conversely, a low sparsity level ($\alpha = 0.2$) means that most loadings are nonzero, so each factor influences many variables. Adjusting sparsity allows the simulation to model both simple (sparse) and complex (dense) underlying structures, reflecting scenarios commonly encountered in social science research [40, 18].

Decay (γ): Decay show the simplicity/complexity of the data by determining how quickly the contribution, which is the variance explained, of successive factors diminishes. A low decay value ($\gamma = 0.0$) means that all factors contribute similarly to the variance, resulting in a more complex and evenly distributed factor structure. A high decay value ($\gamma = 2.0$) implies that the first few factors account for most of the variance, with remaining factors contributing very little, yielding a simpler and more dominant factor structure. This parameter reflects the eigenvalue decay rate, which influences factor separability and recovery difficulty [32].

Noise (σ^2): Noise represents the variance of the random error added to the data, simulating measurement error or unexplained variance typical in real-world datasets. Higher noise levels ($\sigma^2 = 5.0$) introduce more randomness, making the recovery of the true factor structure more challenging, while lower noise levels ($\sigma^2 = 0.1$) result in cleaner data and easier factor identification. Incorporating noise variance is essential to assess method robustness under varying data quality [28].

Variables (p): This parameter specifies the number of manifest variables in the simulated data. Varying the number of variables ($p = 50, 100, 200$) ensures that the simulation covers different data dimensionalities, from moderate to high-dimensional settings. Such variation tests the scalability and performance of each method under realistic social science data conditions [18].

Sample size (n): These are number of observations in each simulated dataset. Simulations were conducted with both small ($n = 100$) and large ($n = 500$) samples to evaluate how method performance scales with data availability. Sample size is a critical factor influencing the stability and accuracy of factor extraction [29].

Factors (m): The number of factors defines the dimensionality of the latent structure underlying the observed data. Simulations with $m = 3$ and $m = 5$ factors allow for the evaluation of method performance across structures of varying complexity. This variation helps assess how well methods recover simpler versus more complex factor models [9].

Parameter	**Values**
Sparsity (α)	0.2, 0.5, 0.8
Decay (γ)	0.0, 1.0, 2.0
Noise (σ^2)	0.1, 1.0, 5.0
Variables (p)	50, 100, 200
Sample size (n)	100, 500
Factors (m)	3, 5

Table 3.1: Simulation parameter grid

3.2 Evaluation of the Methods on Simulated Data

Once the data is generated, we fit PCA, SPCA, and EFA. For an effective comparison of these methods, we define several evaluation metrics. The goal is to assess how each method performs under varying conditions data conditions, including different sample sizes, dimensionalities, factor complexities, sparsity levels, factor correlations and noise levels as discussed previously. Our focus is on evaluating reconstruction accuracy, loading accuracy, sparsity recovery, and latent factor score correlation. In this section, we briefly discuss the metrics used to achieve these objectives.

3.2.1 Mean squared error. Mean squared error (MSE) is a metric used to calculate reconstruction error [20]. It quantifies how well a method can reproduce the observed data using the estimated factors and loadings. Specifically, MSE measures the average squared difference between the original observed values and the values reconstructed from the model. A lower MSE indicates that the factor solution captures more of the underlying structure and explains more variance in the data, while a higher MSE suggests that important information is being missed.

For each method, we evaluate how well the reconstructed data $\hat{\mathbf{X}}$ approximates the original simulated data $\mathbf{X}$, defined as:

$$\text{MSE} = \frac{1}{np}\sum_{i=1}^{n}\sum_{j=1}^{p}\left(X_{ij} - \hat{X}_{ij}\right)^2.$$

3.2.2 False positive/negative rate. To assess how effectively each method distinguishes between zero and non-zero loadings, we use false positive rate (FPR) and false negative rate (FNR). FPR is the proportion of estimated factor loadings that are non-zero when the true loading is actually zero. In other words, it measures how often a method falsely identifies a variable as contributing to a latent factor when it does not. A low FPR indicates that the method avoids falsely including irrelevant variables in the factor model. On the other hand, FNR is the proportion of estimated loadings that are zero when the true loading is non-zero. It indicates how often a method fails to detect a true association between a variable and a latent factor. A high FNR implies that the method may miss important variables, thereby failing to capture the underlying structure of the data. These metrics are discussed below.

$$\text{FPR} = \frac{\sum_{i,j}\mathbf{1}\left[|\hat{L}_{ij}| > \varepsilon \wedge L_{ij} = 0\right]}{\sum_{i,j}\mathbf{1}\left[L_{ij} = 0\right]}. \tag{3.2.1}$$

$$\text{FNR} = \frac{\sum_{i,j}\mathbf{1}\left[|\hat{L}_{ij}| \leq \varepsilon \wedge L_{ij} \neq 0\right]}{\sum_{i,j}\mathbf{1}\left[L_{ij} \neq 0\right]}. \tag{3.2.2}$$

Equations 3.2.1 and 3.2.2 show these metrics, using a threshold $\varepsilon = 0.5$ to distinguish estimated zero values from non-zero values. The symbol $\mathbf{1}[\cdot]$ denotes an indicator function, which takes the value 1 when the condition inside the brackets is true, and 0 otherwise. For instance, in FPR equation 3.2.1, the indicator function counts the number of loadings that are incorrectly estimated as non-zero ($|\hat{L}_{ij}| > \varepsilon$) when the true loading is actually zero ($L_{ij} = 0$). The sum of these counts is then divided by the total number of true zero loadings, yielding the proportion of false positives among all true zeros. On the other hand, for FNR equation 3.2.2, the indicator function counts how many loadings are incorrectly estimated as zero or near zero ($|\hat{L}_{ij}| \leq \varepsilon$) when the true loading is non-zero ($L_{ij} \neq 0$). Dividing this count by the total number of true non-zero loadings provides the proportion of false negatives. Together,

these metrics quantify the accuracy of sparsity recovery by measuring the rates of incorrect inclusion and exclusion of loadings in the estimated factor structure.

3.2.3 Loading accuracy. To compare the accuracy of PCA, SPCA, and EFA in estimating the true factor loading matrix, we evaluate the loading error, which quantifies the discrepancy between the estimated loadings $\hat{\mathbf{L}}$ and the true loadings $\mathbf{L}$. This metric reflects a method's ability to recover the underlying latent structure of the data.

However, factor models are identifiable only up to an orthogonal rotation, meaning that direct element-wise comparison between $\hat{\mathbf{L}}$ and $\mathbf{L}$ is not meaningful without first aligning them [3]. To address this, we apply Procrustes rotation, which finds the optimal orthogonal transformation $\mathbf{R}$ that minimizes the Frobenius norm of the difference between the rotated estimated loadings and the true loadings:

$$\mathbf{R}^* = \arg \min_{\mathbf{R}^\top \mathbf{R} = \mathbf{I}} \|\hat{\mathbf{L}}\mathbf{R} - \mathbf{L}\|_F^2. \tag{3.2.3}$$

After alignment, the loading error is computed as the Frobenius norm of the difference between $\hat{\mathbf{L}}\mathbf{R}^*$ and $\mathbf{L}$, as defined in Equation 2.1.2.

3.2.4 Factor score correlation. Another key metric for evaluating method's performance is the correlation between the true latent factor scores $\mathbf{F}$ and the estimated factor scores $\hat{\mathbf{F}}$. This is particularly important in social science research, where estimated factor scores are often used for substantive interpretation and further analysis. To achieve this, after alignment using Procrustes rotation as defined previously 3.2.3, and then apply the same optimal rotation ($\mathbf{R}^*$) to the estimated factor scores. We then compute the correlation matrix between the columns of $\mathbf{F}$ and $\hat{\mathbf{F}}$, where each entry represents the Pearson correlation coefficient between a true and an estimated factor. This matrix contains the Pearson correlation coefficients between each pair of corresponding columns (factors) from the true and estimated factor score matrices. The degree of agreement between the estimated factor scores produced by each method and the true latent factors (i.e., the underlying values of the latent variables for each observation) is then assessed using these correlations. Higher positive values indicate better recovery of the true latent structure.To summarize the overall recovery, we compute the mean of the diagonal elements of the correlation matrix, which represent the correlations between each true factor and its corresponding estimated factor. This mean factor score correlation is calculated as follows:

$$\text{Mean factor score correlation} = \frac{1}{m} \sum_{j=1}^{m} \text{cor}\left(\mathbf{F}_j, \hat{\mathbf{F}}_j\right),$$

where $\text{cor}(\mathbf{F}_j, \hat{\mathbf{F}}_j)$ denotes the pearson correlation between the jth true and estimated factor scores. Higher values indicate better recovery of the true latent structure.

3.3 Results of the Methods on Simulated Data

So far, we have generated synthetic data and fitted the data in the three methods. Evaluation methods discussed in the previous section have been employed to assess performance of each method. We now present results of the analysis that we are going to use further to understand the nature of each method in factor extraction.

3.3.1 Results on reconstruction accuracy. Reconstruction accuracy results for PCA, SPCA, and EFA are presented in Figure 3.1. The results reveal varying patterns regarding the reconstruction accuracy of each method. Across all methods, increasing the sample size from $n = 100$ to $n = 500$ generally leads to a reduction in MSE. This improvement is most pronounced for PCA and to a lesser extent, SPCA.

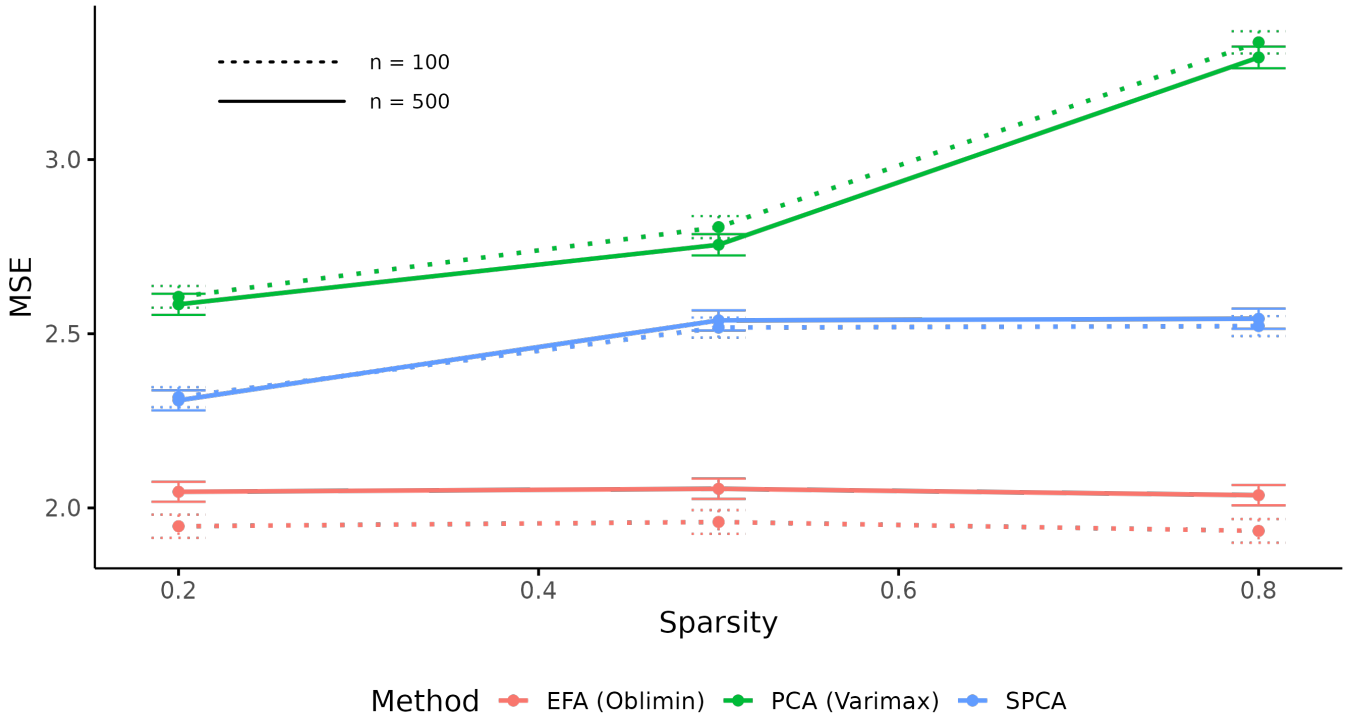

Figure 3.1: Mean squared error by sparsity and sample size

The reduction in error with larger n reflects the greater reliability in estimating the underlying factor structure when more data are available [32]. Interestingly, EFA exhibits consistently low MSE regardless of sample size, with a slight tendency for lower error at smaller samples. This robustness suggests that EFA's estimation procedures and rotation methods are less sensitive to sample size in these simulation conditions, possibly due to its focus on modeling shared variance and its adaptability to different loading structures.

When the factor loading matrix is dense (low sparsity), all three methods yield relatively similar MSE values, indicating comparable reconstruction accuracy. This suggests that when most variables load on each factor, both PCA, SPCA, and EFA approaches are able to capture the data structure effectively. As sparsity increases, the differences among methods become more pronounced. The MSE for PCA increases sharply with higher sparsity, reaching approximately 3.3 at $\alpha = 0.8$. This is because PCA, by design, produces dense components that do not align well with underlying sparse structures. As a

result, it fails to accurately reconstruct data when only a few variables are truly associated with each factor, leading to greater reconstruction error [40]. SPCA, which enforces sparsity via an elastic net penalty, demonstrates improved reconstruction accuracy over PCA in sparse settings. Its MSE rises only moderately with increasing sparsity (from 2.2 to 2.5), reflecting its ability to better capture the true sparse structure. However, SPCA does not match the accuracy of EFA. EFA consistently achieves the lowest MSE across all conditions, with values ranging narrowly from 1.9 to 2.1. This highlights EFA's robustness to varying sparsity levels, likely due to its focus on modeling shared variance. EFA's ability to adapt to both dense and sparse loading patterns makes it particularly effective for reconstructing the underlying data structure.

3.3.2 Results on sparsity recovery. Figure 3.2 illustrates how the methods differ in their ability to recover the true sparsity structure, assessed by FPR and FNR across sparsity (α) and sample size (n).

The left panel of Figure 3.2 displays the FPR. PCA starts with a relatively low FPR of about 0.30 at low sparsity ($\alpha = 0.2$), but the FPR increases steadily with higher sparsity, reaching approximately 0.50 at $\alpha = 0.8$. This indicates that PCA is more prone to including irrelevant variables as sparsity increases, reflecting its tendency to produce dense components that do not align with sparse structures [28]. EFA (Oblimin) follows a similar trend, starting at around 0.45 for low sparsity, dipping at moderate sparsity, and increasing again to about 0.50 at high sparsity. This suggests EFA also struggles to exclude irrelevant variables in highly sparse settings, consistent with its design to capture broad patterns of shared variance [20]. Sparse PCA and Regularized EFA show the opposite pattern: both begin with higher FPRs at low sparsity (around 0.55 and 0.60, respectively) but their FPRs decrease as sparsity increases, reaching about 0.52 (Sparse PCA) and 0.48 (Regularized EFA) at high sparsity. This demonstrates that these methods are more effective at correctly identifying zero loadings in sparse settings, due to their explicit sparsity constraints [40]. Sample size has minimal effect on FPR for all methods, indicating that sparsity recovery is primarily determined by the method's underlying assumptions rather than the amount of data.

Figure 3.2: False positive/negative rate by sparsity

The right panel of Figure 3.2 shows FNR. PCA exhibits the highest FNR at low sparsity (about 0.50 at $\alpha = 0.2$), which decreases to around 0.25 at moderate sparsity, then rises slightly to 0.45 at high

sparsity. This pattern suggests that PCA frequently misses true nonzero loadings in dense settings, likely because its dense component structure may overlook significant relationships [20]. Sparse PCA, EFA, and Regularized EFA all start with lower FNRs at low sparsity (0.30, 0.25, and 0.20, respectively) and steadily improve as sparsity increases, dropping to approximately 0.05 at $\alpha = 0.8$. This reflects their greater ability to detect true nonzero loadings in sparse conditions, with Sparse PCA and Regularized EFA benefiting from sparsity-inducing penalty while EFA (Oblimin) leveraging shared variance modeling [13, 28]. Again, sample size exerts little influence on FNR, reinforcing that method-specific properties dominate performance in sparsity recovery.

From the results we see that, Sparse PCA and Regularized EFA excel in sparse settings, reducing both false positives and false negatives as sparsity increases, making them well-suited for applications where interpretability and correct identification of zero loadings are crucial. PCA and EFA (Oblimin) are more effective in dense settings but become less reliable as sparsity increases, due to their inherent preference for dense representations. Across all methods, changes in sample size have a negligible impact on sparsity recovery, highlighting the importance of methodological choice over data volume for this aspect of performance.

3.3.3 Results on loading recovery. The figure 3.3 shows loading error for PCA, SPCA and EFA across varying eigenvalue decay rate (γ) and sample size (n).

Across all methods and sample sizes, loading error decreases as the eigenvalue decay rate increases. This trend is most pronounced for PCA and SPCA.

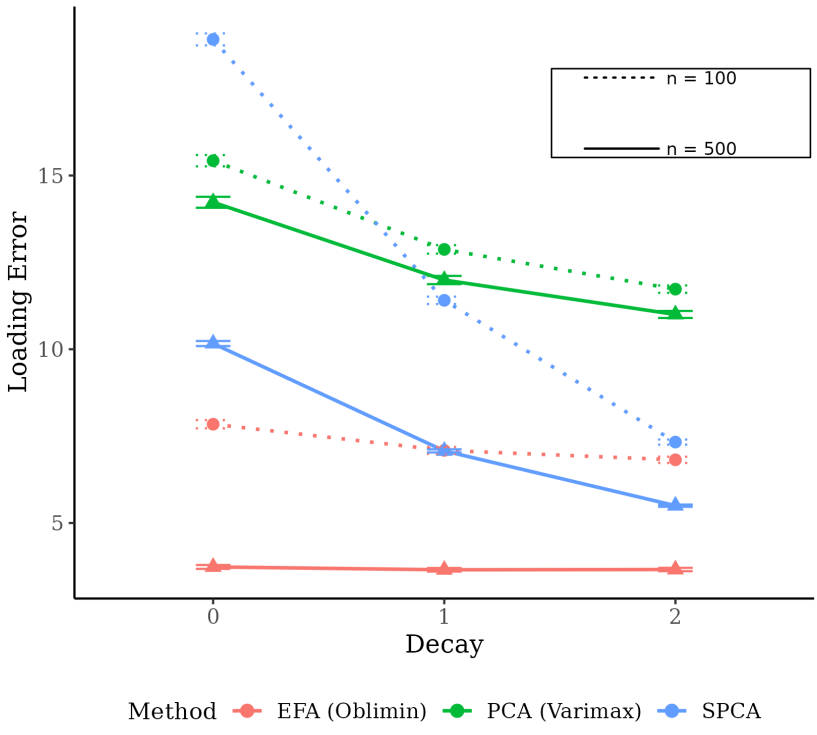

Figure 3.3: Loading error by decay and sample size

At $\gamma = 0$, which corresponds to a complex structure, both PCA and SPCA exhibit high loading errors. As γ increases to 2, indicating a simpler structure, loading errors for these methods decline substantially.

This pattern demonstrates that all methods recover factor loadings more accurately when the data are dominated by a few strong factors and less accurately when variance is spread across many factors.

EFA consistently achieves the lowest loading error across all γ, especially at larger sample. Its error remains low and stable, demonstrating robust performance even in complex scenarios. This highlights EFA's strength in modeling latent factors and their relationships with observed variables [9]. SPCA exhibits intermediate performance. At smaller sample size and low sparsity, SPCA exhibits the highest loading error. However, as sample size increases and the structure becomes simpler, SPCA's loading error decreases more sharply than PCA's, approaching EFA's performance at the highest decay rate. This improvement reflects SPCA's ability to impose sparsity on the loadings, which helps in simpler structures but can distort the true structure when data are limited or the true structure is not strictly sparse [40]. Although PCA's loading error decreases as γ increases in both sample sizes, PCA remains less accurate than EFA and SPCA. This is because PCA seeks directions of maximum variance, which aligns poorly with the true structure when variance is spread across many factors (low decay), leading to mixed or split factors and increased loading error [19].

Therefore, the results show that EFA provides the most robust and accurate recovery of factor loadings across varying levels of structural complexity and sample size. SPCA shows improved performance with larger samples and simpler structures, while PCA is the most sensitive to both data complexity and sample size, consistently yielding the highest loading error. These findings underscore the importance of method selection and sample size considerations in factor recovery tasks.

3.3.4 Result on factor score correlation. Figure 3.4 presents boxplots of the mean score correlation of PCA, SPCA and EFA, evaluated at varying sample sizes.

At the $n = 100$, EFA demonstrates superior performance compared to both PCA and SPCA.The median mean factor score correlation for EFA is near 0.4, with the interquartile range (IQR) spanning from about 0.2 to 0.4. Notably, several results for EFA approach the upper bound of 1.0, indicating that, in some cases, EFA is able to nearly perfectly recover the true factor scores even with limited data. In contrast, both PCA and SPCA exhibit lower and more variable correlations. For PCA, the median is around 0.1, with an IQR ranging from approximately -0.1 to 0.2. SPCA shows a similar pattern, with a median near 0.3 and an IQR from about 0.0 to 0.3. Both methods display a wider spread and more negative outliers, indicating less stable and less accurate factor score recovery at this sample size.

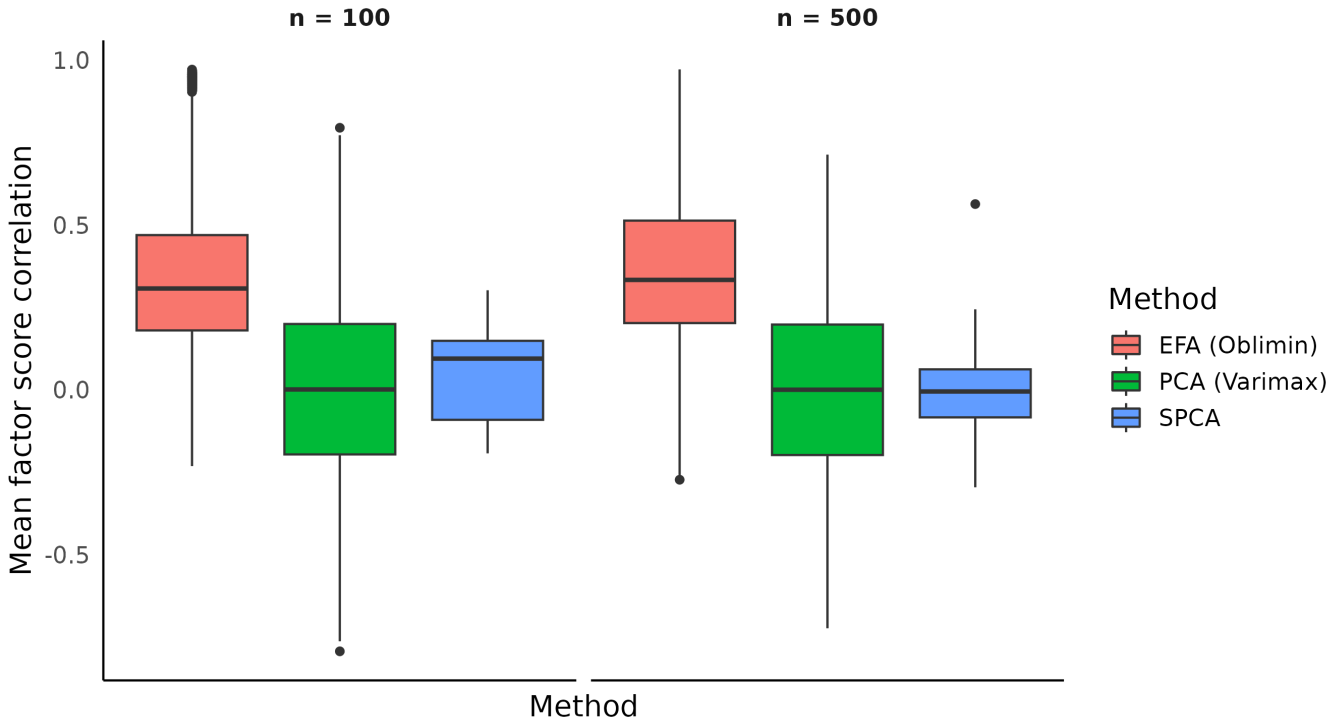

Figure 3.4: Factor score correlation across methods

Increasing n to 500 leads to a marked improvement in EFA's performance. The median mean factor score correlation for EFA rises to approximately 0.6, with the IQR now ranging from 0.4 to 0.6. The upper whisker extends close to 1.0, highlighting greater accuracy and stability in factor score recovery as sample size increases. PCA and Sparse PCA, however, maintain a median correlation near 0.1, with IQRs similar to those observed at $n = 100$. Low correlations suggests limitations of these methods for factor score estimation, recovery of the underlying latent structure.

Clearly the results shows that EFA provides the most accurate and stable recovery of factor scores across both small and large sample sizes. The improvement in EFA's performance with increased sample size is consistent with statistical theory, emphasizing the importance of sufficient data for reliable latent variable estimation. In contrast, PCA and Sparse PCA are substantially less effective, with low median correlations and high variability regardless of sample size. This finding underscores the limitations of these methods for factor score estimation, particularly in scenarios where accurate recovery of the

underlying latent structure is critical.

3.4 Guidelines for Method Selection

Selecting the most appropriate dimensionality reduction method is crucial for reliable results. The choice of method should be informed by the underlying characteristics of the data and the specific goals of the analysis. Based on our simulation results, we offer the following guidelines.

PCA is most effective in scenarios where the data structure is dense and variance is dominated by a few strong factors. The method benefits from larger sample sizes, which help stabilize the estimates. However, PCA becomes less effective when the eigenvalue decay rate is low and performs poorly in recovering factor scores. Consequently, PCA is less suitable in contexts where the accurate identification of zero loadings, interpretability, or precise factor score recovery is required, particularly in sparse or highly complex data settings.

SPCA is most effective when the true underlying structure contains many zero loadings and interpretability through sparsity is a priority. In terms of sparsity recovery, SPCA achieves superior identification of zero loadings compared to PCA. However, its effectiveness depends on having a moderate to large sample size; with small samples, SPCA may distort loadings and yield poor factor score recovery. Additionally, when the true structure is not strictly sparse or is highly complex, the imposed sparsity constraints can introduce bias and reduce overall accuracy.

EFA (Oblimin-rotated) is most effective when accurate recovery of both factor loadings and factor scores is essential, regardless of whether the data structure is dense, sparse, or complex. It adapts well to varying sparsity levels and performs robustly even in limited data conditions. EFA is especially suitable when the goal is to model shared variance and latent constructs. While regularized EFA incorporates penalty terms that encourage sparsity in the factor loadings, it combines the strengths of EFA in modeling latent structure with the interpretability benefits of sparsity.

4. Empirical Analysis

Having examined the performance of principal component analysis (PCA), sparse principal component analysis (SPCA), and exploratory factor analysis (EFA) in the simulation study (Chapter 3), we gained important insights into how well each method recovers latent structure under controlled conditions. In this chapter, we present a practical application in a real-world scenario. Since the simulated study relied on generated data under predefined conditions, it may not fully capture the complexities of real-world datasets, including measurement errors and response variability commonly observed in social science research [28]. To assess whether the results from the simulation study generalize to real data, an empirical application is essential. This allows us to evaluate the real-world applicability of these techniques.

We apply the methods to a social science dataset – a domain in which techniques such as PCA, SPCA, and EFA are commonly used to uncover latent factors within data structures [12]. In particular, the Big Five personality dataset is a widely used example in social science applications. Using this dataset, we aim to demonstrate the effectiveness of PCA, SPCA, and EFA in identifying the Big Five traits: Extraversion (E), Agreeableness (A), conscientiousness (C), neuroticism (N) and openness (O).

4.1 Description of Big Five

In this application, we use the Big Five personality dataset from the `qgraph` package in R. The dataset comprises responses to the Dutch translation of the NEO Personality Inventory (NEO-PI-R), a widely used tool for assessing the Big Five personality traits. The dataset contains 500 observations (students) and 240 variables (items). All variables are numeric, recorded on a 5-point Likert scale (1 to 5), indicating the degree of agreement or disagreement with each statement [14]. Each observation, corresponds to a student, and the 240 items are divided into 48 items per trait, resulting in a data matrix of size 500×240. This structure makes the dataset well-suited for factor analysis, which aims to uncover the underlying latent structure of personality traits [6]. Additionally, the dataset is appropriate for this study due to its reliability for testing factor recovery techniques supported by its previous use in psychometric research [6].

4.2 Preparation of Big Five

The `qgraph` package in R was used to load the Big Five dataset. An initial check for missing values was performed using, since missing data can distort covariance structures leading to biases in factor analysis [28]. Zero missing entries were confirmed. We also ensured consistency with the dataset's documentation.

Given that the original dataset contains 240 items, this causes items within the same facet (sub-trait) to be highly correlated [12]. Directly analyzing these items could lead to multicollinearity, which destabilizes factor solutions and make them difficult to interpret. Therefore, to address this, we reduced the dataset to facet-level composites by averaging the eight items within each facet. This yielded 30 facet scores, comprising six facets for each of the five traits.

The data were standardized by centering each variable to have a mean of 0 and scaling it to have a standard deviation of 1. This standardization ensures comparability across variables by converting

the covariance matrix into a correlation matrix, where all variables contribute equally to the analysis. Standardization is critical for multivariate techniques such as PCA and EFA, which rely on these matrices to derive components or factors. [20, 10].

As in the simulation study, the standardized Big Five dataset was fitted to PCA, SPCA, and EFA methods to extract the five traits of the dataset. For PCA, a varimax rotation was applied after component extraction to improve interpretability. SPCA employed both ℓ_1 (lasso) and ℓ_2 (ridge) penalties to produce sparse loadings. Cross-validation was used to select the optimal penalty parameter, balancing sparsity with the amount of variance explained. For EFA, maximum likelihood estimation was used, followed by oblimin rotation. Additionally, a regularized version of EFA was fitted using an ℓ_1 penalty to induce sparsity in the factor loadings.

4.3 Evaluation Using Tucker Congruence Coefficient

To evaluate how well each method preserves the latent factors of the Big Five personality dataset, we computed Tucker congruence coefficient to assess the degree of similarity between the factor solutions [37]. Tucker's congruence coefficient between two factor loading vectors $\mathbf{a}$ and $\mathbf{b}$, for example those obtained from PCA and SPCA respectively, is defined as:

$$\phi = \frac{\sum_{i=1}^{p} a_i b_i}{\sqrt{\sum_{i=1}^{p} a_i^2}\sqrt{\sum_{i=1}^{p} b_i^2}},$$

where p is the number of observed variables. The coefficient ϕ typically ranges from -1 to 1, with values near 1 indicating high similarity, values near zero indicating no similarity, and negative values indicating an opposite relationship between the factors [27].

Tucker congruence coefficient provides a consistent and interpretable measure of factor similarity, allowing direct comparison across methods. According to established guidelines, values above 0.85 indicate good similarity, while values above 0.95 suggest that the factors are nearly identical [27]. This coefficient is widely used in psychometric research to evaluate the replicability and comparability of factors across different analytical approaches [27].

4.4 Empirical Results

From the extracted components (for PCA and SPCA) and factors (for EFA), we obtained the corresponding loading matrices (Tables 4.1, 4.3, 4.2). These loadings were then used to compute Tucker's congruence coefficients, which allowed us to generate three congruence heatmaps.

The first heatmap compares components extracted by PCA and SPCA (Figure 4.1); the second compares EFA and SPCA (Figure 4.3); and the third compares EFA and PCA (Figure 4.2). These heatmaps illustrate the similarity between factors and components for each pairwise comparison. The values in the heatmaps represent Tucker's congruence coefficients: low off-diagonal values indicate that non-corresponding factors or components are distinct, while high values along the diagonal suggest strong similarity between corresponding factors or components.

4.4.1 Results for PCA and SPCA. In Figure 4.1, the diagonal congruence values between PCA and SPCA components show a maximum value of 0.80, which falls below the conventional threshold for "excellent" congruence (typically above 0.85 [27]). This indicates only moderate similarity between the corresponding components extracted by the two methods. At the same time, the consistently low off-diagonal values demonstrate that non-corresponding components remain distinct, with minimal overlap between unrelated pairs.

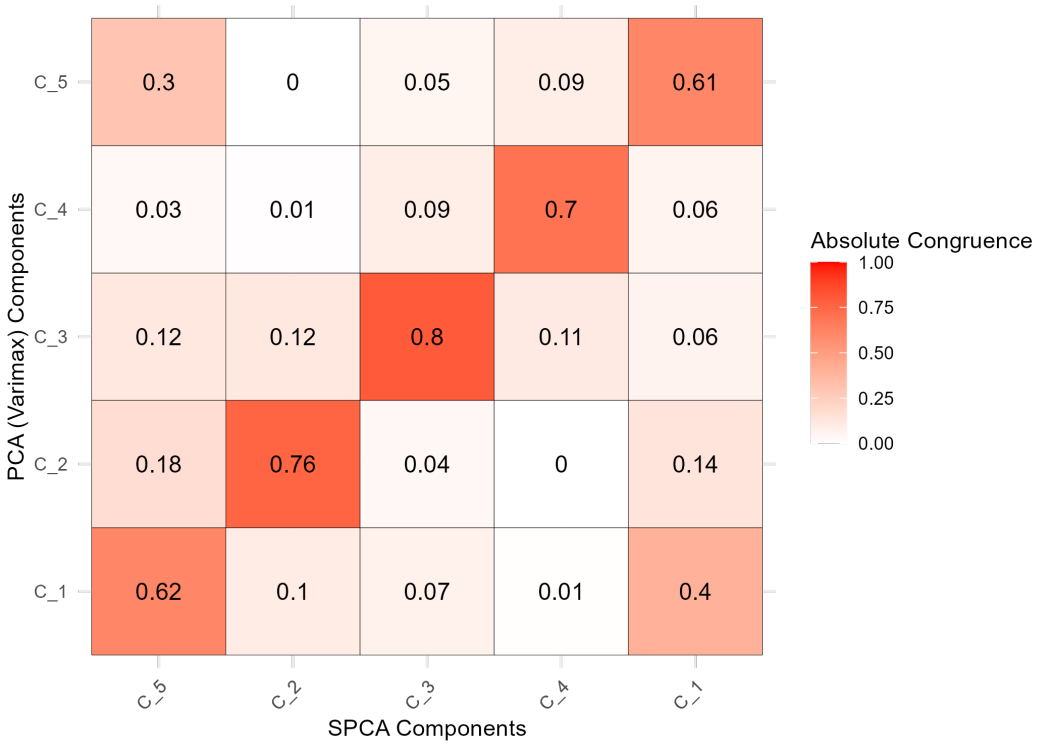

Figure 4.1: PCA (Varimax) vs SPCA congruence coefficients

This pattern aligns with the methodological differences between PCA and SPCA. SPCA imposes a sparsity constraint, resulting in components defined by a smaller subset of variables with nonzero loadings, thereby enhancing interpretability and making the components easier to label. In contrast, PCA seeks to maximize the total variance explained, often yielding components that are more comprehensive but less sparse, and thus less directly interpretable. The moderate congruence observed here suggests that

while SPCA improves interpretability by producing more clearly defined components, it may sacrifice some of the subtle variance captured by PCA, resulting in only moderate rather than high congruence between the two sets of components.

4.4.2 Results for EFA and PCA. The congruence values for EFA (Oblimin rotation) versus PCA (Varimax rotation) in Figure 4.2 indicate a generally high degree of similarity between several factors extracted by EFA and the corresponding PCA components. Notably, multiple diagonal coefficients exceed 0.90, with the highest values reaching 0.94, which approaches the conventional threshold for "nearly identical" factors (typically set at 0.95 [27]). This suggests that, for most factors, the underlying structure recovered by both methods is highly consistent.

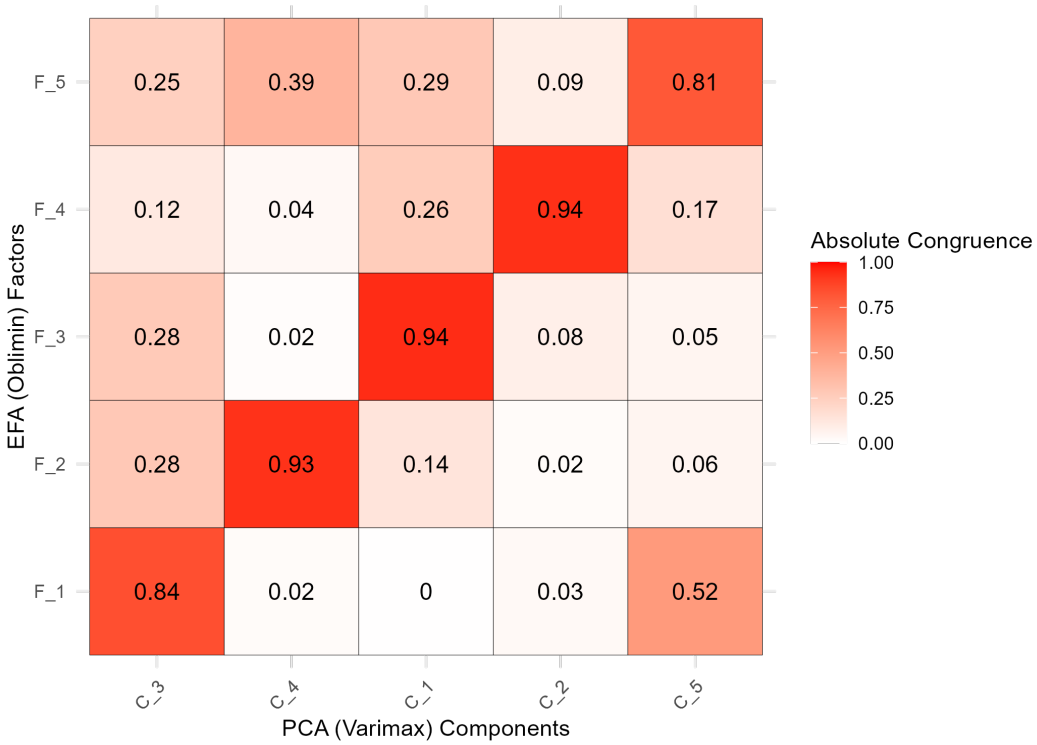

Figure 4.2: EFA (Oblimin) vs PCA (Varimax) congruence coefficients

Both PCA and EFA, especially when paired with appropriate rotations to improve interpretability, are designed to capture the major sources of variance in the data. PCA maximizes explained variance, while EFA models latent variables underlying the observed correlations. The high congruence values observed here reflect that, in this dataset, both approaches effectively recover the Big Five personality structure. This strong similarity across methods supports the empirical robustness of the five-factor model and demonstrates its stability under different analytical frameworks.

It is also worth noting that the off-diagonal values are generally low, indicating that non-corresponding factors remain distinct and that there is little overlap between unrelated components. This further reinforces the validity of the factor solutions and their interpretability.

4.4.3 Results for EFA and SPCA. . Conversely, the congruence values between EFA (Oblimin rotation) factors and SPCA components (Figure 4.3) are generally lower than those observed between EFA and PCA. The heatmap shows that only a few factor-component pairs reach moderate similarity, with congruence coefficients ranging from 0.41 to 0.76. Her, we see that the highest congruence observed is 0.76, while several other values fall below 0.60, indicating limited overlap.

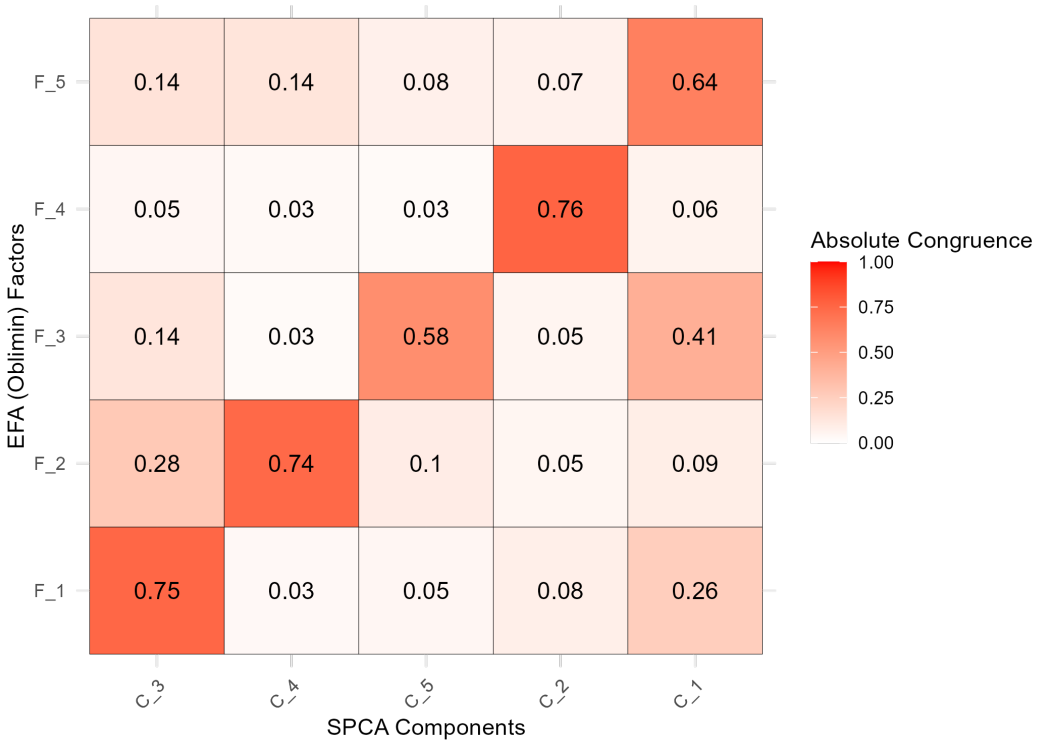

Figure 4.3: EFA (Oblimin) vs SPCA congruence coefficients

This pattern reflects the methodological impact of SPCA's sparsity constraint. By enforcing sparsity, SPCA produces components defined by fewer variables with nonzero loadings, which often results in factors that are more distinct and easier to interpret, but less aligned with the broader latent structure captured by EFA. In contrast, EFA allows for more complex loading patterns, including potential cross-loadings, which can better capture the richness of the underlying latent constructs.

The variability in congruence values highlights that some EFA factors and SPCA components still share a reasonable degree of similarity, but many do not. This reduced overall congruence suggests that SPCA prioritizes interpretability and simplicity – making the factor structure clearer, at the expense of fully recovering the more nuanced relationships present in the data, as revealed by EFA. Thus, while SPCA's simplified structure can aid interpretation, it may miss some of the complexity inherent in the latent constructs.

Variable	PC1	PC2	PC3	PC4	PC5
N Facets					
N_Facet1	−0.364	0.000	0.000	0.000	0.000
N_Facet2	0.261	0.000	−0.247	0.250	0.000
N_Facet3	−0.310	−0.209	0.174	0.000	0.000
N_Facet4	0.277	0.140	0.000	−0.259	0.000
N_Facet5	0.000	0.000	−0.317	0.000	0.000
N_Facet6	0.259	0.000	0.000	0.000	0.000
E Facets					
E_Facet1	0.476	0.000	0.000	0.000	−0.271
E_Facet2	0.000	0.000	−0.347	0.000	0.223
E_Facet3	0.314	0.134	0.000	−0.235	−0.139
E_Facet4	0.000	0.000	0.000	0.109	0.347
E_Facet5	0.000	−0.360	0.000	−0.177	0.000
E_Facet6	0.000	0.274	−0.306	0.000	0.000
O Facets					
O_Facet1	0.000	0.000	−0.170	−0.262	0.000
O_Facet2	0.356	−0.178	0.000	0.000	0.000
O_Facet3	0.279	0.117	−0.163	0.276	0.213
O_Facet4	−0.125	0.000	−0.454	0.000	0.000
O_Facet5	0.116	0.000	0.109	0.000	0.339
O_Facet6	0.000	−0.477	0.000	0.000	0.000
A Facets					
A_Facet1	0.000	0.000	0.000	0.000	0.361
A_Facet2	0.122	−0.112	−0.118	−0.107	−0.204
A_Facet3	0.136	0.110	0.000	−0.319	0.000
A_Facet4	0.000	−0.235	0.000	0.000	0.282
A_Facet5	0.000	0.000	−0.232	0.115	0.309
A_Facet6	0.000	0.000	−0.131	−0.489	−0.123
C Facets					
C_Facet1	0.225	0.000	0.119	−0.200	0.170
C_Facet2	0.000	−0.371	−0.117	0.000	−0.108
C_Facet3	0.000	0.000	0.000	−0.235	0.233
C_Facet4	0.000	−0.100	0.000	−0.388	0.000
C_Facet5	0.332	−0.170	0.175	0.000	0.000
C_Facet6	0.233	0.000	0.000	0.000	0.163

Table 4.1: PCA (Varimax) loadings

Variable	PC1	PC2	PC3	PC4	PC5
N Facets					
N_Facet1	0.000	0.000	−0.414	0.000	0.000
N_Facet2	−0.314	0.000	0.000	0.000	0.000
N_Facet3	0.000	−0.376	0.000	0.000	0.000
N_Facet4	0.000	0.000	0.000	0.000	−0.305
N_Facet5	0.000	0.000	−0.316	0.000	0.000
N_Facet6	0.000	0.000	0.000	0.000	0.000
E Facets					
E_Facet1	0.000	0.000	0.000	0.000	−0.884
E_Facet2	0.000	0.000	−0.464	0.000	0.000
E_Facet3	0.000	0.000	0.000	0.000	−0.337
E_Facet4	−0.309	0.000	0.000	0.000	0.000
E_Facet5	0.000	−0.339	0.000	0.000	0.000
E_Facet6	0.000	0.000	−0.323	0.000	0.000
O Facets					
O_Facet1	0.000	0.000	0.000	0.000	0.000
O_Facet2	0.000	0.000	0.000	0.000	0.000
O_Facet3	−0.374	0.000	0.000	0.000	0.000
O_Facet4	0.000	0.000	−0.514	0.000	0.000
O_Facet5	−0.442	0.000	0.000	0.000	0.000
O_Facet6	0.000	−0.606	0.000	0.000	0.000
A Facets					
A_Facet1	0.000	0.000	0.000	0.000	0.000
A_Facet2	0.000	0.000	0.000	0.000	0.000
A_Facet3	0.000	0.000	0.000	−0.364	0.000
A_Facet4	0.000	0.000	0.000	0.000	0.000
A_Facet5	0.000	0.000	0.000	0.000	0.000
A_Facet6	0.000	0.000	0.000	−0.653	0.000
C Facets					
C_Facet1	−0.309	0.000	0.000	0.000	0.000
C_Facet2	0.000	−0.467	0.000	0.000	0.000
C_Facet3	0.000	0.000	0.000	0.000	0.000
C_Facet4	0.000	0.000	0.000	−0.453	0.000
C_Facet5	0.000	0.000	0.000	0.000	0.000
C_Facet6	−0.305	0.000	0.000	0.000	0.000

Table 4.2: SPCA loading

Variable	F1	F2	F3	F4	F5
N Facets					
N_Facet1	0.217	0.163	−0.439	0.182	0.113
N_Facet2	0.425	−0.176	0.000	0.390	0.136
N_Facet3	0.000	0.408	−0.237	0.000	0.101
N_Facet4	0.221	0.000	−0.315	0.112	0.348
N_Facet5	0.160	0.181	−0.333	0.195	0.000
N_Facet6	0.276	0.285	−0.150	0.212	0.259
E Facets					
E_Facet1	0.000	0.000	0.000	0.000	0.388
E_Facet2	0.456	0.122	−0.549	0.159	0.000
E_Facet3	0.128	0.000	−0.102	0.342	0.327
E_Facet4	0.350	0.157	−0.116	0.000	0.000
E_Facet5	0.153	0.373	0.000	0.136	0.000
E_Facet6	0.264	−0.115	−0.353	0.170	0.176
O Facets					
O_Facet1	0.261	0.219	−0.303	0.405	0.142
O_Facet2	0.220	0.353	0.000	0.000	0.331
O_Facet3	0.369	0.000	−0.246	0.000	0.385
O_Facet4	0.000	0.169	−0.610	0.236	0.000
O_Facet5	0.450	0.136	0.000	0.213	0.106
O_Facet6	0.000	0.615	0.000	0.000	0.000
A Facets					
A_Facet1	0.389	0.127	−0.155	0.157	0.000
A_Facet2	0.000	0.132	0.000	0.130	0.113
A_Facet3	0.199	0.000	0.000	0.364	0.128
A_Facet4	0.278	0.300	−0.116	0.000	0.000
A_Facet5	0.491	0.205	−0.340	0.000	0.000
A_Facet6	0.000	0.000	−0.258	0.668	0.000
C Facets					
C_Facet1	0.395	0.285	0.000	0.333	0.198
C_Facet2	0.000	0.407	−0.155	0.000	0.000
C_Facet3	0.365	0.000	−0.125	0.314	0.000
C_Facet4	0.165	0.200	−0.151	0.463	0.000
C_Facet5	0.000	0.270	0.000	0.125	0.212
C_Facet6	0.258	0.000	0.000	0.177	0.280

Table 4.3: EFA (Oblimin) loadings

5. Concluding Remarks

This thesis assessed the performance of three popular methods for uncovering hidden structures in high-dimensional datasets: principal component analysis (PCA), sparse PCA (SPCA), and exploratory factor analysis (EFA). A simulation study was conducted to identify scenarios where each method works best and when it is less suitable to provide guidelines to practitioners. An empirical application to these methods is demonstrated using the Big Five personality dataset.

In the simulation study, EFA (oblimin-rotated) consistently outperformed both PCA and SPCA in terms of reconstruction accuracy, loading recovery, and factor score correlation. EFA's strength held true with small or large sample sizes, sparse and dense settings, and simple and complex settings. This is consistent with the literature, which emphasizes EFA's ability to recover latent factors accurately through modeling shared variance and allowing for complex loading patterns [9, 28]. EFA combined with appropriate regularization technique demonstrates its ability to combine the interpretability benefits of sparsity to EFA. In the empirical analysis, EFA showed a very high similarity with PCA (with congruence coefficients up to 0.94), confirming that EFA can accurately recover well-established structures like the Big Five personality traits. These results underscore EFA's reliability for modeling underlying factors, even under challenging conditions [9].

SPCA performed well in situations where the true structure was sparse. In the simulations, SPCA achieved better identification of zero loadings and had lower false positive rates, aligning with prior research that highlights the method's ability to enhance interpretability by enforcing sparsity [40, 11]. However, its performance dropped in small samples. SPCA's performance also drops at dense setting. In the empirical analysis, SPCA showed moderate similarity with EFA (with congruence coefficients upto 0.76), reflecting a simpler and more interpretable factor structure. While SPCA's components were easier to interpret, they sometimes missed subtle relationships that EFA could capture. Therefore, SPCA is most useful when interpretability is a priority, especially in high-dimensional and sparse datasets with enough samples [11].

PCA performed well in the simulation only when the data were dense, the structure was simple, and the sample size was large. In sparse or complex setting, PCA showed higher reconstruction errors and poor factor score recovery [19]. In the empirical study, PCA closely matched EFA in recovering the Big Five structure (with congruence coefficients above 0.90), likely because the dataset's simple structure. However, PCA is less suitable for studies that require clear interpretation of factors or precise recovery of underlying latent variables [10, 20].

To practitioners, use EFA with appropriate rotation when the goal is to accurately uncover and model hidden factors, regardless of data complexity or sparsity. Furthermore, regularized EFA combines EFA's benefits by adding interpretability. We recommend SPCA particularly for high-dimensional, sparse datasets with adequate sample size where the aim is interpretation. Use PCA when the goal is to prioritize variance explanation especially in dense and simple datasets. By following these guidelines, researchers can make better choices and draw more meaningful conclusions from their data.

Finally, like many simulation studies limitations, we only relied on specific parameter choices: level of sparsity, sample size, and eigenvalue decay rate which, although representative, may not generalize to all research contexts. This could be improved by exploring more diverse conditions. Also, the empirical analysis was limited to a simple, well-structured dataset focused on the Big Five personality traits, which may not fully represent the diversity and complexity found in other social science domains – more data sets could be explored.

Acknowledgements

I would like to express my deepest gratitude to my supervisor, Prof. Gabriel Wallin, for his invaluable guidance and support throughout this research. My sincere thanks also go to my mentor, Dr. Miriam Chepleting Sitienei, and to Idene Mantho for their mentorship and encouragement.

I am especially grateful to the staff of AIMS Cameroon. I thank the Centre President, Prof. Mama Foupouagnigni, the Academic Director, and all AIMS Cameroon staff members for providing a supportive and inspiring environment. I also appreciate the lecturers, tutors and classmates with whom I have had the privilege to learn and interact during my research. Special thanks to my fellow Kenyans; Sofia and Patrick for your unwavering support.

A heartfelt thank you goes to my parents, Harrison, Veronicah, and Janefffer, for their unwavering support, love, and encouragement throughout my academic journey.

Above all, I give glory to God for His guidance, strength, and blessings.

References

[1] Hervé Abdi and Lynne J. Williams. Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(4):433–459, 2010.

[2] Kohei Adachi and Nickolay T Trendafilov. Sparse principal component analysis subject to prespecified cardinality of loadings. *Computational Statistics*, 31:1403–1427, 2016.

[3] David J Bartholomew, Martin Knott, and Irini Moustaki. *Latent variable models and factor analysis: A unified approach*. John Wiley & Sons, 2011.

[4] Michael W Browne. An overview of analytic rotation in exploratory factor analysis. *Multivariate behavioral research*, 36(1):111–150, 2001.

[5] Alexandre d'Aspremont, Laurent El Ghaoui, Michael I. Jordan, and Gert R. G. Lanckriet. A direct formulation for sparse pca using semidefinite programming. *SIAM Review*, 49(3):434–448, 2007.

[6] Conor V. Dolan, Frans J. Oort, Reinoud D. Stoel, and Jelte M. Wicherts. Testing measurement invariance in the target rotated multigroup exploratory factor model. *Structural Equation Modeling*, 16(2):295–314, 2009.

[7] Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.

[8] Ahmed Mohamed Mohamed Elsayed. Comparison between principal components and factor analysis for different data. *International Journal of Statistical Distributions and Applications*, 8(4):65–79, 2022.

[9] Leandre R. Fabrigar and Duane T. Wegener. *Exploratory Factor Analysis*. Oxford University Press, New York, 2012.

[10] Leandre R Fabrigar, Duane T Wegener, Robert C MacCallum, and Erin J Strahan. Evaluating the use of exploratory factor analysis in psychological research. *Psychological methods*, 4(3):272, 1999.

[11] Rosember Guerra-Urzola, Katrijn Van Deun, Juan C Vera, and Klaas Sijtsma. A guide for sparse pca: Model comparison and applications. *psychometrika*, 86(4):893–919, 2021.

[12] H.H. Harman. *Modern Factor Analysis*. University of Chicago Press, 1976.

[13] Trevor Hastie, Robert Tibshirani, and Martin Wainwright. Statistical learning with sparsity. *Monographs on statistics and applied probability*, 143(143):8, 2015.

[14] H.A. Hoekstra, J. Ormel, and F. De Fruyt. *NEO-PI-R/NEO-FFI: Big 5 persoonlijkheidsvragenlijst. Handleiding [Manual of the Dutch version of the NEO-PI-R/NEO-FFI]*. Swets and Zeitlinger, Lisse, The Netherlands, 2003.

[15] Robert I Jennrich. Rotation to simple loadings using component loss functions: The orthogonal case. *Psychometrika*, 69(2):257–273, 2004.

[16] Robert I Jennrich. Rotation to simple loadings using component loss functions: The oblique case. *Psychometrika*, 71:173–191, 2006.

[17] Robert I Jennrich and PF Sampson. Rotation for simple loadings. *Psychometrika*, 31(3):313–323, 1966.

[18] 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.

[19] I. T. Jolliffe and J. 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.

[20] 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.

[21] Karl G Jöreskog. A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, 34(2):183–202, 1969.

[22] Karl G Jöreskog and Arthur S Goldberger. Factor analysis by generalized least squares. *Psychometrika*, 37(3):243–260, 1972.

[23] Henry F Kaiser. The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, 23(3):187–200, 1958.

[24] Henk AL Kiers. Simplimax: Oblique rotation to an optimal target with simple structure. *Psychometrika*, 59(4):567–579, 1994.

[25] Wojtek J Krzanowski. Between-groups comparison of principal components. *Journal of the american statistical association*, 74(367):703–707, 1979.

[26] Xinyi Liu, Gabriel Wallin, Yunxiao Chen, and Irini Moustaki. Rotation to sparse loadings using lp losses and related inference problems. *arXiv preprint arXiv:2206.02263*, 2022.

[27] Urbano Lorenzo-Seva and Jos M. F. ten Berge. Tucker's congruence coefficient as a meaningful index of factor similarity. *Methodology*, 2(2):57–64, 2006.

[28] Stanley A Mulaik. *Foundations of factor analysis*. CRC press, 2009.

[29] Daniel J Mundfrom, Dale G Shaw, and Tian Lu Ke. Minimum sample size recommendations for conducting factor analyses. *International journal of testing*, 5(2):159–168, 2005.

[30] S Park, Eva Ceulemans, and Katrijn Van Deun. A critical assessment of sparse pca (research): why (one should acknowledge that) weights are not loadings. *Behavior Research Methods*, 56(3):1413–1432, 2024.

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

[32] Kristopher J Preacher and Robert C MacCallum. Exploratory factor analysis in behavior genetics research: Factor recovery with small sample sizes. *Behavior genetics*, 32:153–161, 2002.

[33] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2023.

[34] Haipeng Shen and Jianhua Z Huang. Sparse principal component analysis via regularized low rank matrix approximation. *Journal of multivariate analysis*, 99(6):1015–1034, 2008.

[35] John Spores. *Psychological assessment and testing: a clinician's guide*. Routledge, 2022.

[36] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288, 1996.

[37] Ledyard R. Tucker. A method for synthesis of factor analysis studies. *Personnel Psychology*, 5(3):233–245, 1951.

[38] Peter Whittle. On principal components and least square methods of factor analysis. *Scandinavian Actuarial Journal*, 1952(3-4):223–239, 1952.

[39] Allen Yates. *Multivariate exploratory data analysis: A perspective on exploratory factor analysis*. Suny Press, 1987.

[40] Hui Zou, Trevor Hastie, and Robert Tibshirani. Sparse principal component analysis. *Journal of computational and graphical statistics*, 15(2):265–286, 2006.