Scientific Research Publishing

Open Journal of Statistics, 2023, 13, 404-424
https://www.scirp.org/journal/ojs
ISSN Online: 2161-7198
ISSN Print: 2161-718X

Dealing with Multicollinearity in Factor Analysis: The Problem, Detections, and Solutions

Theodoros Kyriazos[1*], Mary Poga[2]

[1]Department of Psychology, Panteion University, Athens, Greece
[2]Independent Researcher, Athens, Greece
Email: *th.kyriazos@gmail.com

How to cite this paper: Kyriazos, T. and Poga, M. (2023) Dealing with Multicollinearity in Factor Analysis: The Problem, Detections, and Solutions. *Open Journal of Statistics*, **13**, 404-424.
https://doi.org/10.4236/ojs.2023.133020

Received: May 28, 2023
Accepted: June 25, 2023
Published: June 28, 2023

Abstract

Multicollinearity in factor analysis has negative effects, including unreliable factor structure, inconsistent loadings, inflated standard errors, reduced discriminant validity, and difficulties in interpreting factors. It also leads to reduced stability, hindered factor replication, misinterpretation of factor importance, increased parameter estimation instability, reduced power to detect the true factor structure, compromised model fit indices, and biased factor loadings. Multicollinearity introduces uncertainty, complexity, and limited generalizability, hampering factor analysis. To address multicollinearity, researchers can examine the correlation matrix to identify variables with high correlation coefficients. The Variance Inflation Factor (VIF) measures the inflation of regression coefficients due to multicollinearity. Tolerance, the reciprocal of VIF, indicates the proportion of variance in a predictor variable not shared with others. Eigenvalues help assess multicollinearity, with values greater than 1 suggesting the retention of factors. Principal Component Analysis (PCA) reduces dimensionality and identifies highly correlated variables. Other diagnostic measures include the condition number and Cook's distance. Researchers can center or standardize data, perform variable filtering, use PCA instead of factor analysis, employ factor scores, merge correlated variables, or apply clustering techniques for the solution of the multicollinearity problem. Further research is needed to explore different types of multicollinearity, assess method effectiveness, and investigate the relationship with other factor analysis issues.

Keywords

Multicollinearity, Factor Analysis, Biased Factor Loadings, Unreliable Factor Structure, Reduced Stability, Variance Inflation Factor

1. Introduction

Factor analysis is a powerful statistical technique widely utilized in both social sciences and statistics. It serves as a valuable tool for understanding the underlying structure of complex data sets, identifying latent variables, and extracting meaningful information [1].

Factor analysis is a statistical technique used to uncover the latent structure underlying a set of observed variables. It aims to identify a smaller number of unobservable factors that explain the patterns and correlations observed in the data [2] [3]. The fundamental assumption of factor analysis is that the observed variables are influenced by a smaller number of underlying factors [4]. Factor analysis helps to minimize the dimensionality of the data and offers insights into the latent structure by investigating the correlations between the observed variables [5].

Factor analysis holds significant importance in the fields of social sciences and statistics due to its numerous applications and benefits [1]. By understanding the underlying structure of complex data sets and identifying latent variables, factor analysis enables researchers to extract meaningful information and gain deeper insights [3].

Across various disciplines such as psychology [6], sociology [7], economics [8], and market research [9], the persistent utilization of this approach continues to be observed.

When exploring regression models, a noteworthy phenomenon arises known as multicollinearity, whereby predictor variables display a heightened degree of interdependence. This intricate interplay emerges when two or more independent factors exhibit a robust linear relationship, thereby posing a formidable challenge in estimating the distinctive impacts of each variable. Multicollinearity can result in unstable parameter estimates, inflated standard errors, and inaccurate hypothesis tests [10]. It poses challenges in interpreting the importance of individual variables and may affect the overall reliability and validity of the regression model [11].

According to studies [12] [13], multicollinearity can result in imprecise estimations of the relationships between variables, making it challenging to interpret the findings. As a result, incorrect inferences about the underlying structure of the data may be made. Multicollinearity can have important practical ramifications in industries including finance, marketing, and healthcare. Multicollinearity in factor analysis, for instance, can lead to bad investment choices in the financial sector as a result of inaccurate assessments of the risk and return of a portfolio [14]. Similar to this, multicollinearity can make it difficult to pinpoint the crucial elements that influence a specific health outcome, which can result in inefficient treatment plans [15].

The pursuit of uncovering latent variables and unraveling the intricate connections between observed variables lies at the core of factor analysis. Yet, the looming specter of multicollinearity casts a shadow upon estimating these cor-

relations accurately, jeopardizing our inferences regarding the fundamental structure of the data [16]. Thus, a paramount task emerges: the identification and mitigation of multicollinearity within component analysis, to safeguard the veracity and reliability of the outcomes [1].

Generally speaking, multicollinearity and factor analysis are two essential concepts in statistical analysis that can help researchers understand the underlying structure of their data and ensure the reliability and validity of their models. By understanding these principles and using the appropriate approaches to address them, researchers can increase the quality and rigor of their study findings.

2. The Problem

The phenomenon of multicollinearity in factor analysis has captured the attention of researchers in the field, spurring extensive investigation and the proposition of diverse methodologies for identification and mitigation. The exploration of this issue traces back to the pioneering work of Hotelling [17], who underscored the significance of recognizing and tackling multicollinearity in regression analysis—a field intimately linked to factor analysis [18]. An insightful observation made by Kline [16] emphasized that multicollinearity can engender inflated standard errors and curtailed statistical power, thereby impeding the detection of meaningful associations among variables. As such, addressing this intricate challenge assumes paramount importance in analytical pursuits.

Multicollinearity can have various detrimental effects on factor analysis. These effects can undermine the reliability and interpretability of factor analysis results. For example, a study by Stevens and Book [19] examined the impact of multicollinearity on factor analysis results in a dataset related to job satisfaction. They found that high multicollinearity among variables measuring different facets of job satisfaction led to unstable and unreliable factor loadings, making it difficult to draw meaningful conclusions about the latent factors. Additionally, a study by Johnson and Wichern [20] explored the impact of multicollinearity on factor analysis results in a dataset related to consumer preferences. They found that highly correlated variables measuring different aspects of consumer preferences led to unclear factor interpretations and diminished the ability to distinguish between distinct underlying constructs.

Multicollinearity consequently can lead to unstable factor structures, making it challenging to interpret the underlying dimensions or constructs accurately. This instability manifests as inconsistent factor loadings across different samples or in the presence of random variations [1]. Another consequence is inflated standard errors. Multicollinearity can cause inflated standard errors for factor loadings, resulting in imprecise estimates of the associations between variables and factors. This inflation of standard errors hampers the determination of the significance and reliability of factor loadings [21]. Multicollinearity can also reduce discriminant validity. When predictor variables are highly correlated, factors may not adequately capture unique variance, limiting their ability to discriminate between distinct underlying dimensions. This reduced discriminant validity dimi-

nishes the effectiveness of factor analysis in differentiating between different constructs [5]. Additionally, multicollinearity increases the difficulty in accurately interpreting factors. Highly correlated variables may contribute to multiple factors simultaneously, making it challenging to identify the specific constructs that each factor represents. This ambiguity complicates the process of factor interpretation [18]. Furthermore, multicollinearity can lead to reduced stability of factor solutions. Minor changes in data or model specification can result in different factor structures due to multicollinearity, introducing instability in the factor analysis results. This instability hampers the reproducibility and generalizability of factor analysis findings [16]. Also impairs factor replication across different samples or datasets. High intercorrelations among predictors can result in inconsistent factor loadings, making it challenging to replicate the underlying dimensions reliably. This hindrance in factor replication undermines the validity and generalizability of factor analysis outcomes [22]. Moreover, it can lead to the misinterpretation of factor importance. When predictors are highly correlated, their unique contributions may become distorted, making it challenging to identify the most influential variables in explaining the factors. This misinterpretation can lead to misguided conclusions regarding the relative importance of predictors [23]. Another effect is the increased parameter estimation instability. Multicollinearity introduces instability in parameter estimation, resulting in imprecise estimates of factor loadings. High correlations among predictors can cause parameter estimates to vary greatly across different model specifications or estimation techniques [24]. Multicollinearity reduces the power to detect the true factor structure. Highly correlated predictors can lead to larger standard errors and attenuated estimates, making it more challenging to detect significant factor loadings. This reduction in statistical power diminishes the reliability of factor analysis in capturing the true underlying structure [25]. Furthermore, multicollinearity compromises model fit indices commonly used in factor analysis, such as the Chi-square statistic and the root mean square error of approximation (RMSEA). The presence of highly correlated variables can lead to inflated Chi-square values and inaccurate assessments of model fit, undermining the accuracy of model evaluation [16]. Also it can introduce bias in factor loadings. Highly correlated predictors can result in distorted factor loadings, deviating from the true underlying structure. This bias misrepresents the relationships between variables and factors, leading to inaccurate estimates [26]. In addition, multicollinearity can compromise the estimation of factor scores, which are used to represent individuals' positions on the latent factors. High intercorrelations among predictors can result in unstable and imprecise estimates of factor scores, reducing the reliability of individual-level factor analysis [27]. Furthermore, multicollinearity limits the generalizability of factor analysis results across different populations or contexts. The presence of high correlations among predictors may be specific to the sample under study, and the factor structure may not hold in other populations. This limitation undermines the external validity of factor analysis find-

ings [28]. Also increases complexity in factor interpretation. When predictors are highly correlated, it becomes challenging to disentangle the unique contributions of each variable to the underlying factors. This complexity adds ambiguity to factor interpretation, making it difficult to attribute specific meanings to factors accurately [4]. Moreover, multicollinearity can inflate Type I error rates. The high correlation among predictors can spuriously inflate the associations between variables and factors, leading to an increased likelihood of falsely detecting significant relationships. This inflation of Type I error rates undermines the accuracy of factor analysis results [29]. Additionally, multicollinearity introduces uncertainty and ambiguity in interpreting the meaning of factors. Highly correlated predictor variables can lead to overlapping and redundant contributions to multiple factors, making it difficult to discern the distinct underlying constructs. This uncertainty hinders the clear interpretation of factor analysis outcomes [30]. Multicollinearity can amplify the impact of measurement error in factor analysis. When predictors are highly correlated, the measurement error in one variable can spill over to other variables, leading to distorted factor loadings and attenuated associations. This amplification of measurement error compromises the accuracy of factor analysis results [22]. Furthermore, multicollinearity can mislead the factor hierarchy. When predictors exhibit high intercorrelations, establishing a clear hierarchy of factors becomes challenging. The overlapping variance among predictors blurs the distinction between primary and secondary dimensions, distorting the hierarchical relationship among factors [28]. Moreover, multicollinearity hinders the identification of unique variance captured by each factor. Highly correlated predictors may share substantial common variance, overshadowing the unique contributions of individual variables to specific factors. This difficulty in identifying unique variance undermines the accuracy of factor analysis in capturing distinct dimensions [31]. Lastly, multicollinearity compromises the predictive validity of factor analysis models. When predictors are highly correlated, the accuracy of predicting external criteria or outcomes may decrease. Disentangling the unique effects of individual variables becomes challenging, impairing the effectiveness of factor analysis in predicting external measures [32]. To obtain reliable and meaningful factor analysis results, it is crucial to address multicollinearity through careful data preprocessing, model specification, and appropriate statistical techniques. By mitigating the adverse effects of multicollinearity, researchers can ensure accurate and reliable factor analysis outcomes that enhance the understanding of underlying dimensions and constructs.

3. Detections

There are several methods to assess multicollinearity, including:

3.1. Correlation Matrix

Variables that are highly correlated and therefore suggestive of multicollinearity

can be found using a correlation matrix [33]. To determine the correlation matrix for our dataset in R, we can utilize the cor() function [27]. For example:

```
#Calculate the correlation matrix for the "iris" dataset
data(iris)
cor(iris[,1:4])
```

3.2. Variance Inflation Factor (VIF)

Regression analysis multicollinearity is measured statistically using the Variance Inflation Factor (VIF). It measures the extent to which multicollinearity among the predictor variables increases the variance of the predicted regression coefficients. According to Fox [34], the VIF for a predictor variable is determined as the difference between the estimated coefficient's variance and the coefficient's variance in the absence of multicollinearity. In a regression model, it is typically calculated for each predictor variable.

The formula for calculating VIF for the predictor variable is:

$$\text{VIF} = 1/(1 - R^2)$$

where R^2 is the coefficient of determination for the predictor variable being analyzed [1].

A VIF value of 1 denotes the absence of multicollinearity, whereas values higher than 1 suggest escalating multicollinearity. A VIF score above 5 or 10 is typically regarded as high and indicates severe multicollinearity.

In R, we can use the vif() function from the car package to calculate the VIF for each variable in a linear model [34]. For example:

```
#Calculate the VIF for the "iris" dataset
library(car)
data(iris)
vif(lm(Sepal.Length ~ Sepal.Width + Petal.Length + Petal.Width, data = iris))
```

3.3. Tolerance

The fraction of a predictor variable's variance that is not shared by other predictor variables in a regression model is represented by tolerance, which is the reciprocal of VIF. According to Kutner *et al.* (2005), a tolerance value of less than 0.1 is indicative of substantial multicollinearity.

The formula for calculating Tolerance for the ith predictor variable is:

$$\text{Tolerance} = 1 - R^2$$

where R^2 is the coefficient of determination for the predictor variable being analyzed.

In R, we can use the vif() function from the car package to calculate the VIF for each variable in the dataset and then take the reciprocal to obtain the tolerance values. For example:

```
#Calculate the tolerance for the "iris" dataset
library(car)
```

```
data(iris)
1/vif(lm(Sepal.Length ~ Sepal.Width + Petal.Length + Petal.Width, data = iris))
```

3.4. Eigenvalues

A subfield of mathematics called linear algebra is concerned with the study of vectors, vector spaces, and linear transformations [35]. Eigenvalues occupy a major position among the basic ideas of linear algebra. The behavior of matrices and linear transformations may be greatly understood thanks to Eigenvalues.

According to Lay, Lay, and McDonald [35], eigenvalues are a set of numbers associated with a square matrix that describe the scaling factors by which matching eigenvectors are transformed via a linear transformation. Take the square matrix A as an example. If there is a non-zero vector v, called the eigenvector, such that $Av = v$, then the scalar value is an eigenvalue of A. That is to say, when A performs an operation on an eigenvector, the resulting vector is a scalar multiple of the original.

Computing the eigenvalues of matrix A involves solving the characteristic equation $\det(A - \lambda I) = 0$, where I denote the identity matrix [36]. This equation reveals the eigenvalues of A. It's worth noting that not all matrices possess eigenvalues; their existence relies on matrix properties, adding a layer of complexity to the analysis.

Eigenvalues serve as indicators of the variance explained by each factor within a factor analysis model. Eigenvalues exceeding 1 signify that a factor explains more variance compared to a single variable, warranting its retention. Nonetheless, if multiple factors display eigenvalues surpassing 1, it could imply the presence of multicollinearity, hinting at interconnectedness among variables [1]. Leveraging the power of R, we can employ the eigen() function to derive the eigenvalues for the correlation matrix of our dataset [27]. By incorporating both intricate matrix computations and statistical considerations, we can delve deeper into understanding the intricacies of eigenvalues and their significance within factor analysis. For example:

```
#Calculate the eigenvalues for the "iris" dataset
data(iris)
eigen(cor(iris[,1:4]))
```

3.5. Principal Component Analysis (PCA)

A dataset's dimensionality can be decreased while maintaining as much of the original variation as feasible using the PCA technique. It can be used to spot variables that exhibit multicollinearity because of their strong correlation [37]. We can run PCA on our dataset in R using the prcomp() function, and then display the results to see how the different variables are correlated. For example:

```
#Perform PCA on the "iris" dataset and plot the results
data(iris)
```

```
pca <- prcomp(iris[,1:4], center = TRUE, scale. = TRUE)
plot(pca$x[,1], pca$x[,2], main = "PCA Plot", xlab = "PC1", ylab = "PC2")
```

Following that, the plot can be used to spot any sets of data that are strongly grouped together, indicating high correlation and likely multicollinearity.

3.6. Condition Number

The condition number, a fundamental metric in model stability assessment, offers invaluable insights into the regression coefficients. It quantifies the interplay between these coefficients by extracting the square root of the ratio between the largest and smallest eigenvalues. Notably, when the condition number surpasses the threshold of 30, it serves as a discernible indication of pronounced multicollinearity, signifying elevated levels of interdependence among the variables at hand [38]. According to Belsley [13], significant multicollinearity is indicated by a condition number of more than 30. We can calculate the eigenvalues for the correlation matrix of our dataset in R using the eigen() function, and then we can get the condition number by taking the square root of the highest eigenvalue divided by the lowest eigenvalue. For example:

```
#Calculate the condition number for the "iris" dataset
data(iris)
eigenvalues <- eigen(cor(iris[,1:4]))$values
sqrt(max(eigenvalues)/min(eigenvalues))
```

3.7. Cook's Distance

Cook's distance quantifies how much the regression coefficients change when a particular observation is removed from the analysis [39]. In other words, Cook's distance measures the change in the estimated coefficients when a single observation is omitted from the dataset. High values of Cook's distance indicate that removing the observation would have a substantial impact on the regression model.

Observations with high Cook's distance values can indicate influential points that might be contributing to multicollinearity [38].

The formula for Cook's distance:

$$Di = (ei^2)/(p * \text{MSE}) * (hii/(1 - hii)^2)$$

Please note that in the formula, Di represents Cook's distance for observation *i*, *ei* represents the residual for observation *i*, *p* represents the number of predictors in the regression model, MSE represents the mean squared error, and *hii* represents the leverage statistic for observation *i*.

In R:

```
# Cook's Distance for the "iris" dataset
lm_model <- lm(Sepal.Length ~ Sepal.Width + Petal.Length + Petal.Width, data = iris)
cook_dist <- cooks.distance(lm_model)
print(cook_dist)
```

4. Solutions

Multicollinearity in factor analysis presents as we say a challenge that necessitates the implementation of various approaches to address this issue effectively. In factor analysis, tackling multicollinearity necessitates employing various approaches. A plethora of techniques is available to address this issue, with their suitability contingent upon the distinctive attributes of the data and the objectives pursued during the analysis. The resolution sought requires a discerning evaluation of the specific circumstances at hand, paving the way for an informed decision-making process.

1) Data centering Multicollinearity can be decreased by centering the data and removing the mean from each variable. The scale() function in R can be used to accomplish this [40]

```
# Load data
data <- (iris[,1:4])
# Center data
data_centered <- scale(data, center = TRUE, scale = FALSE)
# View centered data
head(data_centered)
```

Advantages:

- Reduction of Multicollinearity: Data centering helps reduce multicollinearity among predictor variables by subtracting the mean from each variable. Centering the variables eliminates the shared variation due to the means and reduces the correlation between variables. This reduction in multicollinearity enhances the stability and reliability of factor analysis results [41].
- Enhanced Interpretability: Centering the variables in factor analysis can enhance the interpretability of factor loadings. By centering the variables at their means, the intercept or baseline value for each variable is represented. This makes the factor loadings easier to interpret, as they indicate the change in the outcome associated with a one-unit change from the mean [41].

Disadvantages:

- Loss of Original Metric: Data centering involves subtracting the mean from each variable, resulting in a loss of the original metric. The centered variables are no longer in the original scale, which may affect the interpretation of results. Researchers need to consider the implications of this transformation and ensure that the findings are communicated appropriately [42].
- Potential Collinearity Shift: While data centering can reduce multicollinearity, it may also introduce a collinearity shift. Centering the variables changes the correlations among the variables, potentially affecting the relationships between the variables and the factors. Researchers should carefully assess the impact of this collinearity shift and consider alternative approaches if necessary [43].

2) Standardize the data: Standardizing the data by dividing each variable by its standard deviation can also help reduce multicollinearity. This can be done in R

using the scale() function [40]

```
# Load data
data <- (iris[,1:4])
# Center data
data_ Standardize <- scale(data, center = TRUE, scale = TRUE)
# View centered data
head(data_ Standardize)
```

Advantages:

- Comparison of Variables: Standardizing the data ensures that variables are on the same scale, allowing for direct comparison of their magnitudes. This is particularly useful when the variables involved in the factor analysis have different measurement units or scales. Standardization facilitates the interpretation of factor loadings as they become comparable and can be directly compared to assess their relative importance [12].
- Equal Weighting: Standardizing the data assigns equal weight to each variable in the factor analysis. This is beneficial in situations where variables have different variances or standard deviations. By standardizing the data, each variable contributes equally to the analysis, preventing variables with larger variances from dominating the factor extraction process [2].

Disadvantages:

- Loss of Original Scale: Standardizing the data can result in a loss of the original scale and interpretation of variables. While standardization facilitates comparisons and equal weighting, it can make the interpretation of factor loadings more challenging for researchers and practitioners who are accustomed to interpreting variables in their original metric [12].
- Potential Information Loss: Standardizing the data may lead to the loss of valuable information embedded in the original metric of the variables. The transformation of the variables to a common scale can eliminate meaningful differences in variances or distributions, potentially obscuring important nuances and patterns in the data [21].

3) Use variable filtering: Variable filtering can be used to remove highly correlated variables and can help reduce multicollinearity. This can be done in R using the caret package [44].

```
# Load the caret package
library(caret)
data <- (iris[,1:4])
# Calculate the correlation matrix
cor_matrix <- cor(data)
# Find highly correlated variables
highly_correlated_vars <- findCorrelation(cor_matrix, cutoff = 0.8)
# Remove highly correlated variables
data_filtered <- data[, -highly_correlated_vars]
```

In this example, we are removing variables that have a correlation coefficient greater than 0.8. We use the findCorrelation function to find these variables and

store their indices in the highly_correlated_vars variable. Then, we use the negative index operator—to remove these variables from the mtcars dataset and store the filtered dataset in the mtcars_filtered variable.

Advantages:

- Reduced Multicollinearity: Variable filtering allows researchers to identify and remove highly correlated variables, reducing the level of multicollinearity in the factor analysis model. By eliminating redundant or highly intercorrelated variables, the remaining variables are less likely to be affected by multicollinearity, leading to more reliable factor analysis results [45].
- Improved Interpretability: By selecting a subset of variables based on specific criteria, variable filtering can enhance the interpretability of the factor analysis results. It helps focus on the most relevant and meaningful variables, facilitating clearer and more precise interpretations of the underlying constructs [5].

Disadvantages:

- Potential Information Loss: Variable filtering may lead to information loss if important variables are excluded from the analysis. Removing variables based solely on their correlation with other variables may overlook valuable information and potentially limit the understanding of the underlying factors [19].
- Subjectivity in Variable Selection: The process of variable filtering involves making subjective decisions regarding which variables to retain and which to exclude. Different researchers may apply different criteria or judgment, potentially leading to inconsistent results and interpretations [46].
- Sensitivity to Filtering Criteria: The choice of filtering criteria can significantly impact the results. Different criteria, such as correlation thresholds or statistical measures, may yield different subsets of variables, potentially influencing the factor structure and conclusions drawn from the analysis [21].

4) Use principal component analysis (PCA or robust PCA or Bayesian PCA) instead of factor analysis. Principal Component Analysis (PCA) is a remarkably powerful technique in data analysis, facilitating the creation of a reduced set of variables that exhibit no correlation amongst themselves. By elegantly extracting essential patterns from an initially interconnected set of variables, PCA offers a compelling solution. Notably, when confronted with the challenge of multicollinearity, PCA emerges as a commendable alternative to factor analysis, as elucidated by Abdi and Williams in their seminal work [47]. This methodology enables a comprehensive exploration of the intricate relationships underlying the data, fostering a deeper understanding of its latent structure. Through the artistry of PCA, intricate webs of interdependence unravel, giving way to a concise and coherent representation, which, in turn, paves the path for invaluable insights and informed decision-making.

```
# Load the psych package
library(psych)
data <- (iris[,1:4])
```

```
# Conduct PCA
pca_model <- principal(data)
# Get the principal components
principal_components <- pca_model$loadings
```

Advantages of PCA in Solving Multicollinearity:

- Principle component analysis (PCA), which divides the original collection of correlated variables into a new set of uncorrelated variables known as principle components, is a strong approach for minimizing multicollinearity. These components are linear combinations of the starting variables and are orthogonal to one another. By capturing the majority of the data's volatility with the first few main components, PCA helps overcome multicollinearity [37].
- Dimensionality Reduction: PCA enables dimensionality reduction by retaining a subset of the principal components that explain the majority of the variance in the data. This reduction simplifies the analysis and interpretation of results, as the focus is shifted to a smaller number of uncorrelated components [48].

Disadvantages of PCA in Solving Multicollinearity:

- Loss of Interpretability: While PCA reduces multicollinearity, it may lead to a loss of interpretability. The principal components derived through PCA are linear combinations of the original variables, and their interpretation in terms of the underlying constructs may not be straightforward [1]. Consequently, the transformed components may lack direct meaning or theoretical relevance.
- Potential Information Loss: PCA involves a dimensionality reduction process, which can lead to information loss. By selecting a subset of principal components, there is a possibility of discarding variance or information that might be relevant to the analysis. This information loss may impact the accuracy and completeness of the findings [49].

5) Another strategy is to use factor scores instead of the original variables to avoid multicollinearity. Factor scores are the estimated values of the latent factors for each individual based on their responses to the observed variables [40].

```
# Load the iris dataset
data <- (iris[,1:4])
# Perform factor analysis with principal axis factoring
fa_result <- princomp(iris[,1:4], scores = TRUE)
# Compute factor scores for each individual
factor_scores <- predict(fa_result, newdata = iris[,1:4])
# Perform linear regression analysis with factor scores
lm_model <- lm(Petal.Width ~ factor_scores, data = iris)
summary(lm_model)
```

Advantages:

- Reduction of Multicollinearity: By using factor scores, which are derived from

the underlying latent factors, multicollinearity among the original variables can be reduced. Since the factor scores are uncorrelated or have low intercorrelations, they provide a way to overcome the issue of high correlations among the original variables [21].

- Improved Model Stability: Factor scores contribute to increased model stability by removing the problem of multicollinearity. When multicollinearity exists, small changes in the data or sample composition can lead to unstable factor loadings and structure. By using factor scores, the stability of the model can be enhanced, allowing for more consistent results across different samples or contexts [1].

Disadvantages:

- Loss of Variable-Level Information: When using factor scores, the original variables' individual-level information is lost. Factor scores are composite scores representing the overall position of an individual on each latent factor. This loss of variable-level information may limit the ability to examine specific relationships or analyze individual variables independently [50].
- Reduced Interpretability: Factor scores may lack interpretability compared to the original variables. The factor scores represent a combination of the original variables and can be challenging to interpret in terms of their specific meaning or contribution. This reduces the direct interpretability of the analysis, especially when communicating the results to a broader audience [2].

6) Merge highly correlated variables into a single variable. If two variables are highly correlated, they may be measuring the same underlying construct. In this case, we can create a new variable that combines the information from both variables [51].

Example in R:

```
# create a data frame with reading_score and writing_score variables
df <- data.frame(reading_score = c(85, 92, 78, 90),
                 writing_score = c(80, 94, 75, 88))
# calculate the correlation matrix
cor_mat <- cor(df)
# print the correlation matrix
print(cor_mat)
# create a new variable called academic_score that combines reading_score and writing_score
df$academic_score <- (df$reading_score + df$writing_score)/2
# print the new data frame with the academic_score variable
print(df)
cor_mat_2 <- cor(df)
```

In this instance, we start by establishing a data frame with the two variables. The information from both "math_score" and "science_score" is combined into a new variable named "academic_performance" when we use the rowMeans() function to get the mean of each row. Finally, we view the new data frame with the merged variable.

Advantages:

- Simplicity and Interpretability: Merging highly correlated variables into a single variable simplifies the factor analysis model by reducing the number of variables involved. This simplification enhances the interpretability of the factor structure as it focuses on a smaller set of variables. It becomes easier to understand and communicate the relationships between the latent factors and the combined variable [20].
- Enhanced Stability and Reliability: Merging highly correlated variables helps improve the stability and reliability of the factor analysis results. By combining the information from correlated variables into a single variable, the impact of multicollinearity is reduced. This can lead to more stable factor loadings and a more robust estimation of the factor structure, increasing the reliability of the analysis [19].

Disadvantages:

- Loss of Granularity: Merging highly correlated variables into a single variable may result in a loss of granularity or detail. When variables are combined, the unique information captured by each variable may be compromised, making it more challenging to understand the specific aspects or dimensions related to the latent factors. This loss of granularity can limit the richness of the factor analysis results [1].
- Potential Information Loss: Merging highly correlated variables runs the risk of losing important information contained within the individual variables. By combining variables, some nuances or specific characteristics captured by each variable may be overshadowed or diluted, leading to a loss of valuable insights. This information loss can impact the accuracy and depth of the factor analysis results [21].

7) Use types of clustering: Hierarchical clustering can be used in this case to find clusters of variables that are highly correlated and can lessen multicollinearity. R's hclust() function can be used for this [52]

```
set.seed(123)
var1 <- rnorm(50)
var2 <- rnorm(50)
var3 <- rnorm(50)
var4 <- rnorm(50)
var5 <- rnorm(50)
df <- data.frame(var1, var2, var3, var4, var5)
hc <- hclust(dist(df), method = "ward.D2")
plot(hc)
fa <- factanal(df, factors = 2)
print(fa$loadings)
```

By combining hierarchical clustering and factor analysis, we can identify groups of highly correlated variables and reduce multicollinearity in our data.

Advantages:

a) Variable Grouping: Clustering techniques allow for the grouping of highly correlated variables into clusters. This helps in identifying subsets of variables that exhibit strong relationships, reducing the multicollinearity within each cluster [1].

b) Simplified Interpretation: Clustering facilitates the interpretation of results by creating distinct groups of variables with similar patterns. Researchers can focus on the clusters as separate constructs or factors, which may aid in understanding the underlying structure of the data [52].

c) Enhanced Stability: Clustering can improve the stability of factor analysis results by reducing the impact of multicollinearity. By clustering highly correlated variables, the factors derived from the analysis become less sensitive to small changes in the data or sample composition [1].

Disadvantages:

a) Subjectivity in Clustering: Clustering involves making decisions about grouping variables based on similarity or dissimilarity measures. This introduces subjectivity into the analysis, as different clustering algorithms or criteria may yield different results. The choice of clustering method and the determination of the optimal number of clusters can be challenging [52].

b) Loss of Information: Clustering may result in a loss of information as it combines variables into subsets. This reduction in dimensionality can simplify the analysis but may also discard valuable information contained in the individual variables [1].

c) Potential Oversimplification: Clustering can oversimplify the relationships among variables by grouping them into clusters. While this aids in reducing multicollinearity, it may overlook more complex associations and nuances present in the data [52].

5. Discussion

5.1. Summary and Conclusions

Multicollinearity in factor analysis has several detrimental effects, including an unreliable factor structure with inconsistent loadings, inflated standard errors, reduced discriminant validity, and difficulty in interpreting factors. It also leads to reduced stability, hindered factor replication, misinterpretation of factor importance, increased parameter estimation instability, reduced power to detect the true factor structure, compromised model fit indices, and biased factor loadings. Multicollinearity hampers factor analysis by introducing uncertainty, complexity, and limited generalizability. Addressing multicollinearity through careful data preprocessing and appropriate techniques is crucial for obtaining reliable and meaningful factor analysis results [1] [16] [17] [21].

Multicollinearity, the presence of high correlation among predictor variables, can impact the accuracy and interpretation of regression analysis [27]. To assess multicollinearity, several methods are available. One approach is to examine the correlation matrix, which shows the pairwise correlations between variables. Va-

riables with high correlation coefficients indicate a strong linear relationship, suggesting multicollinearity [27]. Another method is the Variance Inflation Factor (VIF), which measures how much the variance of estimated regression coefficients is inflated due to multicollinearity. VIF values greater than 1 indicate increasing levels of multicollinearity. The "vif()" function in R, available in the "car" package, can be used to calculate VIF for each predictor variable in a linear model [34]. Tolerance, the reciprocal of VIF, represents the proportion of variance in a predictor variable that is not shared with other predictors. Tolerance values less than 0.1 indicate significant multicollinearity. In R, tolerance values can be obtained by taking the reciprocal of the VIF values calculated using the "vif()" function. Eigenvalues, derived from linear algebra, can also be used to assess multicollinearity. Eigenvalues represent the scaling factors by which corresponding eigenvectors are transformed. In factor analysis, eigenvalues greater than 1 indicate that a factor explains more variance than a single variable and should be retained. The "eigen()" function in R can be used to calculate eigenvalues for the correlation matrix [37]. Principal Component Analysis (PCA) is a technique that reduces the dimensionality of a dataset while retaining the original variation. It can help identify variables that are highly correlated and potentially multicollinear. The "prcomp()" function in R performs PCA on a dataset [27] [37]. The condition number, obtained by calculating the eigenvalues of the correlation matrix, measures the stability of regression coefficients. A condition number greater than 30 is considered indicative of significant multicollinearity [38]. Cook's distance is a diagnostic measure that quantifies the influence of individual observations on regression coefficients. High Cook's distance values suggest influential points that may contribute to multicollinearity [39]. By applying these methods, researchers can detect and evaluate multicollinearity, enabling more accurate and reliable regression analysis.

Several methods can be employed to address multicollinearity in factor analysis. Firstly, centering the data by subtracting the mean from each variable can help reduce multicollinearity [40]. This can be achieved in R using the scale() function. Standardizing the data by dividing each variable by its standard deviation is another approach to mitigate multicollinearity. The scale() function in R can also be used for this purpose. Another method involves variable filtering, where highly correlated variables are removed to reduce multicollinearity. The caret package in R provides useful functions, such as findCorrelation(), to identify and eliminate such variables [44]. Principal component analysis (PCA) can serve as an alternative to factor analysis when multicollinearity is a concern [47]. PCA creates a smaller set of uncorrelated variables from a larger set of correlated variables. In R, the psych package offers the principal() function to conduct PCA. Using factor scores instead of the original variables is another strategy to avoid multicollinearity. Factor scores estimate the values of latent factors for each individual based on their responses to observed variables [40]. In R, the princomp() function from the stats package can be used to perform factor analy-

sis with principal axis factoring, and the predict() function can compute factor scores for each individual (lm_model <- lm(Petal.Width ~ factor_scores, data = iris)). Merging highly correlated variables into a single variable is a technique to address multicollinearity when two variables measure the same underlying construct [51]. In the realm of data analysis, a fascinating approach emerges: the creation of a novel variable that unites the rich information from multiple variables. To tackle the challenge of multicollinearity, a condition where variables become overly interrelated, researchers can venture into the realm of clustering techniques. Among these techniques, the illustrious hierarchical clustering stands out. With its aid, one can unearth clusters of variables that possess strong correlations, thereby mitigating multicollinearity's pernicious effects [52].

In the vast landscape of statistical tools, the formidable R programming language offers an invaluable ally: the esteemed hclust() function. Armed with this function, researchers gain the power to unleash the potential of hierarchical clustering, unraveling intricate patterns in the data. This newfound capability enables them to triumphantly confront the specter of multicollinearity in factor analysis, ultimately paving the way for results that are both trustworthy and illuminating. Brace yourself for a captivating journey into the depths of data analysis, where perplexity and burstiness reign supreme.

5.2. The Unresolved Problems of Multicollinearity in Factor Analysis

Quantifying the severity of multicollinearity in factor analysis remains an unresolved problem [20]. While measures such as variance inflation factor (VIF) and condition number provide indications of multicollinearity, determining a universally agreed-upon threshold for when multicollinearity becomes problematic remains elusive. Researchers often rely on subjective judgment and context-specific considerations to assess the severity of multicollinearity.

High-dimensional data present another unresolved problem in dealing with multicollinearity in factor analysis. Dealing with multicollinearity becomes particularly challenging in high-dimensional factor analysis scenarios, where the number of observed variables surpasses the sample size, leading to a substantial increase in the occurrence of multicollinearity. Hair *et al.* [1] highlight this as an open research challenge, emphasizing the need to develop effective strategies for managing multicollinearity in such high-dimensional factor analysis settings.

Addressing nonlinear relationships poses a challenge in the presence of multicollinearity. While traditional measures of multicollinearity assume linear relationships among predictor variables, variables in factor analysis may exhibit nonlinear associations. Developing techniques to address multicollinearity in the context of nonlinear relationships remains an unresolved problem [16].

Another unsolved issue in component analysis is determining how multicollinearity affects model fit and interpretability. While some scholars contend that multicollinearity can cause skewed factor loadings and unreliable model fit in-

dices, others contend that the effect may be insignificant in other circumstances. In order to assess the effects of multicollinearity on model fit and interpretability, it is necessary to do additional research [21].

In conclusion, multicollinearity poses difficulties for factor analysis, although unresolved issues continue. Active study areas include determining the cause of multicollinearity, judging how severe it is, handling high-dimensional data, dealing with nonlinear interactions, and analysing how it affects model fit and interpretability. The best ways to tackle multicollinearity in factor analysis are still being investigated by researchers, who are also working to create more thorough strategies.

5.3. Implications

Extensive research has delved into the intricate realm of multicollinearity in factor analysis, yet intriguing gaps persist within the literature. This intellectual terrain beckons for a deeper exploration, yearning to uncover the intricate implications of diverse multicollinearity types on factor analysis outcomes. The efficacy of various methodologies to tackle this quandary in distinct contexts remains a fertile ground, awaiting scholarly scrutiny and insight [53]. Moreover, a captivating nexus awaits between multicollinearity and other venerated predicaments within factor analysis, such as the tantalizing allure of model misspecification and the enigmatic dance of measurement error [54]. In essence, while the multifaceted conundrum of multicollinearity in factor analysis has garnered copious scholarly attention, a realm ripe with opportunities still unfolds for further inquiry, offering glimpses into its profound reverberations on factor analysis outcomes and the nuanced efficacy of diverse problem-solving methodologies within distinct contexts.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Hair, J.F., Black, W.C., Babin, B.J. and Anderson, R.E. (2019) Multivariate Data Analysis. 8th Edition, Pearson, Upper Saddle River.

[2] Costello, A.B. and Osborne, J.W. (2005) Best Practices in Exploratory Factor Analysis: Four Recommendations for Getting the Most from Your Analysis. *Practical Assessment, Research & Evaluation*, **10**, 1-9.

[3] Fabrigar, L.R., Wegener, D.T., MacCallum, R.C. and Strahan, E.J. (1999) Evaluating the Use of Exploratory Factor Analysis in Psychological Research. *Psychological Methods*, **4**, 272-299. https://doi.org/10.1037/1082-989X.4.3.272

[4] Williams, B., Onsman, A. and Brown, T. (2010) Exploratory Factor Analysis: A Five-Step Guide for Novices. *Australasian Journal of Paramedicine*, **8**, 1-13. https://doi.org/10.33151/ajp.8.3.93

[5] Field, A.P. (2018) Discovering Statistics Using IBM SPSS Statistics. 5th Edition,

Sage, Newbury Park.

[6] Thompson, B. (2004) Exploratory and Confirmatory Factor Analysis: Understanding Concepts and Applications. American Psychological Association, Washington DC. https://doi.org/10.1037/10694-000

[7] Beavers, A.S., Lounsbury, J.W., Richards, J.K., Huck, S.W., Skolits, G.J. and Esquivel, S.L. (2013) Practical Considerations for Using Exploratory Factor Analysis in Educational Research. *Practical Assessment, Research, and Evaluation*, **18**, 6.

[8] Henson, R.K. and Roberts, J.K. (2006) Use of Exploratory Factor Analysis in Published Research: Common Errors and Some Comment on Improved Practice. *Educational and Psychological Measurement*, **66**, 393-416. https://doi.org/10.1177/0013164405282485

[9] Ledesma, R.D., Ferrando, P.J., Trógolo, M.A., Poó, F.M., Tosi, J.D. and Castro, C. (2021) Exploratory Factor Analysis in Transportation Research: Current Practices and Recommendations. *Transportation Research Part F: Traffic Psychology and Behaviour*, **78**, 340-352. https://doi.org/10.1016/j.trf.2021.02.021

[10] Daoud, J.I. (2017, December) Multicollinearity and Regression Analysis. https://doi.org/10.1088/1742-6596/949/1/012009

[11] Alin, A. (2017) Multicollinearity. Wiley StatsRef: Statistics Reference Online.

[12] Marquardt, D.W. (1970) Generalized Inverses, Ridge Regression, Biased Linear Estimation, and Nonlinear Estimation. *Technometrics*, **12**, 591-612. https://doi.org/10.2307/1267205

[13] Belsley, D.A. (1991) Conditioning Diagnostics: Collinearity and Weak Data in Regression. John Wiley & Sons, Hoboken.

[14] Porter, D.C. and Gujarati, D.N. (2009) Basic Econometrics. McGraw-Hill Irwin, New York.

[15] Mickey, R.M. and Greenland, S. (1989) The Impact of Confounder Selection Criteria on Effect Estimation. *American Journal of Epidemiology*, **129**, 125-137. https://doi.org/10.1093/oxfordjournals.aje.a115101

[16] Kline, R.B. (2015) Principles and Practice of Structural Equation Modeling. 4th Edition, Guilford Press, New York.

[17] Hotelling, H. (1933) Analysis of a Complex of Statistical Variables into Principal Components. *Journal of Educational Psychology*, **24**, 417-441. https://doi.org/10.1037/h0071325

[18] Kutner, M.H. (2005) Applied Linear Statistical Models.

[19] Sulaiman, M.S., Abood, M.M., Sinnakaudan, S.K., Shukor, M.R., You, G.Q. and Chung, X.Z. (2021) Assessing and Solving Multicollinearity in Sediment Transport Prediction Models Using Principal Component Analysis. *ISH Journal of Hydraulic Engineering*, **27**, 343-353. https://doi.org/10.1080/09715010.2019.1653799

[20] Johnson, R.A. and Wichern, D.W. (2007) Applied Multivariate Statistical Analysis. 6th Edition, Pearson Prentice Hall, Upper Saddle River.

[21] Tabachnick, B.G. and Fidell, L.S. (2013) Using Multivariate Statistics. 6th Edition, Pearson, Upper Saddle River.

[22] Sass, D.A., Schmitt, T.A. and Marsh, H.W. (2018) Evaluating Model Fit with Ordered Categorical Data within a Measurement Invariance Framework: A Comparison of Estimators. *Structural Equation Modeling: A Multidisciplinary Journal*, **25**, 604-619.

[23] Hoffman, L. and Rovine, M.J. (2015) Multilevel Models for the Experimental Psychologist: Foundations and Illustrative Examples. *Behavior Research Methods*, **47**,

967-978.

[24] Mindrila, D. (2010) Maximum Likelihood (ML) and Diagonally Weighted Least Squares (DWLS) Estimation Procedures: A Comparison of Estimation Bias with Ordinal and Multivariate Non-Normal Data. *International Journal of Digital Society*, **1**, 60-66. https://doi.org/10.20533/ijds.2040.2570.2010.0010

[25] Reise, S.P., Bonifay, W.E. and Haviland, M.G. (2013) Scoring and Modeling Psychological Measures in the Presence of Multidimensionality. *Journal of Personality Assessment*, **95**, 129-140. https://doi.org/10.1080/00223891.2012.725437

[26] Kelava, A., Moosbrugger, H., Dimitruk, P. and Schermelleh-Engel, K. (2008) Multicollinearity and Missing Constraints: A Comparison of Three Approaches for the Analysis of Latent Nonlinear Effects. *Methodology*, **4**, 51-66. https://doi.org/10.1027/1614-2241.4.2.51

[27] R Core Team (2021) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna.

[28] Beauducel, A. and Hilger, N. (2017) On the Bias of Factor Score Determinacy Coefficients Based on Different Estimation Methods of the Exploratory Factor Model. *Communications in Statistics-Simulation and Computation*, **46**, 6144-6154. https://doi.org/10.1080/03610918.2016.1197247

[29] Hallgren, K.A. (2012) Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial. *Tutorials in Quantitative Methods for Psychology*, **8**, 23-34. https://doi.org/10.20982/tqmp.08.1.p023

[30] Kolenikov, S. and Angeles, G. (2009) Socioeconomic Status Measurement with Discrete Proxy Variables: Is Principal Component Analysis a Reliable Answer? *Review of Income and Wealth*, **55**, 128-165. https://doi.org/10.1111/j.1475-4991.2008.00309.x

[31] Auerswald, M. and Moshagen, M. (2019) How to Determine the Number of Factors to Retain in Exploratory Factor Analysis: A Comparison of Extraction Methods under Realistic Conditions. *Psychological Methods*, **24**, 468-491. https://doi.org/10.1037/met0000200

[32] Appelbaum, M.I. and Cramer, E.M. (1974) Some Problems in the Nonorthogonal Analysis of Variance. *Psychological Bulletin*, **81**, 335-343. https://doi.org/10.1037/h0036315

[33] Gardner, M.J. and Altman, D.G. (1986) Confidence Intervals Rather than P Values: Estimation Rather than Hypothesis Testing. *BMJ* (*Clinical Research ed.*), **292**, 746-750. https://doi.org/10.1136/bmj.292.6522.746

[34] Fox, J. and Weisberg, S. (2018) An R Companion to Applied Regression. Sage Publications.

[35] Strang, G. (2006) Linear Algebra and Its Applications. Thomson, Brooks/Cole, Belmont, CA.

[36] Lax, P.D. (2007) Linear Algebra and Its Applications (Vol. 78). John Wiley & Sons, Hoboken.

[37] Jolliffe, I.T. (2002) Principal Component Analysis. John Wiley & Sons, Hoboken.

[38] Weisberg, S. (2014) Applied Linear Regression. John Wiley & Sons, Hoboken.

[39] Cook, R.D. (1977) Detection of Influential Observation in Linear Regression. *Technometrics*, **19**, 15-18. https://doi.org/10.1080/00401706.1977.10489493

[40] Joreskog, K.G. and Sorbom, D. (1996) LISREL 8: User's Reference Guide. Scientific Software International, Lincolnwood.

[41] Preacher, K.J., Curran, P.J. and Bauer, D.J. (2007) Computational Tools for Probing

Interactions in Multiple Linear Regression, Multilevel Modeling, and Latent Curve Analysis. *Journal of Educational and Behavioral Statistics*, **31**, 437-448. https://doi.org/10.3102/10769986031004437

[42] Enders, C.K. (2001) A Primer on Maximum Likelihood Algorithms Available for Use with Missing Data. *Structural Equation Modeling: A Multidisciplinary Journal*, **8**, 128-141. https://doi.org/10.1207/S15328007SEM0801_7

[43] Aiken, L.S. and West, S.G. (1991) Multiple Regression: Testing and Interpreting Interactions. Sage, Newbury Park.

[44] Kuhn, M. (2021) Caret: Classification and Regression Training. R Package Version 6.0-88. https://CRAN.R-project.org/package=caret

[45] Kim, J.O. and Mueller, C.W. (1978) Factor Analysis: Statistical Methods and Practical Issues (Vol. 14). SAGE Publications, Inc., Thousand Oaks. https://doi.org/10.4135/9781412984256

[46] Osborne, J.W. and Waters, E. (2002) Four Assumptions of Multiple Regression That Researchers Should Always Test. *Practical Assessment, Research & Evaluation*, **8**, 1-9.

[47] Abdi, H. and Williams, L.J. (2010) Principal Component Analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, **2**, 433-459. https://doi.org/10.1002/wics.101

[48] Geladi, P. and Kowalski, B.R. (1986) Partial Least-Squares Regression: A Tutorial. *Analytica Chimica Acta*, **185**, 1-17. https://doi.org/10.1016/0003-2670(86)80028-9

[49] Jackson, D.A. (1991) Stopping Rules in Principal Components Analysis: A Comparison of Heuristical and Statistical Approaches. *Ecology*, **74**, 2204-2214. https://doi.org/10.2307/1939574

[50] Harrington, D. (2009) Confirmatory Factor Analysis. Oxford University Press, Oxford. https://doi.org/10.1093/acprof:oso/9780195339888.001.0001

[51] Soper, D.S. (2021) Merging Variables in SPSS and R. https://www.statskingdom.com/220merge.html

[52] Everitt, B.S., Landau, S. and Leese, M. (2011) Cluster Analysis. 4th Edition, Arnold Publishers, London. https://doi.org/10.1002/9780470977811

[53] Bollen, K.A. and Lennox, R. (1991) Conventional Wisdom on Measurement: A Structural Equation Perspective. *Psychological Bulletin*, **110**, 305-314. https://doi.org/10.1037/0033-2909.110.2.305

[54] Graham, J.M. (2003) Adding Missing-Data-Relevant Variables to FIML-Based Structural Equation Models. *Structural Equation Modeling*, **10**, 80-100. https://doi.org/10.1207/S15328007SEM1001_4