Süleyman Demirel Üniversitesi
Fen Bilimleri Enstitüsü Dergisi
Cilt 27, Sayı 3, 401-410, 2023

Süleyman Demirel University
Journal of Natural and Applied Sciences
Volume 27, Issue 3, 401-410, 2023

DOI: 10.19113/sdufenbed.1250855

Comparison of Classical and Robust Factor Analyses Methods

Barış ERGÜL[*1], **Zeki YILDIZ**[2]

[1,2]University of Eskişehir Osmangazi, Faculty of Science, Department of Statistics, 26040, Eskişehir, Türkiye

(Alınış / Received: 13.02.2023, Kabul / Accepted: 14.09.2023, Online Yayınlanma / Published Online: 25.12.2023)

Keywords
MLE, GLS, MCD, M, Classical Factor Analysis, Robust Factor Analysis

Abstract: Factor analysis is a multivariate statistical analysis technique that has become very popular in recent years. In the factor analysis model, the error covariance matrix is assumed to be the multivariate normal distribution, and outliers are likely to be accounted for. Various estimation methods were compared with Monte Carlo simulation for the factor analysis model. The performances of the estimation methods were evaluated based on the ratio of the total variance explained and the criterion fit values. Considering the MLE, PCA, WLS, and GLS methods for classical factor analysis and the MCD, M, and S methods for robust factor analysis, the ratio of total variance explained, and fit values decreased as the sample size increased. When the number of variables increases, the ratio of total variance explained, and fit values increase at different sample sizes. It can be said that the WLS and GLS methods are better than others for classical factor analysis and the MCD and M methods are better than others for robust factor analysis.

Klasik ve Sağlam Faktör Analizleri Yöntemlerinin Karşılaştırılması

Anahtar Kelimeler
MLE, GLS, MCD, M, Klasik Faktör Analizi, Sağlam Faktör Analizi

Öz: Faktör analizi, son yıllarda popüler hale gelen çok değişkenli istatistiksel analiz tekniklerinden biridir. Bu çalışmada, hata kovaryans matrisinin çok değişkenli normal dağılım ve aykırı değerler olması durumunda faktör analizi modeli kullanılmıştır. Faktör analizi modeli için farklı tahmin yöntemleri Monte Carlo simülasyonu ile karşılaştırılmıştır. Tahmin yöntemlerinin performansı, açıklanan toplam varyans oranı ve uyum değerleri kriterine göre değerlendirilmiştir. Klasik faktör analizi için MLE, PCA, WLS ve GLS yöntemleri ve sağlam faktör analizi için MCD, M ve S yöntemleri dikkate alındığında, toplam varyansın açıklama oranı ve fit değerleri, farklı örneklem büyüklüklerinde artarak, her bir örneklem büyüklüğünde azalmıştır. Değişken sayısı arttıkça açıklanan toplam varyans oranı ve fit değerleri farklı örneklem büyüklüklerinde artmaktadır. Klasik faktör analizi için WLS ve GLS yöntemlerinin, sağlam faktör analizi için MCD ve M yöntemlerinin daha iyi yöntemler olduğu söylenebilir.

1. Introduction

Today, many variables shed light on problems, events, facts or perceptions, attitudes, and behaviors. It is no longer sufficient to examine a single variable to solve the problems arising from these events, phenomena, perceptions, attitudes, and behaviors. However, as the number of variables increases, the study of events, phenomena, perceptions, attitudes, and behaviors becomes even more complex.

Factor analysis is a multivariate statistical analysis technique that has become very popular in recent years. Factor analysis aims to determine the original (independent) variables in the data set with linear combinations called factors. The first step is to create the covariance matrix (or correlation matrix) when the number of original variables is p. The factor analysis model contains many parameters, including the variances of the error components. The error components are the parts of the observed variables that are not explained by the factors. The variances of the error components are important because they determine the amount of variance in the observed variables that are not explained by the factors [1].

p is the independent variable, assuming that $x_1, x_2, \dots, x_p$, and k associate the latent factors $f_1, f_2, \dots, f_k$ with the following statistical model:

$$x_j - \mu_j = \lambda_{j1} f_1 + \lambda_{j2} f_2 + \cdots + \lambda_{jk} f_k + \varepsilon_j \tag{1}$$

$\lambda_{j1}, \lambda_{j2}, \dots \lambda_{jk}$ refers to the factor loadings, ε_j refers to the error terms.

[*] Corresponding author: bergul@ogu.edu.tr

Factor Analysis Model with k factor is defined as follows with matrix notation when defined by $x = (x_1, x_2, \dots, x_p)', f = (f_1, f_2, \dots, f_k)$ and $\varepsilon = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_p)'$.

$$x - \mu = \Lambda f + \varepsilon \qquad (2)$$

Λ refers to the matrix of factor loadings, f refers to the factor score vector and ε refers to the error vector.

The assumptions of the factor analysis model are as follows: [2]

1. The factors and the vector of the error terms are independent of each other and their mean is a zero vector. $(E(f) = E(\varepsilon) = 0, Cov(x, f) = \Lambda)$

2. The covariance matrix of the factors is equal to the unit vector. $(Cov(f) = I_k)$

3. The joint distribution of the factors is a multivariate normal distribution.

These assumptions are strong, and they may not always be met in real data. However, it has been shown that the classical estimates have good asymptotic properties under some weak assumptions. This means that the estimates will be approximately correct as the sample size increases.

Some of the weak assumptions that are sufficient for the classical estimates to have good asymptotic properties include:

i. The factors are not perfectly correlated.
ii. The error terms are not perfectly correlated.
iii. The error terms have a finite variance.

These assumptions are more likely to be met in real data than the strong assumptions listed above. Therefore, the factor analysis model can be a useful tool for data analysis even if the strong assumptions are not met.

It is important to note that the factor analysis model is a statistical model, and as such, it is only an approximation of reality. The estimates from the factor analysis model will never be perfect, but they can be a useful tool for understanding the data [3-4].

The main purpose of factor analysis is to obtain the matrix **Λ** and the covariance of the error matrix (**Ψ**) obtained by orthogonal transformation. The maximum likelihood method and the basic factor methods are obtained by separating the matrix **Σ** where $\Sigma = Cov(x)$.

$$\Sigma = \Lambda\Lambda' + di \qquad (3)$$

The factor analysis model is shown in equations (1) and (2). The variance of the variables to which the common factor contributes is called the common variance. The common variance can be replaced in the equation by h_i^2 in the equation $\Sigma = \Lambda\Lambda' + \Psi$ and $\Sigma = H^2 + \Psi$ can be written. The common variance is the sum of the loadings of the variables on the common factor. When $k > 1$, there is always some natural uncertainty associated with the factor model.

Let T which is any mxm dimensional orthogonal matrix and consider $TT' = T'T = I$. The equality in (2) can be written as follows:

$$\begin{gathered} x - \mu = \Lambda f + \varepsilon = \Lambda TT' + \varepsilon = \Lambda^* f^* + \varepsilon \\ \Lambda^* = \Lambda T \text{ ve } f^* = T'f \text{ and,} \\ E(f^*) = T'E(f) = 0 \text{ and,} \\ Cov(f^*) = T'Cov(f)T = T'T = I \end{gathered} \qquad (4)$$

Based on the observations on x, it is impossible to distinguish Λ factor loadings and Λ^* factor loadings. That is, f ve $f^* = Tf$ factors have the same statistical properties. In general, although Λ^* factor loadings and Λ factor loadings are different, they were both obtained from the same covariance matrix.

$$\Sigma = \Lambda\Lambda' + \Psi = \Lambda\mathrm{TT}'\Lambda' + \Psi = (\Lambda^*)(\Lambda^*)' + \Psi \qquad (5)$$

Since orthogonal matrices correspond to coordinate system transformations, the uncertainty structure is removed by "factor rotation". Λ factor loadings are determined by an orthogonal matrix T.

$$\Lambda^* = \Lambda T \text{ and } \Lambda \qquad (6)$$

The common variance is determined by the diagonal elements of the matrix $\Lambda\Lambda' = (\Lambda^*)(\Lambda^*)'$ is not affected by the choice of the orthogonal matrix T.

The factor analysis progresses by identifying conditions that allow the estimation of λ and ψ matrices. The matrix of factor loadings is then rotated (multiplied by an orthogonal matrix), with the rotation determined by some of the "ease of interpretation" criteria. Once the factor loadings and error terms have been determined, the factors are determined and the estimated values of the factors themselves (called factor scores) are produced [5].

The factor analysis model for this study assumed that the covariance matrix error terms had a multivariate normal distribution, and outliers are likely to be considered. Various estimation methods were compared with the Monte Carlo simulation for the factor analysis model. The performance of the estimation methods was evaluated based on the ratio of the total variance explained and fit values. In the second phase of the study, estimation methods were presented. Later, the estimation methods were compared with the simulation study for different sample sizes.

The remainder of this paper is arranged as follows. The Factor Analysis models are described in Section 2. Section 3 describes simulation results for Classical and Robust estimation of the Factor Analysis models. Section 4 considers an application of the Classical and Robust Factor Analysis models to Women Track Records data. Different estimation techniques are compared in terms of computational efficiency. Conclusions and a few remarks are given in Section 5.

2. Material and Methods

The sample covariance matrix (S) is an estimator of the unknown population covariance matrix Σ. If the out-of-diagonal elements of the S matrix obtain small values, the variables (or if the sample correlation matrix is essentially close to zero or zero value) are unrelated and it is not useful to analyze a factor. However, the main purpose of the factor analysis is to determine common factors.

Three of the most commonly used methods for parameter estimation in factor analysis are the principal component (and the corresponding basic factor), the maximum likelihood method and the robust estimation method.

2.1. Principal Component Method

The principal component factor analysis of the S sample covariance matrix is indicated by $\hat{\lambda}_1 \geq \hat{\lambda}_2 \dots \geq \hat{\lambda}_p$, the eigenvalues/eigenvectors pairs $(\hat{\lambda}_1, \hat{e}_1), (\hat{\lambda}_2, \hat{e}_2), \dots, (\hat{\lambda}_p, \hat{e}_p)$. To show the number of common factors given with $k < p$, the prediction matrix of the estimated factor loadings $\tilde{l}_{ij}$ is given as follows [5]:

$$\widetilde{\Lambda} = \left[\sqrt{\hat{\lambda}_1}\hat{e}_1 \quad \vdots \quad \sqrt{\hat{\lambda}_2}\hat{e}_2 \vdots \cdots \sqrt{\hat{\lambda}_k}\hat{e}_k\right] \tag{7}$$

The estimated error term matrix is provided by the diagonal elements of the $S - \widetilde{\Lambda}\widetilde{\Lambda}'$ matrix:

$$\widetilde{\Psi} = \begin{bmatrix} \widetilde{\Psi}_1 & 0 \dots & 0 \\ 0 & \widetilde{\Psi}_2 \dots & 0 \\ 0 & 0 \dots & \widetilde{\Psi}_p \end{bmatrix} \tag{8}$$

Here, it is expressed by $\widetilde{\Psi}_i = s_{ii} - \sum_{j=1}^{k} \tilde{l}_{ij}^2$. The common variance is estimated as follows:

$$\tilde{h}_i^2 = \tilde{l}_{i1}^2 + \tilde{l}_{i2}^2 + \cdots + \tilde{l}_{ik}^2 \tag{9}$$

The principal component factor analysis for the correlation matrix found by the sample is obtained starting with R instead of S.

$$S - \widetilde{(\Lambda\Lambda'} + \widetilde{\Psi}) \tag{10}$$

The diagonal elements of the S matrix will be equal to the diagonal elements of the $\widetilde{\Lambda}\widetilde{\Lambda}' + \widetilde{\Psi}$ matrix. To prevent this situation, this problem is solved by taking the factor as the number of principal components of the S matrix. This raises several factors. This allows the selection of eigenvalues larger than the value of 1, as in the principal component analysis [5].

2.2. Maximum Likelihood Method, Weighted Least Squares, and Generalized Least Squares Method

If the distribution of factors and error terms are assumed to be normal, the maximum likelihood estimation of factor loadings and error variance can be written. When the joint probability functions of the f_j and ε_j are normally distributed, $x_j - \mu = \Lambda f_j + \varepsilon_j$ also has the normal distribution. In this case, the maximum likelihood function can be written. The maximum likelihood function varies between Λ and Ψ, $\Sigma = \Lambda\Lambda' + \Psi$. Estimation of $\widehat{\Lambda}$ and Ψ is resolved by providing the following conditions [5-6]:

$$\mathrm{S}\widehat{\Psi}\widehat{\Lambda} = \widehat{\Lambda}(I + \widehat{\Lambda}'\widehat{\Psi}^{-1}\widehat{\Lambda}) \tag{11}$$

$$\widehat{\Psi} = diag(S - \widehat{\Lambda}\,\widehat{\Lambda}') \tag{12}$$

$$\widehat{\Psi} = diag(S - \widehat{\Lambda}\,\widehat{\Lambda}') \tag{13}$$

These equations are resolved iteratively until they converge. For all of the Weighted Least Squares (WLS), Generalized Least Squares (GLS), and Maximum Likelihood (MLE) estimation, gradient algorithms have been developed: those with the Fletcher-Powell and Newton-Raphson methods have been proposed for the MLE [7-8], while the algorithms using the Newton-Raphson and Gauss-Newton methods have been developed for the GLS [9-10] with the gradient algorithms. On the other hand, inequality-based algorithms have been developed for the MLE excluding the GLS. The GLS solution weights the residual matrix by the inverse of the correlation matrix. This has the effect of weighting those variables with low loadings even more than those with high loadings. The WLS solution weights the residual matrix by 1/diagonal of the inverse of the correlation matrix. This has the effect of weighting items with low loadings more than those with high loadings [11].

2.3. Minimum Covariance Determinant (MCD), M and S Estimation Methods

The minimum covariance determinant (MCD) estimator for location and scale can be found using an algorithm implemented by [12]. This algorithm essentially requires step C. In this step, an approximate value for the MCD method is taken, and it is possible to reach another value with a smaller determinant. The MCD algorithm can be summarized as follows: The algorithm aims to find subsets of observations that minimize the determinant of the covariance matrix calculated for a sample of size n. The MCD method is based on the following assumptions. To this end, h observations are examined and the goal is to find a subset of h observations that minimize the determinant of the covariance matrix. Typically, h is taken as h ≈ [0.75*n], where [.] denotes the integer part. h represents the minimum number of observations without outliers. The mean vector calculated for h observations gives the estimate of the location parameter vector for MCD and the covariance matrix calculated for the same values gives the estimate of the scale parameter for MCD [13].

Another method for finding a robust covariance estimate is the M-estimator. The M-estimator aims to minimize the determinant for the multivariate location and scale parameters by finding the S estimator. The S estimator aims to find the weighted mean and the covariance matrix by iterations [14].

The M-estimator for the estimation of μ and Σ uses the S estimation method, which was first introduced in a publication referred to as [15] and then further studied in [16]. For a data set consisting of p-variable observations $\{x_1, \dots, x_n\}$, the S estimator for (μ, Σ) is obtained from the solution of $\sigma(d_1, \dots, d_n) = min$. Here, $(x_i - \mu)'\Sigma^{-1}(x_i - \mu)$ and det(Σ) = 1. Where $\sigma = \sigma(z)$, is the M-estimator of $z = \{z_1, \dots, z_n\}$. It is defined as the solution of $\frac{1}{n}\sum \rho\left(\frac{z}{\sigma}\right) = \delta$ where ρ is non-decreasing, ρ(0) = 0 and ρ(∞) = 1 and δ ∈ (0,1). More simply, the S estimator finds the positive definite symmetric matrix Σ that minimizes the μ vector and det(Σ). S-estimators have a close connection with M estimators, and the solution for (μ, Σ) is also the solution of an equation defining a weighted sample mean and a covariance matrix with an M estimator [13].

3. Results

In this study, the results have been presented based on derived data. For this purpose, the error variances, factor loadings, and the covariance matrix of variables were derived from a multivariate normal distribution. It is assumed that each variable has the same variance and that all the covariance between the variables is equal. All factor loadings are assumed equal in size and set to $\lambda_i = 1$. Thus, the data is derived by the following method: [17]

1. For each observation, construct scores for the construct with the desired number of factors derived from a multivariate normal distribution with a mean value μ, where the variance of each factor is σ_f^2 and the covariance between the two scores is γ.

2. Generate an equal number of variables for each factor, where the score s_i for the variable i is $s_i = \mu + e(i)$, where $e(i) \sim N(0, \sigma_e^2)$.

For the simulation study, the number of repetitions was selected 1000, the number of sample sizes was selected n = 100,500,1000, the number of variables was selected p = 15,20,25,30, and the number of the factor was selected k = 2,3,4. The average of the factors was selected as 5, the variance and covariances of the factors were 1 and 0.5, respectively, and variances of error term 1 were selected. The R program was used for simulation. 7 methods were selected; the MLE (the maximum likelihood estimation), the PCA (the principal components), the WLS (the weighted least squares), the GLS (the generalized least squares) methods, the MCD (minimum covariance determinant), M, and S estimation methods.
Then, 10 outlier observations were added to the dataset and the results were discussed accordingly.

To make comparisons among methods, the method that gives the highest value ratio of the total variance explained and fit values is considered better method. Fit values refer to how well the factor model reproduces the correlation matrix.

Table 1-2 for classical factor analysis shows that the MLE method is better than others for n=100, p=15, and k=2. The GLS method is better than others for n=500, p=15, and k=2. The WLS method is better than others for n=1000, p=15, and k=2. The WLS method is better than others for n=100, p=15, and k=3. The GLS method performs better than others for n=500, p=15, and k=3. The MLE method is better than others for n=1000, p=15, and k=3. The MLE method performs better than others for n=100, p=15, and k=4. The GLS method is better than others for n=500, p=15, and k=4. The WLS method performs better than others for n=1000, p=15, and k=4.

In all classical methods, the ratio of total variance is explained, and the fit values decrease with each increase in sample size n=100,500 and 1000. When the number of variables increases, the ratio of total variance explained and the fit values increase as the sample size increases n=100,500 and 1000. In the classical factor analysis, in the cases where n=100,500 and 1000 and p=15,20,25,30, it can be seen that the ratio of total variance explained and the fit values increase considering the k=2 factor structure. It can be said that MLE, WLS and GLS methods are better than others for classical factor analysis.

Table 1-2 for robust factor analysis shows that the M method is better than others for n=100, p=15, and k=2. The MCD method performs better than others for n=500, p=15, and k=2. The MCD method is better than others for n=1000, p=15, and k=2. The MCD method is better than others for n=100, p=15, and k=3. The MCD method performs better than others for n=500, p=15, and k=3. The MCD method is better than others for n=1000, p=15, and k=3. The MCD method is better than others for n=100, p=15, and k=4. The MCD method is better than others for n=500, p=15, and k=4. The MCD method is better than others for n=1000, p=15, and k=4.

In all robust methods, the ratio of total variance explained and the fit values decrease with each increase in sample size increases n=100,500 and 1000. When the number of variables increases, the ratio of total variance explained and the fit values increase as the sample size increases n=100,500 and 1000. In the robust factor analysis, it can be said that MCD and M methods are better than others for robust factor analyses.

The classical factor analysis methods outperformed other techniques in terms of the ratio of total variance explained and fit values. This is because the dataset used in the analysis was derived from a multivariate normal distribution and did not contain any outliers.

The classical factor analysis methods are based on the assumption that the data follows a multivariate normal distribution. This assumption is not always met in real data, but it is a good approximation for many datasets. The robust factor analysis methods are designed to be more robust to depart from the multivariate normal distribution, but they are not as efficient as the classical factor analysis methods when the data does follow a multivariate normal distribution.

The results of the study support the theoretical framework of classical factor analysis. Classical factor analysis is best suited for datasets that follow a multivariate normal distribution. However, it is important to note that the classical factor analysis methods may not be as accurate for datasets that do not follow a multivariate normal distribution.

In addition to the assumptions about the distribution of the data, the results of the study also depend on the sample size. The classical factor analysis methods are more accurate for larger sample sizes. This is because the classical factor analysis methods rely on maximum likelihood estimation, which is a more efficient estimator for larger sample sizes.

Overall, the results of the study suggest that the classical factor analysis methods are a good choice for estimating the factor analysis model when the data follows a multivariate normal distribution and the sample size is large. However, it is important to note that the classical factor analysis methods may not be as accurate for datasets that do not follow a multivariate normal distribution.

The analysis of the ratio of total variance explained and fit values indicate that classical factor analysis methods outperform other techniques. This is because the dataset used in the analysis was derived from a multivariate normal distribution and does not contain any outliers. These results support the theoretical framework of classical factor analysis, which is best suited for datasets that follow a multivariate normal distribution.

Table 1. The Ratio of Total Variance Explained for Classical and Robust Factor Analysis (1000 repetitions)

		p=15			**p=20**			**p=25**			**p=30**		
Sample size	**Method**	**k=2**	**k=3**	**k=4**	**k=2**	**k=3**	**k=4**	**k=2**	**k=3**	**k=4**	**k=2**	**k=3**	**k=4**
100	MLE	**0,5084**	0,5363	**0,5811**	0,5110	**0,5526**	**0,5764**	0,5544	0,5746	0,6003	0,5754	0,5974	0,6155
	PCA	0,5079	0,5327	0,5601	0,5109	0,5441	0,5672	0,5539	0,5744	0,5924	0,5753	0,5973	0,6153
	WLS	0,5082	**0,5366**	0,5582	**0,5264**	0,5474	0,5694	0,5444	0,5747	0,5936	0,5755	0,5974	0,6158
	GLS	0,5065	0,5321	0,5573	0,5162	0,5464	0,5707	**0,5785**	**0,5794**	**0,6007**	**0,5783**	**0,6002**	**0,6174**
	MCD	0,5074	**0,5354**	**0,5801**	0,5183	**0,5484**	**0,5744**	**0,5749**	**0,5764**	**0,5998**	0,5768	0,5935	0,6168
	M	**0,5078**	0,5353	0,5774	**0,5194**	0,5471	0,5691	0,5634	0,5733	0,5993	**0,5773**	**0,5994**	**0,6171**
	S	0,5069	0,5323	0,5702	0,5071	0,5413	0,5688	0,5538	0,5744	0,5925	0,5712	0,5923	0,6161
500	MLE	0,4914	0,5074	0,5420	0,5027	0,5316	0,5574	0,5125	0,5344	0,5745	0,5356	0,5536	0,5823
	PCA	0,4910	0,5071	0,5415	0,5026	0,5315	0,5564	0,5124	0,5343	0,5741	0,5352	0,5533	0,5819
	WLS	0,4913	0,5075	0,5430	0,5030	0,5356	0,5573	0,5126	0,5345	0,5744	0,5355	0,5537	0,5830
	GLS	**0,4915**	**0,5077**	**0,5432**	**0,5108**	**0,5436**	**0,5626**	**0,5128**	**0,5352**	**0,5749**	**0,5357**	**0,5539**	**0,5832**
	MCD	**0,4911**	**0,5073**	**0,5364**	**0,5092**	**0,5315**	**0,5569**	**0,5117**	**0,5337**	**0,5742**	0,5325	0,5515	0,5811
	M	0,4910	0,5063	0,5360	0,5088	0,5311	0,5560	0,5110	0,5332	0,5731	**0,5327**	**0,5519**	**0,5817**
	S	0,4902	0,5045	0,5332	0,5036	0,5304	0,5547	0,5095	0,5330	0,5711	0,5306	0,5502	0,5800
1000	MLE	0,4823	**0,4936**	0,5211	0,4918	0,5148	0,5448	0,5085	0,5247	0,5539	0,5215	0,5399	0,5712
	PCA	0,4816	0,4928	0,5200	0,4911	0,5144	0,5432	0,5082	0,5244	0,5533	0,5208	0,5398	0,5705
	WLS	**0,4826**	0,4944	**0,5223**	0,4927	0,5156	0,5449	0,5085	0,5248	0,5542	0,5218	0,5404	0,5726
	GLS	0,4821	0,4934	0,5207	**0,4936**	**0,5168**	**0,5464**	**0,5088**	**0,5253**	**0,5549**	**0,5226**	**0,5411**	**0,5733**
	MCD	**0,4805**	**0,4918**	**0,5198**	**0,4892**	**0,5140**	**0,5428**	**0,5066**	**0,5231**	**0,5538**	**0,5195**	**0,5384**	0,5702
	M	0,4800	0,4914	0,5195	0,4889	0,5137	0,5422	0,5061	0,5220	0,5521	0,5192	0,5380	0,5700
	S	0,4791	0,4904	0,5185	0,4877	0,5130	0,5406	0,5038	0,5222	0,5510	0,5177	0,5361	0,5690

Table 2. The Fit Values for Classical and Robust Factor Analysis (1000 repetitions)

		p=15			p=20			p=25			p=30		
Sample size	**Method**	**k=2**	**k=3**	**k=4**	**k=2**	**k=3**	**k=4**	**k=2**	**k=3**	**k=4**	**k=2**	**k=3**	**k=4**
100	MLE	**0,9267**	0,9359	**0,9446**	0,9484	**0,9569**	**0,9614**	0,9636	0,9727	0,9754	0,9764	0,9794	0,9813
	PCA	0,9265	0,9348	0,9432	0,9482	0,9551	0,9610	0,9626	0,9726	0,9744	0,9763	0,9792	0,9811
	WLS	0,9265	**0,9361**	0,9436	**0,9496**	0,9558	0,9612	0,9616	0,9727	0,9753	0,9765	0,9795	0,9817
	GLS	0,9261	0,9345	0,9427	0,9488	0,9556	0,9613	**0,9704**	**0,9733**	**0,9768**	**0,9770**	**0,9798**	**0,9822**
	MCD	0,9262	**0,9354**	**0,9445**	0,9490	**0,9564**	**0,9612**	**0,9701**	**0,9732**	**0,9753**	0,9767	0,9794	0,9819
	M	**0,9263**	0,9353	0,9442	**0,9492**	0,9562	0,9610	0,9675	0,9731	0,9751	**0,9768**	**0,9795**	**0,9820**
	S	0,9261	0,9348	0,9439	0,9473	0,9543	0,9610	0,9625	0,9729	0,9743	0,9759	0,9788	0,9813
500	MLE	0,9168	0,9308	0,9426	0,9337	0,9458	0,9481	0,9554	0,9623	0,9718	0,9761	0,9780	0,9788
	PCA	0,9165	0,9305	0,9427	0,9337	0,9456	0,9467	0,9551	0,9622	0,9714	0,9757	0,9777	0,9786
	WLS	0,9167	0,9310	0,9431	0,9340	0,9462	0,9476	0,9552	0,9624	0,9716	0,9759	0,9782	0,9793
	GLS	**0,9170**	**0,9313**	**0,9435**	**0,9351**	**0,9465**	**0,9485**	**0,9558**	**0,9631**	**0,9726**	**0,9764**	**0,9784**	**0,9798**
	MCD	**0,9164**	**0,9262**	**0,9388**	**0,9347**	**0,9455**	**0,9470**	**0,9540**	**0,9611**	**0,9715**	0,9741	0,9764	0,9782
	M	0,9162	0,9261	0,9385	0,9343	0,9453	0,9468	0,9535	0,9608	0,9706	**0,9744**	**0,9766**	**0,9784**
	S	0,9160	0,9221	0,9354	0,9335	0,9446	0,9456	0,9526	0,9605	0,9697	0,9730	0,9751	0,9777
1000	MLE	0,9071	**0,9243**	0,9310	0,9233	0,9370	0,9453	0,9485	0,9516	0,9689	0,9724	0,9745	0,9771
	PCA	0,9066	0,9239	0,9300	0,9211	0,9367	0,9442	0,9483	0,9513	0,9683	0,9719	0,9742	0,9762
	WLS	**0,9077**	0,9248	**0,9324**	0,9244	0,9376	0,9455	0,9485	0,9517	0,9692	0,9728	0,9757	0,9776
	GLS	0,9072	0,9244	0,9320	**0,9257**	**0,9379**	**0,9463**	**0,9489**	**0,9524**	**0,9698**	**0,9731**	**0,9759**	**0,9788**
	MCD	**0,9055**	**0,9210**	**0,9295**	**0,9204**	**0,9361**	**0,9431**	**0,9469**	**0,9505**	**0,9677**	**0,9604**	**0,9737**	**0,9755**
	M	0,9053	0,9207	0,9292	0,9201	0,9357	0,9427	0,9466	0,9502	0,9671	0,9601	0,9733	0,9752
	S	0,9043	0,9201	0,9281	0,9191	0,9349	0,9413	0,9454	0,9493	0,9654	0,9585	0,9722	0,9746

Table 3-4 for classical factor analysis with 10 outliers shows that the MLE method is better than others for n=100, p=15, and k=2. The MLE method performs better than others for n=500, p=15, and k=2. The MLE method is better than others for n=1000, p=15, and k=2. The MLE method is better than others for n=100, p=15, and k=3. The MLE method performs better than others for n=500, p=15, and k=3. The MLE method is better than others for n=1000, p=15, and k=3. The GLS method is better than others for n=100, p=15, and k=4. The GLS method performs better than others for n=500, p=15, and k=4. The MLE method is better than others for n=1000, p=15, and k=4.

Considering all classical methods, the ratio of total variance explained and the fit values decrease with each increase in sample size n=100,500 and 1000. When the number of variables increases, the ratio of total variance explained and the fit values increase as the sample size increases n=100,500 and 1000. It can be said that MLE and GLS methods are better than others for classical factor analysis.

Table 3-4 for robust factor analysis with 10 outliers shows that the M method is better than others for n=100, p=15, and k=2. The MCD method performs etter than others for n=500, p=15, and k=2. The MCD method is better than others for n=1000, p=15, and k=2. The MCD method is better than others for n=100, p=15, and k=3. The MCD method performs better than others for n=500, p=15, and k=3. The MCD method is better than others for n=1000, p=15, and k=3. The MCD method is better than others for n=100, p=15, and k=4. The MCD method performs better than others for n=500, p=15, and k=4. The MCD method is better than others for n=1000, p=15, and k=4.

When considering all robust factor analysis methods, it can be seen that the ratio of total variance explained and fit values decreases as the sample size increases n=100, 500 and 1000. However, as the number of variables increases, the ratio of total variance explained and fit values increases for each sample size n=100, 500 and 1000. This suggests that the MCD and M methods are better than others for robust factor analysis. The analysis of the ratio of total variance explained and fit values shows that robust factor analysis methods outperform classical methods. This is particularly important since the dataset used in the analysis includes 10 outliers, indicating that robust factor analysis is a more suitable approach for such data, as supported by theoretical expectations.

Table 3. The Ratio of Total Variance Explained for Classical and Robust Factor Analysis with 10 outliers (1000 repetitions)

		p=15			p=20			p=25			p=30		
Sample Size	**Method**	**k=2**	**k=3**	**k=4**	**k=2**	**k=3**	**k=4**	**k=2**	**k=3**	**k=4**	**k=2**	**k=3**	**k=4**
100	MLE	**0,4069**	**0,4793**	0,4993	**0,4212**	**0,4936**	**0,5222**	**0,4822**	0,5064	0,5302	**0,5227**	**0,5469**	**0,5763**
	PCA	0,4050	0,4785	0,4973	0,4202	0,4911	0,5115	0,4716	0,5052	0,5283	0,5195	0,5324	0,5685
	WLS	0,3794	0,4461	0,5211	0,4205	0,4747	0,5068	0,4723	**0,5093**	**0,5468**	0,5214	0,5386	0,5758
	GLS	0,3905	0,4565	**0,5228**	0,4152	0,4651	0,5043	0,4727	0,5061	0,5465	0,5022	0,5345	0,5694
	MCD	**0,5153**	**0,5712**	**0,6046**	0,5498	0,5823	0,6186	0,5985	0,6218	0,6466	**0,5792**	**0,6137**	**0,6237**
	M	0,5067	0,5517	0,6003	**0,5527**	**0,5911**	**0,6225**	**0,6013**	**0,6234**	**0,6532**	0,5764	0,6131	0,6231
	S	0,5134	0,5491	0,5841	0,5234	0,5579	0,5702	0,5642	0,5829	0,6091	0,5666	0,5886	0,6083
500	MLE	**0,4005**	**0,4538**	0,4842	**0,4185**	**0,4757**	**0,5151**	**0,4470**	**0,4823**	**0,5276**	**0,4681**	**0,5126**	**0,5688**
	PCA	0,3921	0,4529	0,5001	0,4171	0,4744	0,5088	0,4464	0,4814	0,5244	0,4606	0,5086	0,5653
	WLS	0,3736	0,4432	0,5012	0,4033	0,4601	0,5065	0,4342	0,4820	0,5253	0,4623	0,5117	0,5675
	GLS	0,3882	0,4537	**0,5119**	0,4029	0,4623	0,5142	0,4355	0,4813	0,5255	0,4642	0,5094	0,5661
	MCD	**0,5116**	**0,5394**	**0,5507**	**0,5482**	**0,5743**	**0,6035**	**0,5662**	**0,5889**	**0,6226**	**0,5735**	**0,6076**	**0,6233**
	M	0,4911	0,5275	0,5464	0,5473	0,5712	0,6004	0,5636	0,5826	0,6213	0,5714	0,6062	0,6228
	S	0,4923	0,5283	0,5421	0,5034	0,5334	0,5587	0,5427	0,5522	0,5746	0,5520	0,5774	0,5956
1000	MLE	**0,3996**	**0,4382**	**0,4768**	**0,4039**	**0,4565**	**0,4888**	**0,4236**	**0,4751**	**0,5077**	**0,4543**	**0,4974**	**0,5228**
	PCA	0,3723	0,4344	0,4731	0,4005	0,4514	0,4849	0,4206	0,4722	0,5038	0,4514	0,4927	0,5201
	WLS	0,3877	0,4327	0,4742	0,4024	0,4543	0,5858	0,4221	0,4737	0,5065	0,4525	0,4951	0,5216
	GLS	0,3898	0,4366	0,4753	0,4013	0,4552	0,4877	0,4227	0,4740	0,5071	0,4531	0,4962	0,5220
	MCD	**0,4976**	**0,5226**	**0,5362**	**0,5222**	**0,5561**	**0,5744**	**0,5421**	**0,5511**	**0,5672**	**0,5517**	**0,5688**	**0,5844**
	M	0,4897	0,5199	0,5347	0,5227	0,5542	0,5727	0,5416	0,5507	0,5666	0,5513	0,5683	0,5837
	S	0,4888	0,5195	0,5322	0,5029	0,5301	0,5723	0,5401	0,5500	0,5643	0,5507	0,5665	0,5807

Table 4. The Fit Values for Classical and Robust Factor Analysis with 10 outliers (1000 repetitions)

		p=15			p=20			p=25			p=30		
Sample size	**Method**	**k=2**	**k=3**	**k=4**	**k=2**	**k=3**	**k=4**	**k=2**	**k=3**	**k=4**	**k=2**	**k=3**	**k=4**
100	MLE	**0,8122**	**0,8479**	0,8618	**0,8391**	**0,9169**	**0,9382**	**0,9323**	0,9381	0,9493	**0,9543**	**0,9725**	**0,9788**
	PCA	0,8113	0,8478	0,8614	0,8380	0,9154	0,9363	0,9311	0,9372	0,9481	0,9536	0,9683	0,9775
	OLS	0,8116	0,8477	0,8611	0,8377	0,9143	0,9311	0,9317	0,9375	0,9489	0,9540	0,9721	0,9780
	WLS	0,7711	0,8012	0,8828	0,8385	0,9137	0,9359	0,9311	**0,9386**	**0,9504**	0,9551	0,9704	0,9783
	GLS	0,8019	0,8424	**0,8855**	0,8362	0,9132	0,9343	0,9319	0,9377	0,9500	0,9528	0,9697	0,9778
	MCD	**0,9282**	**0,9617**	**0,9777**	0,9562	0,9769	0,9821	0,9794	0,9824	0,9902	**0,9821**	**0,9832**	**0,9953**
	M	0,9273	0,9608	0,9769	**0,9623**	**0,9782**	**0,9834**	**0,9816**	**0,9835**	**0,9914**	0,9816	0,9826	0,9947
	S	0,9265	0,9492	0,9734	0,9518	0,9646	0,9703	0,9645	0,9699	0,9816	0,9769	0,9787	0,9901
500	MLE	**0,8105**	**0,8459**	0,8520	**0,8362**	**0,8944**	**0,9227**	**0,8611**	**0,9005**	**0,9352**	**0,9446**	**0,9623**	**0,9715**
	PCA	0,8100	0,8453	0,8532	0,8351	0,8921	0,9209	0,8607	0,8994	0,9340	0,9403	0,9597	0,9688
	OLS	0,8102	0,8451	0,8556	0,8347	0,8907	0,9213	0,8603	0,8998	0,9343	0,9406	0,9604	0,9694
	WLS	0,8075	0,8454	0,8544	0,8334	0,8884	0,9204	0,8593	0,9002	0,9345	0,9411	0,9616	0,9706
	GLS	0,8084	0,8457	**0,8568**	0,8326	0,8893	0,9222	0,8586	0,8996	0,9349	0,9425	0,9601	0,9691
	MCD	**0,9244**	**0,9584**	**0,9612**	**0,9427**	**0,9652**	**0,9781**	**0,9587**	**0,9743**	**0,9815**	**0,9758**	**0,9781**	**0,9836**
	M	0,9201	0,9567	0,9609	0,9412	0,9623	0,9723	0,9576	0,9734	0,9806	0,9749	0,9777	0,9822
	S	0,9203	0,9576	0,9601	0,9356	0,9588	0,9675	0,9531	0,9671	0,9703	0,9736	0,9756	0,9778
1000	MLE	**0,8014**	**0,8354**	**0,8423**	**0,8267**	**0,8846**	**0,9124**	**0,8515**	**0,8907**	**0,9251**	**0,9348**	**0,9528**	**0,9606**
	PCA	0,7962	0,8323	0,8402	0,8211	0,8820	0,9105	0,8477	0,8884	0,9237	0,9308	0,9499	0,9570
	OLS	0,8003	0,8344	0,8409	0,8236	0,8827	0,9112	0,8483	0,8890	0,9241	0,9316	0,9508	0,9581
	WLS	0,7978	0,8335	0,8412	0,8245	0,8834	0,9117	0,8491	0,8892	0,9243	0,9322	0,9518	0,9588
	GLS	0,7984	0,8351	0,8415	0,8225	0,8837	0,9121	0,8497	0,8898	0,9246	0,9329	0,9523	0,9594
	MCD	**0,9143**	**0,9481**	**0,9510**	**0,9324**	**0,9556**	**0,9680**	**0,9488**	**0,9645**	**0,9717**	**0,9666**	**0,9757**	**0,9795**
	M	0,9135	0,9476	0,9504	0,9317	0,9547	0,9675	0,9479	0,9636	0,9708	0,9659	0,9753	0,9792
	S	0,9126	0,9465	0,9490	0,9251	0,9486	0,9572	0,9410	0,9597	0,9634	0,9651	0,9742	0,9786

4. Real-Life Application

Athletics is one of the most widely followed sports events worldwide. Countries prepare for competitions throughout the year. This study used data obtained before competitions on a country basis for female athletics athletes. In this dataset, 100 m (s), 200 m (s), 400 m (s), 800 m (min), 1500 m (min), 3000 m (min), and marathon (min) values were taken as independent variables. The analysis used data from 55 countries. The data was obtained from https://towardsdatascience.com/factor-analysis-on-women-track-records-data-with-r-and-python-6731a73cd2e0 [18].

Table 5 shows the results of evaluating MLE and GLS techniques for classic factor analysis. The table shows that the ratio of total variance explained and the fit values are higher for the MLE method than for the GLS method. This suggests that the MLE method is a better choice for estimating the factor analysis model. However, the table also shows that the GLS method produces similar results to the MLE method. This suggests that the GLS method is a good choice for estimating the factor analysis model when the MLE method is not available or when the data does not meet the assumptions of the MLE method.

Table 5 also shows that there is some uncertainty regarding which factor the 400m variable belongs to when the GLS method is used. This is because the 400m variable has high loadings on both factors. This suggests that the 400m variable is a measure of both speed and endurance.

The results of the study suggest that the MLE method is a better choice for estimating the factor analysis model. However, the GLS method is a good choice for estimating the factor analysis model when the MLE method is not available or when the data does not meet the assumptions of the MLE method.

Table 5. The results of classic factor analysis (CFA) for athletics data (MLE and GLS)

	CFA (GLS)		CFA (MLE)	
	F1	F2	F1	F2
100 m		0,803		0,811
200 m		0,773		0,760
400 m	**0,556**	**0,558**	0,623	
800 m	0,899		0,910	
1500 m	0,564		0,533	
3000 m	0,691		0,669	
Maraton	0,666		0,634	
Var. Exp.	0,388	0,332	0,389	0,332
Total Var. Exp.	0,720		**0,721**	
Fit Value	0,971		**0,972**	

The graphs in Figure 1 show the distribution of the data for each of the seven variables. The number of outliers detected in the dataset is 17.

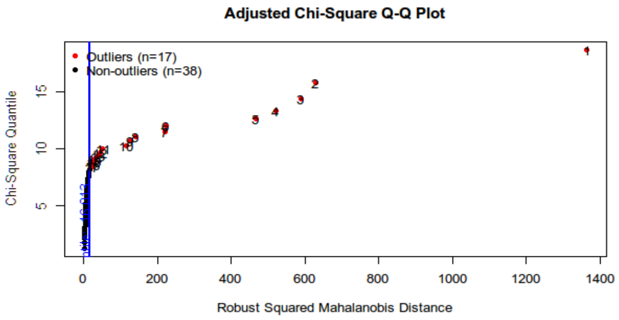

Figure 1. The Adj. Chi-Square Graph for Athletics Data

Based on the results of the robust factor analysis, a 2-factor structure was used and the analysis was continued using the two best methods, MCD and M methods. The results of this analysis are presented in Table 6. The results of the study suggest that the M method is a better choice for factor analysis in this situation. The M method is a robust factor analysis method that is less sensitive to outliers. This is important because the dataset contains outliers. The results of the study also suggest that the 400m variable belongs to the second factor. This is consistent with the theoretical framework of factor analysis. The 400m variable is a measure of both speed and endurance. It is therefore likely to be associated with both factors.

The results of the study can be summarized as follows:

i. The M method is a better choice for factor analysis in this situation.
ii. The 400m variable belongs to the second factor.
iii. The 100m, 200m, and 400m variables belong to the short-distance running factor (F2).
iv. The 800m, 1500m, 3000m, and marathon variables belong to the middle and long-distance running factor (F1).

The results of the study can be used to improve athletic training programs and to help athletes improve their performance. The results suggest that athletes who want to improve their performance in short-distance running should focus on training for speed. Athletes who want to improve their performance in middle and long-distance running should focus on training for endurance.

The results of the study are also interesting from a theoretical perspective. This is an important finding because it suggests that the M method can be used to analyze data that contains outliers.

Table 6. The results of robust factor analysis (RFA) for athletics data (MCD and M)

	RFA (MCD)		RFA (M)	
	F1	F2	F1	F2
100 m		0,906		0,895
200 m		0,902		0,905
400 m		**0,703**		**0,752**
800 m	0,844		0,833	
1500 m	0,936		0,916	
3000 m	0,888		0,872	
Maraton	0,656		0,733	
Var. Exp.	0,491	0,388	0,494	0,418
Total Var. Exp.	0,879		**0,912**	
Fit Value	0,972		**0,973**	

5. Discussion and Conclusion

Factor analysis is a simulation study that is used to test the effectiveness of this method and to determine how accurate the factor analysis results are. These studies are also used to evaluate how factors such as different number of factors, sample sizes or distribution of sample affect the results of factor analysis. It is the results of these studies that help to determine the most suitable conditions for the use of factor analysis methods.

In this study, classical and robust factor analysis methods, and simulation studies carried out in different number of variables, number of factors and sample sizes were evaluated. The study provides valuable insights into the performance of factor analysis methods. The results of the study can be used to help researchers choose the most suitable factor analysis method for their data.

In general, when classic factor analysis is applied to a dataset that is derived from a multivariate normal distribution, the best methods are Maximum Likelihood Estimation (MLE), Weighted Least Squares (WLS), and Generalized Least Squares (GLS). Overall, classic factor analysis has been found to produce better results, likely because the data used in this analysis is derived from a multivariate normal distribution.

On the other hand, the covariance matrix is easily affected by outliers, and the eigenvalue and eigenvector, which are calculated according to the covariance matrix, are sensitive to outliers too, thus leading to deviation in the results.

This study investigated when robust covariance matrix is used that reduces the influence of outliers, the eigenvalue and eigenvector calculated by that are less sensitive to outliers, thus affecting robust factor analysis results.

When robust factor analysis techniques are applied to datasets that contain outliers, they tend to produce better results compared to other techniques. This is because robust methods, such as M and MCD, are designed to handle outliers more effectively and result in a higher the ratio of total variance explained and better-fit values. In situations where outliers are present, using robust factor analysis techniques is likely to produce better results.

Finally, this study investigated robust estimation methods as alternatives to classical estimation. The results of the simulation study show that such methods are available for factor analysis, and give clear evidence that all robust estimation methods under investigation have a high efficiency by outliers. On the contrary, classical factor analysis is strongly influenced by the uncontrolled effects of outliers which makes them often totally unreliable. Especially MCD and M methods turn out to be very appealing estimation methods for robust factor analysis.

Declaration of Ethical Code

In this study, we undertake that all the rules required to be followed within the scope of the "Higher Education Institutions Scientific Research and Publication Ethics Directive" are complied with and that none of the actions stated under the heading "Actions Against Scientific Research and Publication Ethics" are not carried out.

References

[1] Pison, G., Rousseeuw, P. J., Filzmoser, P., Croux, C. 2003. Robust Factor Analysis. Journal of Multivariate Analysis, 84(1), 145-172.

[2] Er, F., Sönmez, H. 2006. Öğrenci Başarı Notları İçin Robust Faktör Analizi Uygulaması. Anadolu Üniversitesi Bilim ve Teknoloji Dergisi, 7(1), 149-155.

[3] Browne, M. W., Shapiro, A. 1988. Robustness of normal theory methods in the analysis of linear latent variable models. British Journal of Mathematical and Statistical Psychology, 41, 193-208.

[4] Mooijaart, A., Bentler, P. M. 1991. Robustness of normal theory statistics in structural equation models. Statistica Nederlandica, 45, 159-171.

[5] Johnson, R. A., Wichern, D.W. 2007. Applied Multivariate Statistical Analysis. Fifth Edition, Pearson Education Int., New Jersey.

[6] Rencher, A. C. 2002. Methods of Multivariate Analysis. Second Edition, John Wiley & Sons, Inc.

[7] Jennrich, R. I., Robinson, S.M. 1969. A Newton-Raphson Algorithm for Maximum Likelihood Factor Analysis,.Psychometrika, 34, 111 -123.

[8] Jöreskog, K. G. 1967. Some Contributions to Maximum Likelihood Factor Analysis. Psychometrika, 32, 443-482.

[9] Jöreskog, K. G., Goldberger, A.S. 1972. Factor Analysis by Generalized Least Squares. Psychometrika, 37, 243.

[10] Lee, S. Y. 1978. The Gauss-Newton Algorithm for the Weighted Least Squares Factor Analysis. Journal of the Royal Statistical Society: Series D (The Statistician), 27, 103-114.

[11] Revelle, W. 2022. How To: Use the psych package for Factor Analysis and data reduction. R package, R Core Team, 1-95.

[12] Rousseeuw, P. J., Van Driessen, K. 1999. A fast algorithm for the minimum covariance determinant estimator. Technometrics, 41(3), 212-223.

[13] Todorov, V., Filzmoser, P. 2009. An Object-Oriented Framework for Robust Multivariate Analysis. Journal of Statistical Software, 32(3), 2-47.

[14] Fan, J., Wang, W., Zhong, Y. 2016. Robust Covariance Estimation for Approximate Factor Models. arXiv:1602.00719v1, 1-31.

[15] Davies, P. L. 1987. Asymptotic Behavior of S-Estimators of Multivariate Location Parameters and Dispersion Matrices. The Annals of Statistics, 15, 1269–1292.

[16] Lopuhaa, H. P. 1989. On the Relation Between S-Estimators and M-Estimators of Multivariate Location and Covariance. The Annals of Statistics, 17, 1662–1683.

[17] Törmanen, J. 2012. Systems intelligence inventory. Student Project, Master's thesis, Aalto University School of Science.

[18] Pramodithha, R. 2023. Web Page Access Adress: https://towardsdatascience.com/factor-analysis-on-women-track-records-data-with-r-and-python-6731a73cd2e0