Bachelor's thesis

NTNU
Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Mathematical Sciences

Martinius Theodor Singdahlsen

# Rotations in Factor Analysis

Bachelor's thesis in BMAT
Supervisor: Robert Brian OHara
Co-supervisor: Bert van der Veen
December 2022

NTNU
Norwegian University of
Science and Technology

Martinius Theodor Singdahlsen

# Rotations in Factor Analysis

Bachelor's thesis in BMAT
Supervisor: Robert Brian OHara
Co-supervisor: Bert van der Veen
December 2022

Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Mathematical Sciences

NTNU
Norwegian University of
Science and Technology

**Abstract**

This thesis will concern factor analysis with its main focus on the rotations of the factor model. Rotations are applied to both be able to draw inference and to better interpret the data. In this context we will look at the criteria constructed for the desired solution of the rotations, and different methods of how we can rotate a model applied with factor analysis. This includes both the orthogonal rotation criteria Quartimax and Varimax, as well as the oblique rotation criteria Yates' Geomin and Simplimax. After the rotations are defined we will look at rotations with two factors such that it is possible to visualize the rotations in $\mathbb{R}^2$. Some results will be shown of a simpler rotation criteria, named Component loss criteria that is in some sense related to the Simplimax rotation.

# Contents

# Chapter 1

## 1.1 Introduction

In the social and natural sciences, factor analysis is a commonly applied statistical method for the analysis of multivariate data [1, . p405]. With factor analysis we strive to explain observed variables by underlying (latent) variables that cannot be observed. The most important aspect of factor analysis is to understand that the observed variables are being caused by the latent variables. In this way an underlying structure of the observed variables are being discovered when a factor analysis is performed. If the observed variables are chosen with caution one can find that far less factors are required to explain the observed variables. Thus its applications has seen popular use in psychology, where for example there are multiple aspects of a persons personality that cannot be directly measured, however some of the persons personality traits can be captured by carefully constructed tests, and the result can therefore be subjected to a factor analysis.

The reasoning behind why the factor model is applied can be separated into two categories, exploratory factor analysis and confirmatory factor analysis [2, . p5]. Confirmatory factor analysis concerns the case of testing whether our hypothesis is supported by what we have observed. Thus hypothesis testing is of greater interest in confirmatory factor analysis. For example if we want to show that the performance of children in school of 5'th grade cannot be solely determined by the child's motivation, we can construct observed variables of test questions the 5'th graders took in their seven subjects in school. Then the factor model can be applied with one factors and a hypothesis test of the fit of the model can be applied. Thus we may conclude that the data does not support our hypothesis.

As the name exploratory factor analysis might suggest, the goal of the researcher is to apply the model to find meaning full underlying structures that explain the observed variables well. Thus we can develop new theories of what the observed variables are measuring [3, . p4]. However, in general the factor model does not have an unique solution, thus rotations which seek to find better interpretations for the researcher or facilitate for performing inference is done. This thesis concerns different types of rotations and what we are looking for in the new solution found. In the following section I will define the model used in factor analysis and show some important

results that we will use to apply the model to a dataset.

### 1.1.1 The Factor model

**Definition 1.1.1** (The Factor model)**.** Let $X(p \times 1)$ be a random vector of normally distributed variables where $\mu(p \times 1)$ is its mean, and $\Sigma(p \times p)$ is the covariance matrix of $X$. The factor model can be expressed as,

$$X = \Lambda f + \mathrm{e} + \mu. \tag{1.1}$$

Where $f(k \times 1)$ is a random normal vector that has the factors of the model as its elements, $\mathrm{e}(k \times 1)$ is a random normal vector of error variables and $\Lambda(p \times k)$ is a matrix of constants that is referred to as the matrix of factor loadings or factor pattern matrix [4, . p256]. Further, we will assume the following of the components of the factor model,

$$\mathbb{E}(\mathrm{e}) = 0 \text{ and } \mathbb{E}(f) = 0, \tag{1.2}$$

this way the expectation of the variable $X$ represented by the factor model is expressed by $\mu$. Lastly it is assumed that,

$$Cov(\mathrm{e}_i, \mathrm{e}_j) = 0 \text{ for all } i \neq j \text{ and } Cov(f_i, \mathrm{e}_j) = 0 \text{ for all } i, j. \tag{1.3}$$

By assuming the above we acquire that each $\mathrm{e}_i$ represents the unique variance of each corresponding factor $f_i$[5, . p168]. Also we see that the factors may be correlated to each other. The correlation matrix is given by,

$$Var(f) = \Phi, \tag{1.4}$$

where the diagonal of this matrix only consists of ones.

If a data matrix, $\mathbb{X}\ (p \times n)$ is given, where $p$ attributions have been measured on $n$ individuals, and the assumptions of our data is as given in definition of the factor model. Then each row of the data matrix represents a sample from a random normal distribution, $X_i \sim \mathcal{N}(\mu_i,\, \sigma_i^2)$ for $i = 1, 2, ..., p$. Let $x_{st}$ be the entries of $\mathbb{X}$, we can then write the classical factor model, which express each individuals measurement through the factor model as [6, . p15],

$$x_{st} = \sum_{i=1}^{k} \lambda_{si}[f_i]_t + [\mathrm{e}_s]_t + \mu_s \qquad s = 1, 2, ..., p\ t = 1, 2, ..., n.$$

Here the notation of brackets is used as to clarify that it is a sample of an individual $s$, from the random variable inside of the bracket and $\lambda_{ij}$ is an element of the factor loading matrix $\Lambda$. Each summand represents the contribution of its respective factor, wihle $[\mathrm{e}_s]_t$ represents the individual variance. Further more we can observe that the expected value stays constant for each observation of an attribute. This reflects how the factor model describes the underlying structure of the observed variables, hence it does not affect its expected value when the same attributions are measured.

By the definition given in (1.1.1), the Factor model lacks an unique solution, so we first make stricter assumptions to acquire one unique solution apart from the signs of the factors. The solution then obtained is called the initial solution. Specifically when fitting the model, I will further assume that,

$$Var(f) = I, \tag{1.5}$$

$$\Lambda^\intercal \Psi^{-1} \Lambda \text{ is diagonal.} \tag{1.6}$$

To show why we need further more criteria than (1.5) consider the following. Assume $M(p \times p)$ is an orthogonal matrix and that $z = Mf$. The following model bellow also satisfies definition (1.1.1), however has a different design matrix:

$$x = (\Lambda M^\intercal)z + \mathrm{e} + \mu. \tag{1.7}$$

We can see that $Var(z) = MM^\intercal = I$, thus (1.5) still holds. Thus, further constraints need to be implemented so there is only one unique solution when maximizing the likelihood. It will later be shown why this condition does produce an unique solution in theorem (2.1.1).

Note that there are other popular restrictions used instead of the latter restriction that also produces one unique solution. An example of another restriction that may be imposed is, $\Lambda^\intercal D \Lambda$ where $D = diag(Var(f_1), Var(f_2), ..., Var(f_p))$ [4, . p258].

Later on we will look at how we can find different solutions that satisfy definition 1.1.1 where (1.5) and (1.6) are not necessarily satisfied. However (1.5) will be assumed to hold from from this point and all the way through chapter three. Thus we will have solutions with uncorrelated factors. However in chapter four we will consider the case when the factors are allowed to be correlated, thus $Var(f) \neq I$. We will now derive a few straight forward results that will be extremely usefull results. These results will later be used to estimate $f$, $\Lambda$ and $\mathrm{e}$. We are often interested in the variance of variables and how they are correlated, therefore it might be obvious to look at how we can express the variance of $X$ in terms of $f$, $\Lambda$ and $\mathrm{e}$:

$$\begin{aligned} Var(X) &= Var(\Lambda f + \mathrm{e} + \mu) \\ &= Var(\Lambda f + \mathrm{e}) \\ &= Var(\Lambda f) + Var(\mathrm{e}) \\ &= \Lambda^\intercal Var(f) \Lambda + Var(\mathrm{e}) \qquad (1.8) \\ &= \Lambda\Lambda^\intercal + \Psi. \qquad (1.9) \end{aligned}$$

From the definition of the model we have that $Cov(\mathrm{e}_i, \mathrm{e}_j) = 0$. This implies that the covariance matrix of $\mathrm{e}$ is a diagonal matrix. We define $Var(\mathrm{e}) = \Psi$ where $[\Psi]_{ii} = \psi_{ii}^2$. Since $\Lambda$ is a $(p \times k)$

matrix and $\Psi$ is a $(p \times p)$ matrix, we obtain the covariance matrix of $X$ expressed in terms of the variables of the factor model that is $(p \times p)$. The variance of each $x_i$ for $i = 1, 2, .., p$ can then be expressed as the following,

$$\sigma_{ii} = \sum_{j=1}^{k} \lambda_{ij}^2 + \psi_{ii}^2. \tag{1.10}$$

The variance of $X$ can thus be split into two parts. The $commonality$ [4, . p257] which is the following part of the above expression, $\sum_{j=1}^{k} \lambda_{ij}^2$. We will later use the commonality to normalize the loading matrix to find a better rotation. The commonality represents the shared variance between the $x_i$ and all the factors $f_i$ for $i = 1, 2, .., p$. This tells us how much $x_i$ depends on the other $x_i$'s since they are also expressed in terms of the factors of the model. This is reflected by the result $Cor(x_i, x_j) = \lambda_{ij}^2$. It is therefore clear that the the commonality is the sum of shared variance between the $x_i$ and all the factors $f_i$ for $i = 1, 2, .., p$. On the other hand $\psi_{ii}^2$ is the individual variance of each $x_i$ called specific or unique variance. This means that each $\psi_{ii}^2$ is the variance of $x_i$ not shared with other observable variables.

We can find a measure of how the factor model simplifies the observed variables by considering how many parameters the factor model we will fit with $k$ factors and $p$ observed variables. we have that $\Sigma$ has $\frac{p}{2}(p+1)$ unique elements, while $\Lambda$ and $\Psi$ represents $pk+p$ parameters. Since we will also assume that equation (1.6) holds which is $\Lambda^\intercal \Psi^{-1} \Lambda$ is diagonal. This introduces further $\frac{k}{2}(k-1)$ constraints to the model. Therefore the reduction of parameters when applying the factor model will be,

$$\begin{aligned} s &= \frac{p}{2}(p+1) - [pk + p - \frac{k}{2}(k-1)] \\ &= \frac{1}{2}(p-k)^2 - \frac{1}{2}(p+k). \end{aligned} \tag{1.11}$$

# Chapter 2

## 2.1 Finding an initial solution

Multiple methods have been developed to obtain an initial solution to the factor model. In this thesis we will only discuss one of these methods, namely the maximum likelihood fit of the factor model. Another popular method for fitting the factor model is closely related to principal component analysis and is done by diagonalizing the symmetrical matrix $\Sigma - \Psi$ after the elements of $\Psi$ are approximated [4, . p261]. This procedure is called Principal factor analysis.

As previously stated we are assuming for the sake of finding an initial solution that equations (1.5) and (1.6) hold. Constraint (1.6) cannot be applied if any $\psi_{ii} = 0$, when this situation occurs other criteria are imposed to arrive at a solution[4, p. 258]. However in this thesis it will be assumed that the unique variance of all factors are strictly positive, such that we do not concern this case. However the case of at least one $\psi_i = 0$ are not a rare occurrence and has the name, Heywood case. It resembles a boundary problem when the maximum likelihood method is applied.

We will find $\Lambda$ and $\Psi$ by considering the marginal log likelihood function of $X$. Assume that we have $n$ observations of $X$. We naturally choose the estimator $\bar{\mu} = \bar{X}$ as the estimator for the expectancy. This results in the log likelihood function to become [5, . p173],

$$l = -\frac{n-1}{2}[\log|\Sigma| + tr(\Sigma^{-1}S)] + K. \tag{2.1}$$

Here $S$ is the estimated covariance matrix of $X$ and $K$ is some function that is independent of $\Sigma$. We will further use $\Sigma$ as a parameter, such that it can vary over the whole parameter space. It can be shown that the maximum likelihood estimates are scale invariant, therefore replacing the covariance matrix with the correlation matrix $R$ does not alter the solution [5, . p173]. Further we will alter the log likelihood function such that it performs to our advantage. Consider the function:

$$F(\Lambda, \Psi) = \log|\Sigma| + tr(\Sigma^{-1}S). \tag{2.2}$$

To arrive at (2.2) from (2.1) we have removed the constant that is independent of $\Sigma$, and multiplied by $-\frac{2}{n-1}$. Multiplying by a negative value results in the global maximum of (2.1) to have the same solution as the global minimum of $F(\Lambda, \Psi)$. It is worth noting that $F(\Lambda, \Psi)$ is often given by $F(\Lambda, \Psi) = log|\Sigma| + tr(\Sigma^{-1}S) + log|S| + p$. This has to do with hypothesis testing of the fit of the model when $p$ factors are chosen, and since it will not be covered in this thesis I have chosen a simpler function to minimize. Our problem now boils down to solving (2.2) for the parameters $\Lambda$ and $\Psi$. This can be done in two steps. First we minimize (2.2) over $\Lambda$ for fixed $\Psi$, whereafter we minimize (2.2) over $\Psi$ [4, . p264]. The first step can be done analytically, however minimizing over $\Psi$ has to be done numerically. I will not further consider the numeric approach for finding $\Psi$.

**Theorem 2.1.1.** *If $\psi_{ii} > 0$ is fixed and we define $S^* := \Psi^{-1/2} S \Psi^{-1/2}$. Consider the eigendecomposition of $S^*$:*

$$S^* = \Gamma C \Gamma^\intercal,$$

*where $\Gamma = (\gamma_1, \gamma_2, ..., \gamma_k)$ is an orthogonal matrix and $C = diag(c_1, c_2, ..., c_k)$ are the eigenvalues of $S^*$. Then $\Lambda^* = \Psi^{-1/2}\Lambda$ and the value of $\Lambda = \Psi^{-1/2}\Lambda^*$ that satisfies (1.6) and minimizes $F(\Lambda, \Psi)$ is given by $\Lambda_i^* = r_i \gamma_i$ where $r_i = [max(c_i - 1, 0)]^{1/2}$ for $i = 1, 2, ..., k$*

*Proof.* Since the maximum likelihood estimates are scale invariant we can define
$S^* := \Psi^{-1/2} S \Psi^{-1/2}$. where $\Psi^{-1/2}$ is the square root of the inverse of the elements of $\Psi$ since it is a positive definite diagonal matrix. We know that $\Lambda^* \Lambda^{*\intercal}$ must be of rank k since $\Lambda(p \times k)$. Consider the eigendecomposition of $\Lambda^* \Lambda^{*\intercal}$ then,

$$\Lambda^* \Lambda^{*\intercal} = G B G^\intercal, \tag{2.3}$$

where $G = (g_1, g_2, ..., g_k)$ has standard unit columns and $B = diag(b_1, b_2, ..., b_k)$ are positive eigenvalues. Notice how the eigenvalues of $\Lambda^* \Lambda^{*\intercal}$ which is a function of $\Psi$ are determined by $B$. To continue the derivation of $\Lambda$ we will let the two arbitrary orthogonal matrices $G$ and $\Gamma$ be expressed as:

$$\begin{aligned} M &= G\Gamma^\intercal \\ G &= \Gamma M^\intercal. \end{aligned}$$

Since $G$ and $\Gamma$ are both orthogonal matrices, their product is an orthogonal matrix. This way we will minimize $F(\Lambda, \Psi)$ over arbitrary choices of $M$ and $B$. First we will find $\Sigma^{*-1}$ and $|\Sigma^*|$ with respect to $M$ and $B$,

$$\begin{aligned} \Sigma^* &= \Psi^{-1/2} \Sigma \Psi^{-1/2} \\ &= \Lambda^* \Lambda^{*\intercal} + I \\ &= \Gamma M^\intercal B M \Gamma^\intercal + I \\ &= \Gamma (M^\intercal B M + I) \Gamma. \end{aligned} \tag{2.4}$$

From this equation we can extract the determinant of $\Sigma^*$ since we know that the determinant of a matrix is equal to the product of all its eigenvalues. Therefore we obtain that:

$$|\Sigma^*| = \prod_{i=1}^{k}(b_i + 1).$$

The inverse can be found from (2.4) and is as follows:

$$\Sigma^{*-1} = \Gamma(I - M^\intercal B(B + I)^{-1}M)\Gamma^\intercal.$$

Substituting into (2.2) thus gives us:

$$\begin{aligned} F(\Lambda, \Psi) &= \sum_{i=1}^{k} \log(b_1 + 1) - tr[M^\intercal B(I + B)^{-1}MC] + D \\ &= \sum_{i=1}^{k} log(b_1 + 1) - \sum_{i=1}^{k} \frac{b_i c_i}{1 + b_i} m_{ii}^2 + D. \end{aligned}$$

Here $D$ only depends on $\Psi$, therefore it is a constant. The minimum of this function is clearly obtained when $M$ is the identity matrix, such that the problem simplifies to minimizing,

$$F(\Lambda, \Psi) = \sum_{i=1}^{k} \log(b_1 + 1) - \sum_{i=1}^{k} \frac{b_i c_i}{1 + b_i}.$$

The minimum of $F(\Lambda, \Psi)$ with respect to $b_i$ occurs when $b_i = max(c_i - 1, 0)$. Since we found that $M = I$ this implies that $G = \Gamma$ and we get from (2.4) that $\Lambda$ has to satisfy,

$$\Lambda^* \Lambda^{*\intercal} = \Gamma B \Gamma^\intercal.$$

To produce an unique solution the criteria (1.6) is used. Thus the theorem follows. □

### 2.1.1 Fitting factor analysis to data from Schmid and Leiman (1957), and Harman H. H 8 physical variables

In this thesis I have chosen to model two datasets both with two factors. The reasoning behind the choice of two factors is such that we can plot the loadings of the design matrix for visualization. Then in chapter three and four we can apply rotations, which finds different solutions that aim to find better interpretations or a solution that allows for drawing inference easier. This way we can compare the different rotations.

The first dataset I used is from Schmid and Leiman (1957), which can be found in the `psych` R-package [7] and was first published in the paper "The development of hierarchical factor solutions" [8, .p 57]. The dataset has twelve overlying variables that were constructed artificially to show the Schmid-Leiman transformation. The correlation matrix is given below to give an insight to the data. This dataset will be referred to as 12 Artificial variables.

| | V1 | V2 | V3 | v4 | V5 | V6 | V7 | v8 | V9 | V10 | V11 | v12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| V1 | 1.000 | - | - | - | - | - | - | - | - | - | - | - |
| V2 | 0.720 | 1.000 | - | - | - | - | - | - | - | - | - | - |
| V3 | 0.314 | 0.353 | 1.000 | - | - | - | - | - | - | - | - | - |
| V4 | 0.269 | 0.302 | 0.420 | 1.000 | - | - | - | - | - | - | - | - |
| V5 | 0.098 | 0.111 | 0.075 | 0.064 | 1.000 | - | - | - | - | - | - | - |
| V6 | 0.049 | 0.055 | 0.038 | 0.032 | 0.320 | 1.000 | - | - | - | - | - | - |
| V7 | 0.129 | 0.145 | 0.099 | 0.085 | 0.134 | 0.067 | 1.000 | - | - | - | - | - |
| V8 | 0.037 | 0.042 | 0.028 | 0.024 | 0.038 | 0.019 | 0.140 | 1.000 | - | - | - | - |
| V9 | 0.290 | 0.327 | 0.222 | 0.190 | 0.109 | 0.054 | 0.143 | 0.041 | 1.000 | - | - | - |
| V10 | 0.161 | 0.181 | 0.124 | 0.106 | 0.060 | 0.030 | 0.079 | 0.023 | 0.450 | 1.000 | - | - |
| V11 | 0.064 | 0.073 | 0.049 | 0.042 | 0.024 | 0.012 | 0.032 | 0.009 | 0.146 | 0.081 | 1.000 | - |
| V12 | 0.075 | 0.085 | 0.058 | 0.049 | 0.028 | 0.014 | 0.037 | 0.011 | 0.170 | 0.094 | 0.420 | 1.000 |

| | Factor one | Factor two | Unique variance | commonality |
|---|---|---|---|---|
| V1 | 0.795 | -0.167 | 0.340 | 0.660 |
| V2 | 0.865 | -0.163 | 0.225 | 0.775 |
| V3 | 0.429 | 0.038 | 0.815 | 0.185 |
| V4 | 0.374 | 0.035 | 0.859 | 0.141 |
| V5 | 0.148 | 0.080 | 0.972 | 0.028 |
| V6 | 0.079 | 0.051 | 0.991 | 0.009 |
| V7 | 0.188 | 0.091 | 0.956 | 0.044 |
| V8 | 0.057 | 0.033 | 0.996 | 0.004 |
| V9 | 0.486 | 0.586 | 0.420 | 0.580 |
| V10 | 0.298 | 0.480 | 0.681 | 0.319 |
| V11 | 0.119 | 0.201 | 0.945 | 0.055 |
| V12 | 0.136 | 0.219 | 0.934 | 0.066 |
| SST | 2.129 | 0.738 | | |

**Table 2.1:** Initial solution of factor analysis fitted to correlation matrix of 12 Artificial variables

Fitting a factor analysis with two latent variables to the correlation matrix gives table (2.1). The factor loadings, commonality and the unique variance is given in table (2.1).

Each factor column of table (2.1) and (2.2) has a summary statistic that shows the sum of squares of the factor loadings. This is how much variance each factor cumulatively describes the overlying variables. This statistic can be very useful if the researcher wants to pick out the factors that mainly describe the overlying variables, much like principal component analysis. The plot of the factor loadings are plotted such that each point represents an overlying variable. Its first coordinate represents the loading of the first factor. Likewise the second coordinate represents the loading of the second factor. It is worth noting that we have fitted a factor model to a correlation matrix where the factors are assumed to be standard normal and independent. We know that the distance that the point is from the origin represents the square root of the commonality of the overlying variable. Since we know that $1 = \sigma_{ii}$, we can visually see that the choice of two factors produces overlying variables where the unique variance is substantial. Be aware that this only holds with orthogonal rotations since the commonality is not equal to the sum of squares

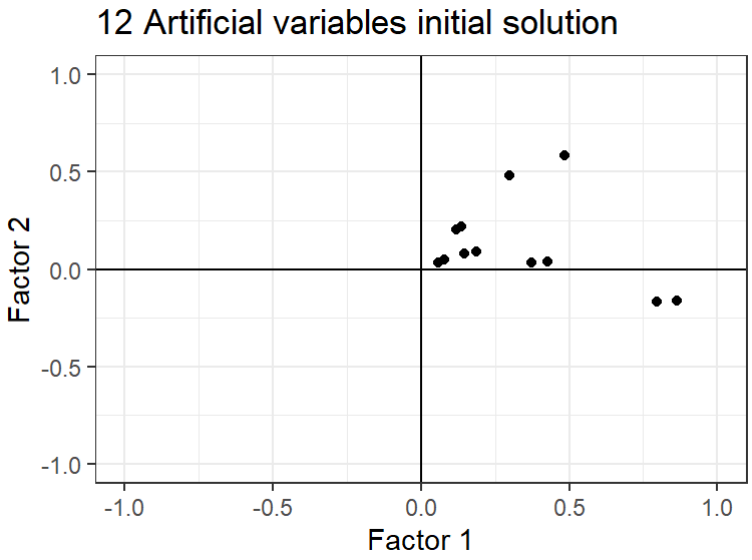



**Figure 2.1:** Plot of the factor loadings, $(\lambda_{i1}, \lambda_{i2})$, from table 1.1

of the loadings corresponding to the overlying variable when an oblique rotation is performed. If we were interested in finding a factor model where the factors would reasonable explain the overlying variables, it would be in interest to consider a factor model with more than two factors. However we will use this visualization as it does not produce a solution where there are clear groupings of the loadings.

On the other hand we will also consider the loading matrix of Harman, H. H 8 physical variables which has two clear groupings of factor loadings. This factor loading matrix can be found in `GPArotation` `R`-package [9] and originates form the book Modern factor analysis [6]. The original dataset has eight overlying variables which measure physiological variables. The dataset has been fitted with two factors, and we include it to demonstrate the different rotations we can achieve. The statistics of the fit can be seen in table (2.2) and the visualization in figure (2.2).

When the factor model with two factors is fitted to the dataset Harman, H. H 8 physical variables it produces a good fit. This can bee seen as the overlying variables have substantially large commonalities, which indicate that the factors explain the overlying variables well. In addition we can see from figure (2.2) that there are two clear groupings of the data. Later on we will see how this will give us great solutions of oblique rotations.

| | Factor one | Factor two | commonality |
|---|---|---|---|
| Height | 0.830 | -0.396 | 0.846 |
| Arm span | 0.818 | -0.469 | 0.889 |
| Length of forearm | 0.777 | -0.470 | 0.825 |
| Lenght of lower leg | 0.798 | -0.401 | 0.798 |
| Weight | 0.786 | 0.500 | 0.868 |
| Bitrochanteric diameter | 0.672 | 0.458 | 0.661 |
| Chest girth | 0.594 | 0.444 | 0.550 |
| Chest with | 0.647 | 0.333 | 0.529 |
| SST | 4.439 | 1.526 | |

**Table 2.2:** Initial solution of factor analysis fitted to correlation matrix of Harman, H. H8 Physical variables

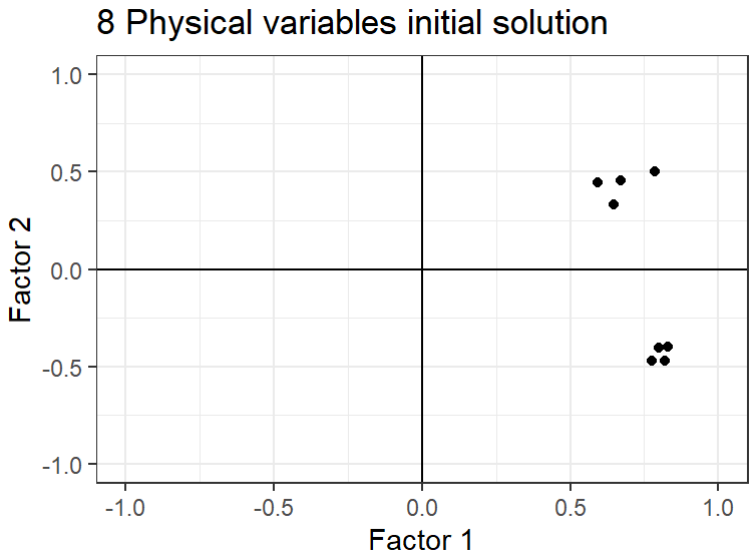



**Figure 2.2:** Plot of the factor loadings from table 1.2

# Chapter 3

## 3.1 Introduction to rotation and orthogonal rotations

We have seen that to obtain a solution to the factor model we assume that the factors are independently identically distributed with unit variance and that $\Lambda^\intercal \Psi^{-1} \Lambda$ is diagonal is retained (this is assumptions (1.5) and (1.6)). In this section we will look at other solutions to the same model, however they will not necessarily satisfy the restraint of (1.6). This is of interest to us as in exploratory factor analysis interpretation is of great importance, as some solutions might simply be easier to interpret than others. Later in this section we will cover what we are looking for in a solution and different methods developed to find the loading matrix that seek to best describe the underlying structure. To find a different solution we will perform a rotation by a matrix $M$ that is orthogonal. Let us first consider If $M(k \times k)$ is a non-singular matrix and we assume the two conditions below hold:

$$\begin{aligned} f_0 &= M^{-1} f, \\ A &= \Lambda M. \end{aligned} \tag{3.1}$$

From these assumptions we obtain the results below:

$$\begin{aligned} \Phi_0 &= M^{-1} M^{-1\intercal}, \\ X &= A f_0 + \mathrm{e} + \mu = \Lambda f + \mathrm{e} + \mu, \\ \Sigma &= A \Phi_0 A^\intercal + \Psi = \Lambda \Lambda^\intercal + \Psi. \end{aligned} \tag{3.2}$$

As we can see from the result above, the variance of the overlying variables is not altered after a rotation has been performed. This is important to point out as performing a rotation redistributes the variance that has already been discovered in the data [3, . p11]. We can see this as the covariance matrix of the new solution found is equal to the initial solution, however we have obtained a new design matrix.

Rotations are divided into two different types of rotations, orthogonal and oblique rotations. An orthogonal rotation is done with an orthogonal matrix $M$, while an oblique rotation is a less stricter form of rotation where the matrix $M$ is only assumed to be non singular. We will first look at the case of the orthogonal rotations. Since we assume that the matrix $M$ is orthogonal when performing an orthogonal rotation we have a special case of equation (3.3).

$$\begin{aligned} \Phi_0 &= M^{-1}M^{-1\intercal} \\ &= I. \end{aligned} \tag{3.3}$$

Thus, by performing an orthogonal rotation the factors are still uncorrelated with unit variance. This is not the case for oblique rotations.

### 3.1.1 Simple Structure Criteria

As there are infinitely many ways of rotating a set of solutions there has been developed a criterion to follow for finding a rotation. This criterion's purpose is to solve the problem of indeterminacy while generating a meaningful solution. A lot of different guidelines for rotations have been proposed, however Thurstone's simple structure criteria has had the greatest impact on the development of rotation, and was developed in 1947 [10, . p335]. Thurstone's simple structure criteria reformulated [6, . p98] are the following five points.

1. Each row of the matrix should have at least one zero.

2. If there are q factors, each column of the matrix should have at least q zeros.

3. For every pair of of columns in the matrix there should be several variables whose entries vanish in one column but not in the other.

4. For every pair of columns in the matrix, a large proportion of the variables should have vanishing entries in both columns when there are four or more factors.

5. For every pair of columns in the matrix there should only be a small number of variables with non-vanishing entries in both columns.

Thurstone describes explicitly that many vanishing entries are desirable when a rotation is performed. And if an orthogonal rotation is performed and we force many loadings to be zero, we will desirably end up with large non zero loadings because of the limitations imposed by an orthogonal matrix. If this situation occurs the authors Mardia, Kent and Bibby, explane this quite well [4, . p268]: "The interpretation of the factor loadings is the most stralghtforward if each variable is loaded highly on at most one factor. and if all the factor loadings are either large and

positive or near zero, with few intermediate values. The variables are then split into disjoint sets, each of which is associated with one factor, and perhaps some variables are left over. A factor j can then be interpreted as an average quality over those variables i for which $\lambda_{ij}$ is large."

It is important to stress that this list is not an absolute truth. The researcher might be looking after a type of rotation that does not satisfy criteria 1, that each row should have at least one zero. This might happen in a situation such as when the overlying variables represent different types of tests and we expect there to be no interfactor correlation. Here some of the tests might be dependent upon all the factors thus creating columns where we do not wish to have zero entries.

Thurstone's simple structure criterion is to be able to visualize the factor loadings in a better way and to solve the problem of indeterminacy. Another way to formulate this is that, Thurstone's simple structure criteria has the main goal of reducing the complexity of the initial loading matrix. To better be able to communicate the different types of relationships between the factors and overlying variables we can have, we will define the most common used definitions for specific factor loading matrices.

Variable complexity is the number of nonzero entries a variable has in the factor pattern matrix [11, . p174]. Since the $r$'th overlying variable is represented by factor loadings of the $r$'th row, the complexity of this overlying variables is the number of nonzero entries in the $r$'th column. This brings us to three classifications of factor loading matrices that we will use, *perfect simple structure*, *approximate simple structure* and *complex structure*. Here the complexity in an intuitive sense is ordered in increasing order, thus as the name perfect simple structure suggests is of least complexity. Perfect simple structure is defined as a factor pattern matrix where all overlying variables have Variable complexity one. Further more approximate simple structure is when the cross loadings are smaller than $0.3$ ($|\lambda_{ij}| < 0.3$). Cross loadings of a factor pattern matrix are all loadings that are smaller than the greatest loading of each respective column. Thus one can view the factors represented by the cross loadings as being the less influencial factors to their respective observed variable. Thus approximate simple structure means that in each row there is only one entry with a value larger than $0.3$. The next step where the factor pattern is even more complex is when at least one of the cross loadings are greater than $0.3$, this is referred to as a complex factor pattern.

In this thesis we will only discuss analytical criteria to arrive at a rotation. The criteria comes in the form of a complexity function that to achieve a solution you minimize or maximize a function $f(\Lambda_0)$, that is a function of its factor loadings. It is possible to use the same criteria to orthogonal rotations as to oblique rotations. We will see that the oblique rotations better optimize the complexity functions, however it is more difficult to find a solution to the oblique rotations. This is because of the restrictions imposed with the different types of rotations.

It is important to say that the complexity functions are designed to try to achieve a solution that satisfies the simple structure criteria as closely as possible. We will see that it is quite hard to satisfy all the points of the simple structure criteria and that different types of rotations focus on different points of the simple structure criteria. For example the first rotation we are going to

define, Quartimax focuses on creating as many zero entries as possible. This is point number 3 and 4 from the simple structure.

### 3.1.2 Quartimax rotation

For the remainder of this chapter we will perform an orthogonal rotation so equation (3.1) will hold where M is an orthogonal matrix. Further more the notation $[\Lambda]_{ij} = b_{ij}$, $[\Lambda_0]_{ij} = a_{ij}$ and $[M]_{qj} = m_{qj}$ from (3.1) will be used to denote the entries of the matrices.

When looking at the simple structure criteria it is clear that we want as many entries of the loading matrix possible to vanish. To develop a method to find a solution that the simple structure criteria describes we need to find some expression we can manipulate to find the best possible fit. A way of doing this is to make a function that "catches up" when the loadings of the rotated structure matrix vanish. This criteria would try to rotate the initial solution in a way such that the third criteria of the simple structure criteria would be satisfied. To arrive at such a solution we make the following observation.

Observe that when a point $p = (x_1, x_2, ..., x_d) \in X \subset R^d$ where $R^d$ is equipped with an orthonormal basis approaches one of the basis vectors we get that $x_j \times x_i$ approaches zero. The restriction that $X$ has to be bounded has to be made, and this is the case as we have finite variance. We can use this observation to our advantage and set up the following expression that we would like to minimize, as it forces the loadings of the rotated factor pattern to become zero. We would like to use this as our complexity function, however it turns out to be to mathematically inconvenient to solve analytically so we will derive a simpler but equivalent complexity function. Consider the equation below:

$$N(A) = \sum_{i=1}^{p} \sum_{j<q=1}^{k} (a_{ij}a_{iq})^2. \tag{3.4}$$

It does not matter if the signs of the resulting factors pattern of our rotation are negative as this would simply represent the factor to be negatively correlated to its corresponding observed variable. However we square the loadings of the rotated matrix such that we do not allow any terms inside of the expression to be negative, as this would result in the complexity function to prefer $a_{ij}a_{iq}$ to be as large as possible and negative above vanishing factor loadings. A comparison to this would be the linear model which minimizes the sum of squares. To arrive to this equation we will consider the following expression,

$$(\sum_{j=1}^{k} a_{ij}^2)^2 = \sum_{j=1}^{k} a_{ij}^4 + 2 \sum_{j<q=1}^{k} (a_{ij}a_{iq})^2. \tag{3.5}$$

We can set this equal to a unknown constant $c$. This can be done because the total variance explained by the factors stays constant when performing a non singular rotation. This is called the commonality of a variable, and we do not include the unique variance as it is not a part of the rotation problem. In the special case of an orthogonal rotation It can be shown that for an orthogonal matrix $M$ we have the following property when (3.1) holds,

$$\sum_{j=1}^{k} a_{ij}^2 = \sum_{j=1}^{k} b_{ij}^2 = h_i^2. \tag{3.6}$$

The last equality holds by definition of the commonality. Setting equation (3.5) equal to a constant and summing over all rows to include all the loadings we arrive at the following:

$$c = \sum_{i=1}^{p}[(\sum_{j=1}^{k} a_{ij}^2)^2] = \sum_{i=1}^{p}\sum_{j=1}^{k} a_{ij}^4 + 2\sum_{i=1}^{p}\sum_{j<q=1}^{k} (a_{ij}a_{iq})^2. \tag{3.7}$$

Thus if we maximize the sum of the loadings to the fourth power we minimize (3.4). This is where the Quartimax rotation has its name from. To apply the Quartimax method of rotation is to find a solution to the problem of maximizing the following expression,

$$Q(A) = \sum_{i=1}^{p}\sum_{j=1}^{k} a_{ij}^4. \tag{3.8}$$

To maximize $Q(A)$ was first proposed by Ferguson in his paper "The concept of parsimony in factor analysis" [12, . p286]. The Quartimax criteria produces solutions that tend to have one main factor [2, .p 17]. This means that there is one factor that mainly correlates all overlying variables. This one factor is often referred to as the G factor. In other words the correlation between the G factor and all overlying variables tends to be maximized. Intuitively this will lead to one column of the loading matrix with significantly larger entries than all the other.

Quartimax has the advantage that the overlying variables generally become easier to interpret. However when all overlying variables mainly becomes correlated to one variable the factors may become harder to interpret as all overlying variables are primarily associated with one factor.

A method to obtain a solution to the Quartimax rotation is to systematical rotate two columns $p$ and $q$ orthogonaly, where we vary the values $p$ and $q$ over all the columns $\{1, 2, .., k\}$. The orthogonal rotation matrix for the rotation of the columns $p$ and $q$ will be denoted $M_{pq}$ and called a pairwise rotation. The result of the pairwise rotation is denoted $C(k \times k)$ with entries $c_{ij}$. Thus to be clear the first rotation will be $C_0 = \Lambda M_{12}$ and the preceding will be $C_1 = \Lambda M_{12}M_{13}$. It would be correct to separate between all the different $C$ matrices obtained, however this will not be done as it does not clarify how the rotation is done. $M_{pq}$ is found by maximizing the criteria underneath that depends upon the angel of rotation $\theta$ of the two columns,

$$Q_{pq}(\theta) = \sum_{i=1}^{p} (c(\theta)_{ip}^4 + c(\theta)_{jp}^4). \tag{3.9}$$

This is the Quartmax criteria, however it is only applied to the two columns $p$ and $q$ that we are rotating. We can continue this process and rotate the matrix we obtain again and again varying systematically over the columns of the initial matrix $\Lambda$. This will look as following:

$$\begin{aligned} A = &\Lambda M_{12} M_{13} ... M_{1(k-1)} M_{1k} \\ &M_{23} M_{24} ... M_{2(k-1)} M_{2k} \\ &... \\ &M_{(k-2)(k-1)} M_{(k-2)k} \\ &M_{(k-1)k}. \end{aligned}$$

Here we first fix the first column and rotate pairwise with the first and all other columns. This is what the first column in the above expression expresses. When this is completed the second column is fixed and we pairwise rotate all preceding columns. We continue this pattern and in total end up doing $k(k-1)/2$ rotations, this is called a cycle [6, . p285].

The Quartimax criterion is clearly bounded above since we have unit variance of the factors. In particular it holds that $Q(\Lambda) \leq p$. Since the criterion $Q_{pq}(\theta)$ is only two parts of the Quartimax criteria we have that after performing a rotation by $M_{pq}$ that,

$$Q(\Lambda) \leq Q(\Lambda M_{pq}).$$

Thus, in practice, to obtain our rotation matrix $M$ such that $Q(M)$ is maximal (This means that $M$ is the solution to the Quartimax rotation), we repeat cycles until the Quartimax criteria does not change by some tolerance. When performing the partial rotation by a $M_{pq}$ matrix, only two basis vectors will be rotated orthogonally. From equation (3.1) we see that we multiply by the matrix $M_{pq}$ on the right hand side. This result in the columns $p$ and $q$ obtaining the new values as shown below. It is important to stress that this is not the solution. The other columns when performing this rotation will remain unchanged,

$$\begin{aligned} c_{jp} &= \lambda_{ip} cos(\theta) + \lambda_{iq} sin(\theta), \\ c_{jq} &= -\lambda_{ip} sin(\theta) + \lambda_{iq} cos(\theta). \end{aligned} \tag{3.10}$$

Now the problem is reduced to finding a $\theta$ that minimizes $Q_{pq}(\theta)$ (this is equation 3.9) as the values $\lambda$ is known form the initial solution or previous pairwise rotations. This can be solved

analytically for $\theta$ by substituting in (3.10) into $Q_{pq}$ and deriving $Q_{pq}$ and setting this term equal to zero. Solving for $\theta$ we obtain the below expression [6, . p285],

$$\tan(4\theta) = \frac{4\sum_{i=1}^{p}(\lambda_{ip}\lambda_{iq})(\lambda_{ip}^2 - \lambda_{iq}^2)}{\sum_{i=1}^{p}[(\lambda_{ip}^2 - \lambda_{iq}^2)^2 - 4(\lambda_{ip}\lambda_{iq})^2]}. \tag{3.11}$$

It can be shown with a bit of calculations and with the help of the properties $sin(x+\pi/2) = cos(x)$ and $cos(x + \pi/2) = -sin(x)$, that $Q_{pq}(\theta)$ is $pi/2$ periodic [6, . p286]. Therefore we only consider solutions for $\theta \in [0, \pi/2)$. However equation (3.11) which is on the form $tan(4\theta) = d$ where $d \in R$ and $\theta \in [0, \pi/2)$, has two solutions for $\theta$. Since $Q_{pq}(\theta)$ is a continuous function, $\pi/2$ periodic and its derivative is equal to zero two times for any period, it follows that one of them has to be a global maximum and the other has to be a global minimum.

Also observe that $tan(4\theta)$ is $\pi/4$ periodic. Thus, if a soluition $\theta_0$ to (3.11) is obtained and $d^2Q_{pq}/(d\theta^2) \geq 0$ we know that $\theta_0 + \pi/4$ is the solution to the rotation of $M_{pq}$ So to arrive at a value $\theta_0$ that maximizes $Q_{pq}$ we have to solve (3.11) and it can be shown that if the numerator of (3.11) and $sin(4\theta_0)$ has the same sign it follow that $d^2Q_{pq}/(d\theta^2) < 0$ thus $\theta_0$ is our solution. If it is not the case that the numerator of (3.11) and $sin(4\theta_0)$ has the same sign we know that $\theta_0 + \pi/4$ would be the solution.

### 3.1.3 Varimax rotation

An alternative way to arrive at Quartimax rotation is to look at the variance of the sum of squares of the loadings of a column of $A$,

$$M(A) = \sum_{i=1}^{p} Var[\sum_{j=1}^{k} a_{ij}^2] = \frac{1}{k}\sum_{i=1}^{p}\sum_{j=1}^{k} a_{ij}^4 - \frac{1}{k^2}\sum_{i=1}^{p}(\sum_{j=1}^{k} a_{ij}^2)^2. \tag{3.12}$$

We have all ready established that the total commonality of an non singular rotation is unchanged therefore the last term is a constant. Thus maximizing M (3.12) is an equivalent statement of maximizing Q (3.8) and minimizing N (3.4) when an orthogonal rotation is performed. Intuitively maximizing the variance is maximizing the distance of the average to the individual loadings squared. This is all done to each row individually and then summed over all the rows. Since the loadings squared can only range from 0 to 1 specifically when the factor model is applied on a correlation matrix, maximizing the variance would then give the result of forcing the loadings to become as close to 0 and 1 under the restriction of an orthogonal transformation.

The Varimax criterion by Kasier [13, . p] on the other hand comes from maximizing the variance of the sum of squares of the columns of $A$. However to account for the factors having different

commonality we normalize the loadings to their corresponding factors. This gives us the Varimax criterion which is as follows,

$$V(A) = \frac{1}{p}\sum_{j=1}^{k}\sum_{i=1}^{p}\frac{a_{ij}^4}{h_i^4} - \frac{1}{p^2}\sum_{j=1}^{k}(\sum_{i=1}^{p}\frac{a_{ij}^2}{h_i^2})^2 \tag{3.13}$$

$$= \sum_{j=1}^{k} Var[\sum_{i=1}^{p}(\frac{a_{ij}}{h_i})^2]. \tag{3.14}$$

Where $h_i^2$ is the commonality of overlying variable $i \in \{1, 2, .., p\}$ (unique variance is not included). This is reflected upon on in equation (3.6). You may observe that the commonality of the overlying variables are all equal to one if the initial solution is found with the assumption of (1.5).

The Varimax rotation is in general considered as a better orthogonal rotation than Quartimax. This is because it tends to be more stable and it does not tend to correlate all overlying variables mainly to a single factor. Of course whether the Varimax or the Quartimax rotation is best depends on what the researcher is after.

### 3.1.4 Worked example for orthogonal rotations

Here the Quartimax rotation has been applied to the initial solution of the 12 artificial variables dataset. The method for finding the solution to the Quartimax criterion used was the Gradient projection algorithm [9] in contrast to the procedure covered in the sub chapter "Quartimax rotation". This was also used for the next orthogonal rotation, the Varimax rotation. The Gradient projection algorithm will briefly be discussed in the chapter of oblique rotations. The loadings of the rotation can be seen in table (3.1) and the visualization in figure (3.1).

We can see that the G factor in this example becomes factor two. However factor two explains quite a large proportion of the variance, with a value of $1.162$ compared to $1.705$ of factor two. This might be because there in not a clear grouping of the variables, which results in the Quartimax not being able to produce a rotation where one factor dominates.

We rotate the initial solution of the 12 overlying variables according to the Varimax criterion. This produces the factor loadings displayed in table (3.2) and figure (3.2). We can see that the Varimax rotation does arrive at a different solution than the Quartimax rotation. Unexpectedly it does produce a solution where one factor dominates with respect to explained variance of the overlying variables.

| | Factor one | Factor two |
|---|---|---|
| V1 | 0.681 | 0.444 |
| V2 | 0.727 | 0.496 |
| V3 | 0.277 | 0.330 |
| V4 | 0.240 | 0.289 |
| V5 | 0.048 | 0.161 |
| V6 | 0.020 | 0.092 |
| V7 | 0.069 | 0.197 |
| V8 | 0.017 | 0.064 |
| V9 | -0.070 | 0.758 |
| V10 | -0.128 | 0.550 |
| V11 | -0.058 | 0.227 |
| V12 | -0.058 | 0.251 |
| SST | 1.162 | 1.705 |

**Table 3.1:** Quartimax rotation of the initial solution of 12 Artificial variables

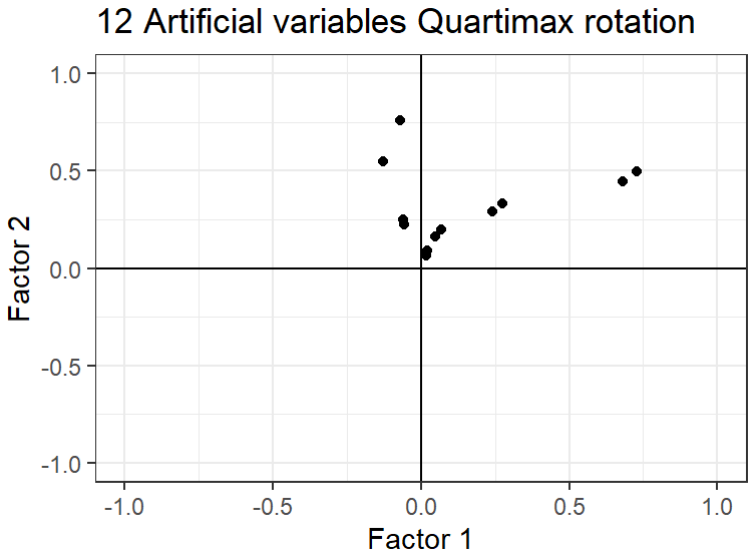



**Figure 3.1:** Plot of the factor loadings from table 2.1

| | Factor one | Factor two |
|---|---|---|
| V1 | 0.805 | 0.112 |
| V2 | 0.869 | 0.140 |
| V3 | 0.391 | 0.181 |
| V4 | 0.340 | 0.159 |
| V5 | 0.112 | 0.125 |
| V6 | 0.057 | 0.075 |
| V7 | 0.147 | 0.149 |
| V8 | 0.043 | 0.051 |
| V9 | 0.259 | 0.716 |
| V10 | 0.118 | 0.553 |
| V11 | 0.044 | 0.230 |
| V12 | 0.054 | 0.252 |
| SST | 1.796 | 1.071 |

**Table 3.2:** Varimax rotation of the initial solution of 12 Artificial variables

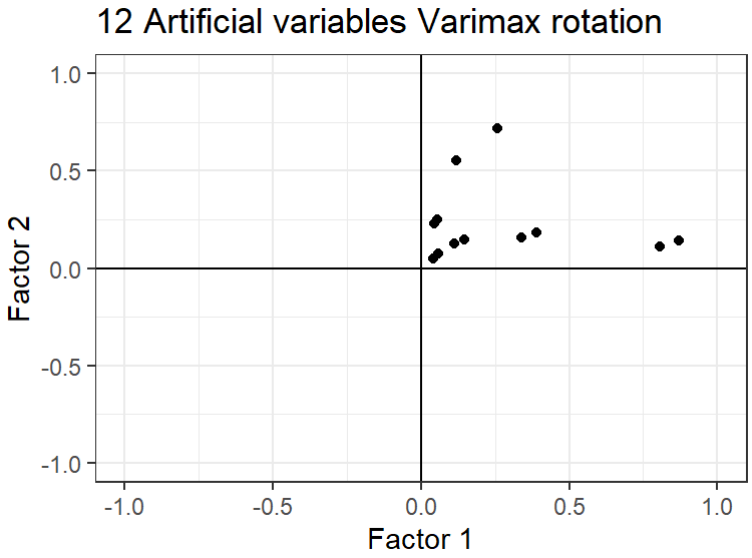



**Figure 3.2:** Plot of the factor loadings from table 2.2

# Chapter 4

## 4.1 Oblique rotations

So far we have only reviewed older popular orthogonal rotations and we have shown how to arrive at the rotation matrix $M$ from maximizing the complexity function of the Quartimax rotation. This turned out to be a long and quite computer intensive calculation that only finds a solution to one type of rotation. This motivates finding an algorithm that can be applied to any complexity function and find an approximate solution. First we will briefly cover how the complexity functions work.

A complexity function of an oblique rotation $f(\Lambda)$, dependent on $\lambda(p \times k)$ which is the factor pattern matrix, is a function that we seek to minimize to arrive at a solution. This is done with an non singular matrix $B(k \times k)$ with column sum equal to one,

$$\Lambda = \Lambda_0 B^{-1}. \tag{4.1}$$

As previously shown in (3.7) we have that $MM^\intercal = I$. This imposes $q(q-1)/2$ restraints. However when performing an oblique rotation we allow the factors to be correlated. We right hand multiply with the inverse of $B$ such that the column of B has to have unit length. This is shown by (4.2), the restriction of an orthogonal rotation. The factors still are assumed to have unit variance. Then in an oblique rotation the following has to hold,

$$\text{Diag}(BB^\intercal) = I. \tag{4.2}$$

Here Diag$(X)$ is a function that sets all of diagonal elements equal to one of $X$. This only imposes m constraint. Observe that the constraints of the orthogonal rotations produces a set that is a sub group of all oblique rotations. Therefor, the oblique rotations have to better or equally optimize the complexity functions.

To arrive at a solution to any complexity function the use of the gradient projection algorithm will be used [9]. This is an algorithm that finds the local minima by an iterative calculation of the gradient and moving with the gradient, until convergence is acquired. Since the solution to the complexity function is a global minimum, and the algorithm finds a local minima, we consider multiple different initial factor matrices as starting points. What is considered sufficient calculations of the algorithm are then performed such that it is highly likely that the global minimum is found. This is what has been used for the computations with the `R`-package `GPArotation` [9] that produce the graphs.

### 4.1.1 Yates' Geomin

It is generally accepted that a factor loading matrix with a fair number of vanishing entries is simpler than a factor loading matrix with mostly intermediate values [11, . p180]. Looking back at the Quartimax and Varimax rotations, we observed that both rotations try to simplify either the columns or the rows of the factor pattern matrix. In for example Quartimax, this was done by looking at the sum of squares of the rows.

We also developed a geometrical interpretation of how the Quartimax rotation tried to force every factor loading to become zero by rotating the initial solution such that each point would come as close to each basis vector as possible. This was done by looking at the product of two factor loadings squared. We will take this one step further, consider the following equation where $X(1 \times k)$ an arbitrary vector and $x_i$ is the $i$'th element of $X$,

$$c(X) = \prod_{i=1}^{k} x_i. \tag{4.3}$$

Our aim is to apply $c(X)$ to a complexity function [14, . p120]. However we first want to modify it such that it will to its best find a solution that satisfies the simple structure criteria. If $c(X)$ was applied to each row of a factor pattern matrix and used as a complexity function, we can see that it would become zero when at least one of the elements of $X$ is zero. So $c(X)$ does have the benefit that its best solution satisfies the first criteria for the simple structure criteria, that each row of the factor pattern matrix should have at least one zero.

However the function $c(X)$ does have problems with indeterminacy as if one element is zero it does not matter what the others are. They can be as far away from zero and one as they want, and this is not what we are looking after. Thus a modification to the function $c(X)$ may be made such that it can be useful in a complexity function. geomin suffers from the problem of indeterminacy. We will therefor focus on an altered version of geomin named Yates' Geomin by Yates [15]. geomin is given as below [14, . p120],

$$f(\Lambda) = \sum_{i=1}^{p} [\prod_{i=1}^{k} \lambda_{ij}^2]^{1/k}.$$

This complexity function has entries as squared which solves the problem of having negative entries in the complexity function. However this proposed complexity function still suffers from indeterminacy. There are multiple ways of solving this problem, as for example finding the solutions to this complexity function and analytically choosing which solution best follows the simple structure. However to be able to arrive at a solution the complexity function was slightly modified such that it is not possible to have an entry to (4.3) that is zero. This was done by adding a small positive constant. This small positive constant does not greatly affect the output of $f(\Lambda)$ however it solves the problem with indeterminacy [14, .p 121]. So the complexity function of Yates' Geomin thus is defined as (4.4) where $\epsilon$ is a small value,

$$f(\Lambda) = \sum_{i=1}^{p}[\prod_{i=1}^{k}(\lambda_{ij}^2 + \epsilon)]^{1/k}. \tag{4.4}$$

Often an $\epsilon = 0.01$ is used for three and four factors [14, .p 121]. When there are more than four factors $\epsilon$ might need to be increased.

For the oblique rotations we fit the initial solution of Harman, H. H 8 Physical variables. A solution for Yates' Geomin rotation was found by fitting with 20 initial sets of initial values, randomly generated. All sets of initial values lead to the same solution up to the sign of the factor pattern matrix. Therefore I choose the solution which resulted in positive large loadings rather than negative large loadings. The result is displayed in figure (4.1) and table (4.1). Table (4.2) is also displayed and it shows the correlation between factor one and factor two.

| | Factor one | Factor two |
|---|---|---|
| Height | 0.891 | 0.059 |
| Arm span | 0.952 | -0.020 |
| Length of forearm | 0.927 | -0.043 |
| Lenght of lower leg | 0.876 | 0.037 |
| Weight | 0.023 | 0.921 |
| Bitrochanteric diameter | -0.009 | 0.817 |
| Chest girth | -0.045 | 0.761 |
| Chest with | 0.093 | 0.680 |

**Table 4.1:** Geomin rotation of the initial solution of Harman H. H 8 Physical variables

| | Factor one | Factor two |
|---|---|---|
| Factor one | 1 | 0.462 |
| Factor two | 0.462 | 1 |

**Table 4.2:** Factor correlation matrix of the rotated factors according to the Geomin criteria of the initial solution of Harman H. H 8 Physical variables

The Geomin rotation produces a solution where the observed variables are mainly correlated to only one factor. This is considered as a good solution because it is easier to interpret. This is

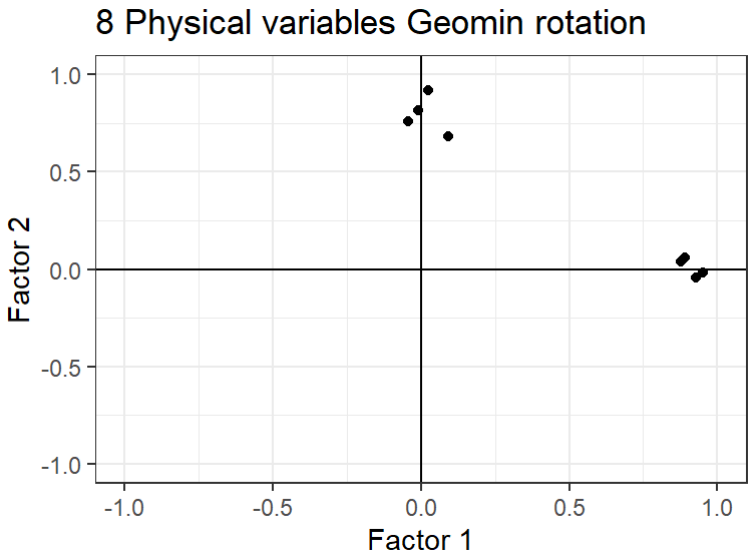



**Figure 4.1:** Plot of factor loadings from table 4.1

only possible because we allow the factors to be correlated. It also has to be mentioned that it was quite simple to obtain a solution, which is in contrast to the next rotation criteria that will be discussed.

### 4.1.2 Simplimax rotation and the component loss criteria

This section will be dedicated to exploring and explaining the component loss rotation criteria introduced by Robert I. Jennrich [11]. The component loss criteria takes the form of a quite simple criteria. However even though the criterion is simple it has some interesting proven properties and can even outperform some of the more complex rotations.

**Definition 4.1.1** (Component loss criteria)**.** Let $\Lambda(p \times k)$ be the target rotation matrix and $\lambda_{ij}$ be its entries (4.1), then a complexity function that takes the form:

$$Q(\Lambda) = \sum_{i=1}^{p} \sum_{j=1}^{k} h(|\lambda_{ij}|). \tag{4.5}$$

is called a component loss criteria. Further more the function $h(|_{ij}|)$ is its defining component loss function. We will be referring to the component loss criteria as CSC and referring to the component loss function as CLF from this point out.

For the result of the CLC's it will be assumed that the CLF's are concave functions of $\lambda$. An intuitive way of looking at concavity is that it is to encourage the factor loadings to be as small as possible. However concavity will not always be assumed, because the component loss function

in interest is not concave on its whole domain. Note that adding a constant or scaling the CLF does not afflict the solution of the component loss criteria. It is therefore often assumed without loss of generality that $h(0) = 0$. This also makes the component loss criteria fall under the categorization as a complexity function. The first theorem concerns the solution of a CLC when a perfect simple structure exists. This theorem will guarantee that when there exists a perfect simple structure of the data, it will be the solution of the component loss criteria when the CLF is concave. To prove this we will need the two following lemmas.

**Lemma 4.1.2.** *If $\Lambda$ and $\tilde{\Lambda}$ are rotations, where the initial solution is given by $\Lambda_0$, and further assume that $\tilde{\Lambda}$ has perfect simple structure where $\tilde{\lambda}_{ir}$ is the non zero element of column $i$. We have that:*

$$|\tilde{\lambda}_{ir}| \leq |\lambda_{i1}| + |\lambda_{i2}| + ... + |\lambda_{ik}|. \tag{4.6}$$

*Proof.* Since $\Lambda$ is a solution of a rotation, we have that, $\Lambda = \Lambda_0 B^{-1}$, $\Lambda B = \Lambda_0$. If $b_i$ is the $i$'th row of the rotation matrix $B$ and $a_i$ is the $i$'th row of $\Lambda_0$, we can express each row of $\Lambda_0$ as the following,

$$a_i = \lambda_{i1} b_1 + \lambda_{i2} b_2 + ... + \lambda_{ik} b_k.$$

As previously established the rows of $B$ have unit length, therefore

$$||a_i|| \leq \lambda_{i1} + \lambda_{i2} + ... + \lambda_{ik}.$$

We will do the same calculation for the case when $\tilde{\Lambda}$ is a perfect simple structure solution of the initial rotation $\Lambda_0$. Then, $\tilde{\Lambda} = \Lambda_0 \tilde{B}^{-1}$, $\tilde{\Lambda}\tilde{B} = \Lambda_0$, and $\tilde{\Lambda}_{ir}$ is the only non zero element of column $i$ of $\tilde{\Lambda}$. This gives the equation,

$$a_i = \tilde{\Lambda}_{ir} \tilde{b}_r.$$

Hence $||a_i|| = |\tilde{\Lambda}_{ir}|$, which gives us the inequality the lemma states. □

**Lemma 4.1.3.** *Assume that $h(\lambda)$ is concave and $u_1, u_2, ..., u_k$ are non negative, then*

$$h(\sum_{i=1}^{k} u_i) \leq \sum_{i=1}^{k} h(u_i). \tag{4.7}$$

*This inequality is strict if $h(\lambda)$ is strictly concave and two or more $u_r$'s are nonzero.*

*Proof.* The inequality clearly holds if all $u_r$ equal zero. Assume at least one $u_r$ is not equal to zero. We can then express $u_r$ as a concave combination of $0$ and $\sum_{i=1}^{k} u_i$ as follows,

$$u_r = (1 - \frac{u_r}{\sum_{i=1}^{k} u_i})0 + (\frac{u_r}{\sum_{i=1}^{k} u_i}) \sum_{i=1}^{k} u_i.$$

Now using the definition of concavity we obtain,

$$h((1-\frac{u_r}{\sum_{i=1}^{k} u_i})0+(\frac{u_r}{\sum_{i=1}^{k} u_i}) \sum_{i=1}^{k} u_i) \geq (1-\frac{u_r}{\sum_{i=1}^{k} u_i})h(0)+(\frac{u_r}{\sum_{i=1}^{k} u_i})h(\sum_{i=1}^{k} u_i) = (\frac{u_r}{\sum_{i=1}^{k} u_i})h(\sum_{i=1}^{k} u_i).$$

Here the left side of the equation equals $h(u_r)$ as the previous equation shows. Now if we sum over all $u_r$ for $r \in \{1, 2, ..., k\}$, we obtain the result of the lemma. To show that the inequality is strict when $h(\lambda)$ is strictly concave and at least two or more $u_r$'s are nonzero, observe that for some $r$, $0 < \frac{u_r}{\sum_{i=1}^{k} u_i} < 1$ which gives a strict inequality. Thus the lemma holds. □

Note that the $u_1, u_2, ..., u_k$ are supposed to represent the elements of a row in a solution of a rotation made by a concave rotation.

**Theorem 4.1.4.** *If there exists a solution $\tilde{\Lambda}$ with perfect simple structure of an oblique rotation using a component loss rotation criteria where the corresponding component loss function $h$ is convex and non decreasing, we have that $\tilde{\Lambda}$ minimizes the component loss criteria. Further more, if $h$ is strictly concave the global minimum must be a solution of only factor matrices with perfect simple structure.*

*Proof.* Assume without loss of generality that $h(0) = 0$. This can be done as adding a constant to the CLF does not alter the solution of the component loss criteria. Now looking at the a row $i$ of the perfect simple structure solution we obtain that,

$$\sum_{j=1}^{k} h(|\tilde{\lambda_{ij}}|) = h(|\tilde{\lambda_{ir}}|).$$

Assume $\Lambda$ is any rotation of $\Lambda_0$. Applying Lemma (4.2.1) and and the monotonicity of $h(\lambda)$

$$h(|\tilde{\lambda_{ir}}|) \leq h(\sum_{j=1}^{k} |\lambda_{ij}|)$$

We can apply lemma (4.2.2) since $h(\lambda)$ is convex and $h(\lambda) = 0$

$$h(\sum_{j=1}^{k} |\lambda_{ij}|) \leq \sum_{j=1}^{k} h(|\lambda_{ij}|) \tag{4.8}$$

Using the first equality and the last two inequalities we obtain

$$\sum_{j=1}^{k} h(|\tilde{\lambda_{ij}}|) \leq \sum_{j=1}^{k} h(|\lambda_{ij}|)$$

If we now sum over all rows we obtain what the theorem states. To prove the last statement of the theorem assume that $h(\lambda)$ is strictly convex and $\Lambda$ does not have perfect simple structure. It therefore follows that for at least one row $r$, of $\Lambda$, that two or more entries are nonzero. Thus by Lemma (4.2.2) we obtain a strict inequality of equation (4.8) for the $r$'th row. Thus $\Lambda$ cannot minimize the component loss function. □

This theorem states that when the component loss function is strictly concave and there exists a perfect simple structure of the data, the component loss criteria must perform a rotation that results in perfect simple structure. There are two simple examples of such functions, the linear CLF, $h_1(\lambda)$ and an exponential CLF named monomolecular growth curve, $h_2(\lambda)$. They are given by,

$$h_1(\lambda) = \lambda, \qquad h_2(\lambda) = 1 - e^{\lambda}.$$

Both the exponential and the linear CLF will produce simple structure solutions if they exist. However only the exponential function will have solutions that only consist of perfect simple structure matrices if they exist, because the linear function is not strictly concave. This is obviously a very useful property however it is extremely unlikely to happen in practice.

We will now direct our focus toward the complexity function Simplimax by Kiers [16] and its similarities with the quadratic right constant, which is a component loss function. First we will prove an interesting result that regards the quadratic right constant when a Thurstone-type simple structure exists. If a factor loading matrix has one or more exact zeroes in each column it has Thurstone-type simple structure.

**Theorem 4.1.5.** *Let $\tilde{\Lambda}$ have Thurstone-type simple structure and $c > 0$ be less than the smallest nonzero absolute loading. $h(u)$ is constant for all $c \leq u$ and $h(0) \leq h(u)$ for all $u \geq 0$, then $\tilde{\Lambda}$ is a local minima of the the CLC defined by $h(u)$.*

*Proof.* We have that $|\tilde{\lambda}| > 0$ if and only if $|\tilde{\lambda}| > c$. Since this is a strict inequality we have a neighborhood of $\tilde{\Lambda}$ where this inequality holds. Let $\Lambda$ be an element of this neighborhood. Then as stated $|\Lambda| > 0$ i.f.f $|\Lambda| > c$ and the following holds,

$$\sum_{i=1}^{p}\sum_{j=1}^{k} h(|\tilde{\lambda}|) = \sum_{|\tilde{\lambda}|>c}\sum h(|\tilde{\lambda}|) = \sum_{|\lambda|>c}\sum h(|\lambda|) \leq \sum_{i=1}^{p}\sum_{j=1}^{k} h(|\lambda|).$$

□

The quadratic right constant will be denoted $H_a(\Lambda)$ and is defined by its quite simple component loss function $h(u)$,

$$h_a(\lambda) = \begin{cases} (\frac{\lambda}{a})^2 & \text{if } |\lambda| < a \\ 1 & \text{else} \end{cases}$$

where $a$ is an arbitrary positive real number smaller than one. Note that $h(\lambda)$ is not concave because of the added constant part. Therefore if there exist a perfect simple structure it may not minimize the quadratic right constant. A similar complexity function is the popular oblique rotation, Simplimax. The simplimax rotation is defined by,

$$S_r(\Lambda) = \sum_{i=1}^{p}\sum_{j=1}^{k} I[\lambda_{ij}^2 \leq \lambda_r^2]\lambda_{ij}^2.$$

where $I[x]$ is the indicator function and $\lambda_r$ is defined to be the $r$'th smallest loading that is chosen beforehand. This complexity function might be called an iterative component loss criteria. This is because $S_r$ is minimized with fixed $\lambda_r$ until the $r$'th smallest factor loadings changes. Thereafter the new $r$'th smallest loading of the resultant factor pattern matrix of the rotation is found, and all larger factor loadings than the $r$'th are set to zero. This is then repeated until convergence is acquired. Keep in mind however that Simplimax is highly dependent on having random initial rotation matrices as it is a complicated complexity function. Therefore the gradient projection algorithm might not obtain the factor loading matrix which which minimizes $S_r$. This is where the quadratic right constant dominates, because it is a simple loss criteria.

The theorem bellow tells us that Simplimax produces the same result as the CLC defined by the quadratic right constant if $a$ and $\lambda_r^2$ is chosen appropriately.

**Theorem 4.1.6.** *Let $\lambda_r^2 < a < \lambda_{r+1}^2$. Then a loading matrix $\Lambda$ is a local minima of $S_r$ if and only if $\Lambda$ is a local minima of $H_a$.*

*Proof.* We can obtain an equality of $S_r$ and $H_a$ in a neighborhood of $\Lambda$ since $a$ is strictly chosen to be between $\lambda_r^2$ and $\lambda_{r+1}^2$. Let $\tilde{\Lambda}$ be an element of this neighborhood. Then, in that neighborhood,

$$S_r(\tilde{\Lambda}) = a^2 H_a(\tilde{\Lambda}) - a^2(k \times p - m).$$

Here we subtract $(k \times p - m)$ because we end up with $(k \times p - m)$ ones from the definition of $H_a$, and the Simplimax criteria gives us zeroes instead. Since we have only scaled and added a constant to the component loss criteria $H_a(\Lambda)$ it does not affect its local minima. □

This theorem does not imply that the component loss criteria defined by the quadratic right constant and Simplimax are equivalent complexity functions because we cannot know which factor loadings are the $r$'th smallest at the last step of the gradient projection algorithm. The two complexity functions simply have different input parameters. Say that there exists a solution of Thurstone-type simple structure that we would like to rotate to using the Simplimax criteria. We can in theory find the corresponding $r$'th smallest loading such that such that the CLC defined by the quadratic right constant's parameter $a$ satisfies $\lambda_r^2 < a < \lambda_{r+1}^2$ where $\lambda_r^2$ is of the final step in the projection algorithm. However this would not guarantee a solution of Thurstone-type simple structure since theorem 2 is week in the sense that it only guarantees a local minima. It is however unlikely as Thunrstone-type simple structure rarely occurs in practice [11, p. 180].

To perform the rotation according to the Simplimax rotation we will also perform gradient decent. However the Simplimax has multiple local minima, such that it is not sufficient to find one local minima and rotate accordingly. Therefor to seek the global minima of the Simplimax complexity function multiple minimizations of the complexity function from random initial starts will be performed to prevent this.

To choose which value of $r$ to be used we will plot a scree plot for each value of $r$ from 1 to 16. The scree plot plots $r$ against the complexity function of Simplimax evaluated in its found global minima with parameter $r$. Kiers (1998) [16, . p572] suggested to choose the value $r$ where the

value after has the first considerably large increase. This way we find the value for $r$ such that the complexity function manages to find as many approximate zeroes as possible. If we look at figure (a), (b) and (c), we can see that figure (a) displays four approximate zeroes, figure (b) displays eight and when $r$ becomes to large the criteria has to consider to many factor loadings such that the resultant rotation does not have any approximate zeroes. The best fit of the Simplimax rotation with parameter $r = 8$ can be seen in table (4.3) and figure (b), with the factor correlation given in table (4.4).

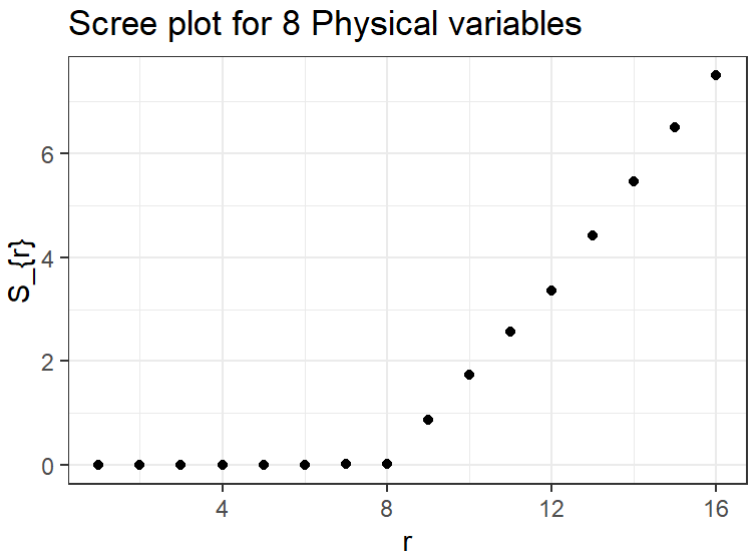
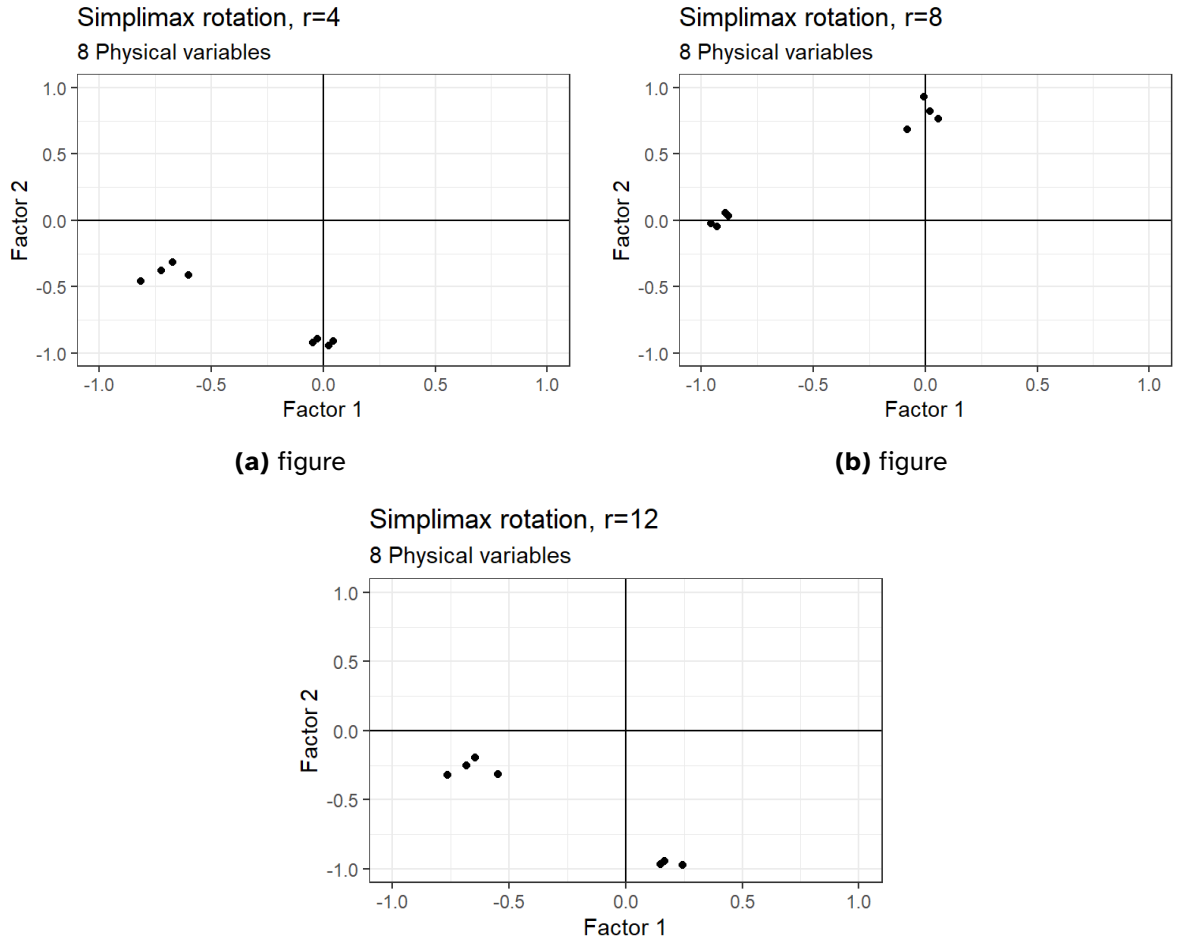



**Figure 4.2:** Scree plot of the Simplimax criteria fited to the initialsolution of Harman H. H 8 Physical variables

| | Factor one | Factor two |
|---|---|---|
| Height | -0.892 | 0.054 |
| Arm span | -0.955 | -0.026 |
| Length of forearm | -0.931 | -0.049 |
| Lenght of lower leg | -0.877 | 0.032 |
| Weight | -0.007 | 0.928 |
| Bitrochanteric diameter | 0.023 | 0.824 |
| Chest girth | 0.058 | 0.768 |
| Chest with | -0.081 | 0.685 |

**Table 4.3:** Simplimax rotation of initial solution of Harman, H. H 8 Physical variables with parameter r=8




**(a)** figure

**(b)** figure

**(c)** figure

**Figure 4.3:** Three different fits of the Simplimax criteria with parameter r = 4,8,12 of the Harman H. H 8 Physical variables

| | Factor one | Factor two |
|---|---|---|
| Factor one | 1 | -0.480 |
| Factor two | -0.480 | 1 |

**Table 4.4:** Factor correlation matrix of Simplimax rotation of initial solution of Harman, H. H 8 Physical variables with parameter r=8

# Chapter 5

## 5.1 Discussion

It might at first seem that the oblique rotations are superior to the orthogonal rotations. This however depends upon what the researcher is after. More often than not the factor model is fitted to try to explain data that is observed. It may be that the researcher knows in advance that the factors that are desired should be correlated and therefore an oblique rotation criteria is chosen. In other situations it might be in interest to find a solution where the factors represent uncorrelated variables. One particular situation that might arise when orthogonal rotation is desired is when the researcher wants to perform dimension reduction. Where as principal component analysis creates a new variable which is a weighted sum of the variables of the dataset, factor analysis assumes that there are underlying variables that explains the observed variables of the dataset. This means that the variables created by principal component analysis are explained by the variables of the dataset. However factor analysis creates underlying variables which explains the variables of the data set. The fundamental difference is of which variables are causing which variables. This assumption of there being underlying variables might be a bad assumption as can be reflected by high values of unique variance and the hypothesis test created for factor analysis. However when it produces a good fit it has the possibility to be simpler that its counter part, principal component analysis [17].

Another interesting part of the theory of factor analysis concerns the uniqueness of a solution. Before the analytical rotations were developed multiple different manual rotations were developed. This however was not favored since it was possible for the researcher to force the factor loadings to fit his own perception of what the factors may represent. This is where the analytical rotation criteria come into play. However the more modern rotation criteria, for example the Simplimax rotation, has a parameter that can be varied which often can result in different solutions of the rotation. Thus it has been recommended by other authors that more attention has to be paid to this problem to be able to use rotation criterion's for theory-construction.

# Bibliography


[1] R. B. Ivimey-Cook and M. C. F. Proctor, 'Factor analysis of data from an east devon heath: A comparison of principal component and rotated solutions,' *Journal of Ecology*, vol. 55, no. 2, pp. 405–413, 1967, ISSN: 00220477, 13652745. [Online]. Available: http://www.jstor.org/stable/2257885 (visited on 14/12/2022).

[2] K. M. Rennie, 'Exploratory and confirmatory rotation strategies in exploratory factor analysis,' *Non-Journal*, pp. 1–28, 1997. [Online]. Available: https://eric.ed.gov/?id=ED406446.

[3] K. M. Kieffer, 'Orthogonal versus oblique factor rotations: A review of the literature regarding the pros and cons,' *Non-Journal*, pp. 1–32, 1998. [Online]. Available: https://eric.ed.gov/?id=ED427031.

[4] K. V. Mardia, J. T. Kent and J. M. Bibby, *Multivariate analysis* (Probability and Mathematical Statistics: A Series of Monographs and Textbooks). Academic Press [Harcourt Brace Jovanovich, Publishers], London-New York-Toronto, Ont., 1979, pp. 3–521, ISBN: 0-12-471250-9.

[5] R. A. Darton, 'Rotation in factor analysis,' *Journal of the Royal Statistical Society*, vol. 29, no. 3, pp. 167–194, 1980.

[6] H. H. Harman, *Modern factor analysis*, revised. University of Chicago Press, Chicago, Ill.-London, 1976, pp. 1–487.

[7] W. Revelle, *Psych: Procedures for psychological, psychometric, and personality research*, R package version 2.2.9, Northwestern University, Evanston, Illinois, 2022. [Online]. Available: https://CRAN.R-project.org/package=psych.

[8] J. M. L. John Schmid, 'The development of hierarchical factor solutions,' *Psychometrika*, vol. 22, no. 1, pp. 53–61, 1957. [Online]. Available: https://link.springer.com/article/10.1007/BF02289209.

[9] C. A. Bernaards and R. I.Jennrich, 'Gradient projection algorithms and software for arbitrary rotation criteria in factor analysis,' *Educational and Psychological Measurement*, vol. 65, pp. 676–696, 2005.

[10] L. L. Thurstone, *Multiple-Factor Analysis. A Development and Expansion of "The Vectors of Mind."* University of Chicago Press, Chicago, Ill., 1947, pp. 1–535.

[11] R. I. Jennrich, 'Rotation to simple loadings using component loss functions: The oblique case,' *Psychometrika*, vol. 71, no. 1, pp. 173–191, 2006, ISSN: 0033-3123. DOI: 10.1007/s11336-003-1136-B. [Online]. Available: https://doi.org/10.1007/s11336-003-1136-B.

[12] G. A. Ferguson, 'The concept of parsimony in factor analysis,' *Psychometrika*, vol. 19, no. 4, pp. 281–290, 1954, ISSN: 1860-0980. DOI: 10.1007/BF02289228. [Online]. Available: https://doi.org/10.1007/BF02289228.

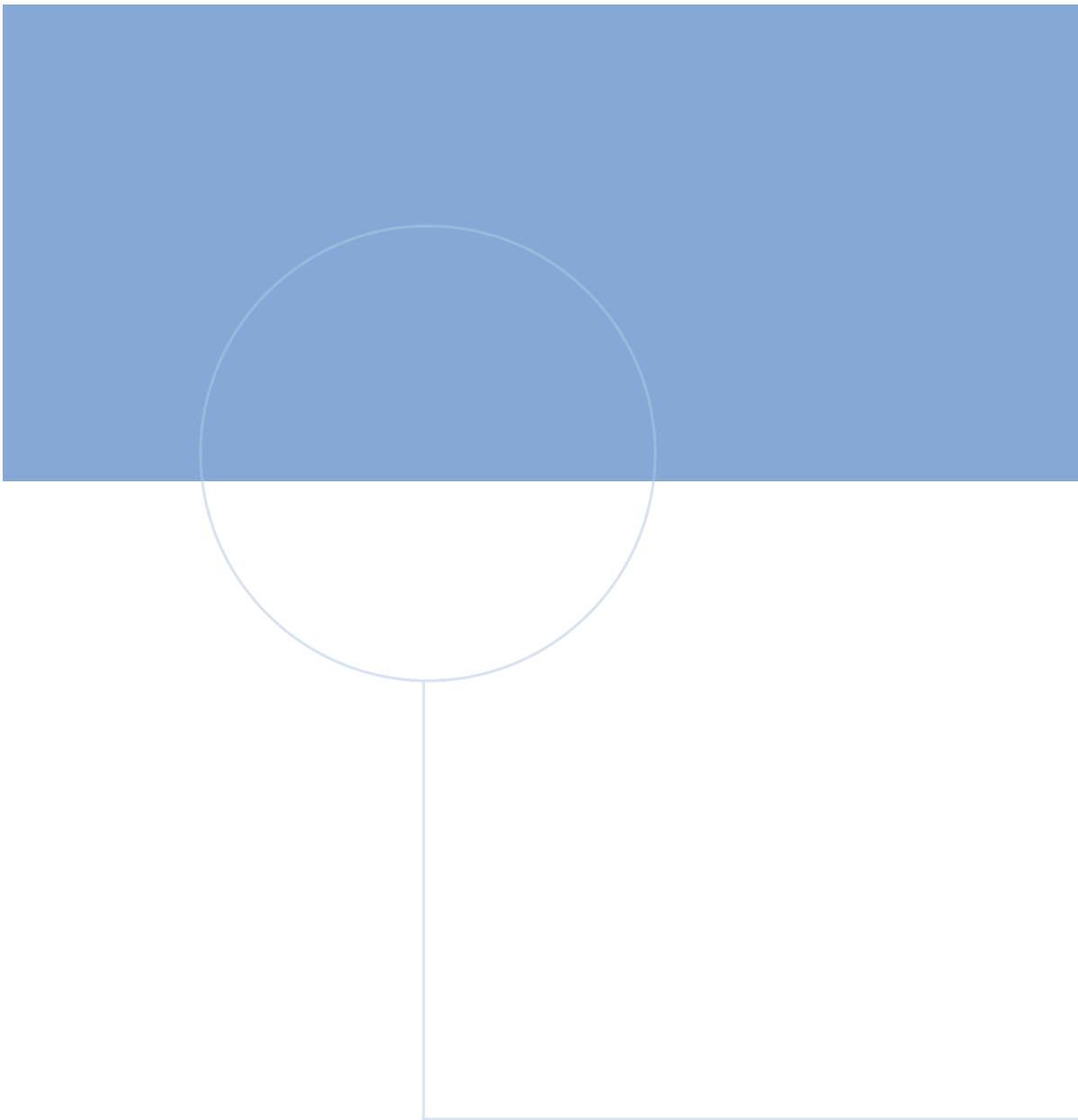
[13] H. F. Kaiser, 'The varimax criterion for analytic rotation in factor analysis,' *Psychometrika*, vol. 23, no. 3, pp. 187–200, 1958, ISSN: 1860-0980. DOI: 10.1007/BF02289233. [Online]. Available: https://doi.org/10.1007/BF02289233.

[14] M. W. Browne, 'An overview of analytic rotation in exploratory factor analysis,' *Multivariate Behavioral Research*, vol. 36, no. 1, pp. 111–150, 2001. DOI: 10.1207/S15327906MBR3601_05. [Online]. Available: https://doi.org/10.1207/S15327906MBR3601_05.

[15] A. Yates, 'Multivariate exploratory data analysis: A perspective on exploratory factor analysis,' *State University of New York Press, Albany*, 1987.

[16] H. A. L. KIERS, 'Simplimax: Oblique rotation to an optimal target with simple structure,' *Psychometrika*, vol. 59, no. 4, pp. 567–579, 1994. DOI: 10.1007/BF02294392. [Online]. Available: https://doi.org/10.1007/BF02294392.

[17] K. P. F.R.S., 'Liii. on lines and planes of closest fit to systems of points in space,' *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 2, no. 11, pp. 559–572, 1901. DOI: 10.1080/14786440109462720. [Online]. Available: https://doi.org/10.1080/14786440109462720.



NTNU

Norwegian University of
Science and Technology