

# **Generating Non-normal Distributions**

## **Methods and Effects**

**Max Auerswald**  
**Dipl.-Psych.**

Inaugural dissertation  
submitted in partial fulfillment of the requirements for the degree  
Doctor of Social Sciences in the Graduate School of Economics and  
Social Sciences at the University of Mannheim.

Dean of the School of Social Sciences: Prof. Dr. Michael Diehl

Academic Director of the CDSS: Prof. Dr. Edgar Erdfelder

Thesis Advisors: Prof. Dr. Edgar Erdfelder

Prof. Dr. Morten Moshagen

Thesis Reviewers: Prof. Dr. Thomas Gschwend

Prof. Dr. Andreas Klein

Date of Oral Defense:

26.10.2017

# Contents

<b>Abstract</b>	<b>1</b>
<b>Manuscripts</b>	<b>2</b>
<b>1 Introduction &amp; Theoretical Background</b>	<b>3</b>
1.1 Assessing Non-normality . . . . .	5
1.1.1 Univariate Measures . . . . .	5
1.1.2 Multivariate Measures . . . . .	11
1.1.3 Tests of normality . . . . .	14
1.2 Effects of Non-Normality . . . . .	15
1.2.1 Effects on the General Linear Model . . . . .	16
1.2.2 Effects on SEM . . . . .	18
1.3 Methods that Generate Non-Normal Data . . . . .	19
1.3.1 Power Constants and its Extensions . . . . .	20
1.3.2 NORTA . . . . .	22
1.3.3 Copulas . . . . .	24
<b>2 Summary of Manuscripts</b>	<b>28</b>
2.1 A method for multivariate non-normality . . . . .	28
2.2 Combining moment- and distribution-based methods . . . . .	33
2.3 Non-normality and exploratory factor analysis . . . . .	37
<b>3 General Discussion &amp; Outlook</b>	<b>42</b>
3.1 Limitations . . . . .	44
3.2 Future Research Questions . . . . .	46
<b>4 Conclusion</b>	<b>48</b>

References	49
Co-Author Statement	58
Statement of Originality	59
Appendix: Manuscripts	60

# Abstract

Many inferential statistical tests require that the observed variables have a normal distribution. Monte Carlo simulations are used to investigate the effects of non-normality by repeatedly applying these tests to samples from a non-normal distribution, for which the correct inference is known. A prerequisite of Monte Carlo studies is an algorithm that generates such samples, thereby controlling three parameters: (1) the correlation among random variables, (2) the marginal distributions, and (3) the multivariate distribution. Most previously used algorithms only allow control over the correlations and the marginals, but recent results show that the robustness of certain methods depends on the multivariate distribution as well.

In my thesis, I suggest a new method to generate samples from non-normal distributions that allows manipulations of all three parameters simultaneously. In the first manuscript, I develop an algorithm that jointly controls the correlation matrix, one central moment of the marginals, and the multivariate distribution. Additionally, I also show that the multivariate distribution has a distinct impact on the robustness of a structural equation model. In the second manuscript, the algorithm is extended to allow control over multiple central moments of the marginals. The third manuscript applies the algorithm to extraction criteria for exploratory factor analysis. Parallel analysis, the extraction criterion with the highest accuracy, was unaffected by the underlying distribution.

Overall, my thesis provides Monte Carlo studies with a powerful tool to reevaluate the robustness of various statistical tests under conditions of non-normality, especially when the assumption of normality pertains to a multivariate distribution. By considering a wider range of plausible data conditions, empirical research can profit from a more accurate assessment of the validity of statistical tests.

# Manuscripts

This thesis is based on three manuscripts which have been published or have been submitted for publication in peer-reviewed journals. The manuscripts are listed below and appended to this thesis in the order in which they will be discussed.

1. Auerswald, M., & Moshagen, M. (2015). Generating correlated, non-normally distributed data using a non-linear structural model. *Psychometrika*, 80, 920-937.
2. Auerswald, M., & Moshagen, M. (2017). *Sampling from arbitrary non-normal distributions with given covariance and central moments*. Manuscript submitted for publication.
3. Auerswald, M., & Moshagen, M. (2017). *How to determine the number of factors to retain in exploratory factor analysis? A comparison of extraction methods under realistic conditions*. Manuscript submitted for publication.

# Introduction & Theoretical Background

One might wonder if even one psychological data set existed, that allowed to test research hypotheses while fulfilling all assumptions underlying the statistical test. Depending on the test, these assumptions include, for example, that missing data are not systematically missing, that the criterion can be expressed as a linear combination of predictors, or that the residuals of the model are independent (e.g. Gelman & Hill, 2007; Tabachnick & Fidell, 2012). One prominent assumption that is commonly violated in empirical data sets is multivariate normality (Blanca, Arnau, López-Montiel, Bono, & Bendayan, 2013; Cain, Zhang, & Yuan, in press; Micceri, 1989). This violation is typically ignored in empirical analyses (Keselman et al., 1998) because most statistical methods are considered to be fairly robust against this violation (e.g. Gelman & Hill, 2007), which means that inferences are still more or less correct even though the assumption of multivariate normality is violated.

The process of determining the robustness of a statistical method with regards to non-normality involves the generation of random samples in which this assumption is violated and the correct inference is known. For example, in a two-sample  $t$ -test, a robustness study could consist of generating multiple random samples from two non-normal distributions with the same population mean  $\mu$  and repeatedly applying  $t$ -tests that compare the sample means to each other. The two-sample  $t$ -test would then be considered as robust if, for example, the empirical Type I error rate (the proportion of significant results in the simulation) is comparable to the nominal  $\alpha$  error. The result of such a Monte Carlo simulation study obviously depends on the choice of distributions. Distributions that are more similar to a normal distribution yield empirical Type I error rates closer to the theoretical  $\alpha$  (Harwell, Rubinstein,

Hayes, & Olds, 1992). Since real data samples show a large variety of distributions (Cain et al., in press), robustness studies should provide results for a wide range of distributions, allowing practitioners to assess if the validity of their analysis is in peril for a given data set and statistical method.

The main goal of this thesis is to provide a more flexible algorithm, called NOTAMO (Normal To Arbitrary MOments), that generates non-normally distributed random variables for Monte Carlo robustness studies. Most previously used algorithms only allowed the manipulation of marginal distributions, either directly (Cario & Nelson, 1997) or by specifying the univariate skewness and kurtosis of the distributions (Vale & Maurelli, 1983). In contrast, NOTAMO allows for the generation of different multivariate distributions with the same marginals, thereby creating data conditions that would be treated as equivalent in other robustness studies. Importantly, the results of robustness studies depend on variations of the multivariate distribution, thus limiting the extent to which other simulation results can be generalized to real data sets.

The introductory chapter is organized as follows: First, I give an overview of measures that describe distributions, such as skewness and kurtosis. These measures are often utilized in robustness studies to set up guidelines, i.e., that a specific method is unaffected by non-normality as long as e.g. the kurtosis is within a certain range. Second, I summarize results on the robustness of the general linear model and structural equation models regarding non-normality. I will end the chapter by outlining three methods that generate non-normal multivariate data. The second chapter gives summaries of the articles this thesis is based on, including a discussion of each article in relation to the central goal of the thesis. The concluding third chapter presents a general discussion and an outlook to future research questions related to robustness studies and multivariate normality.



## 1.1 Assessing Non-normality

Univariate continuous distributions are usually expressed by their probability density function (*PDF*)  $f(x)$ , where

$$\Pr[a \leq X \leq b] = \int_a^b f(x)dx, \quad (1.1)$$

for a random variable  $X$ . Thus, the *PDF* is used to obtain the probability that  $X$  falls into a given interval  $[a, b]$ . Similarly, the cumulative density distribution (*CDF*)  $F(x)$  expresses the random variable  $X$  as

$$\Pr[X \leq x] = F(x), \quad (1.2)$$

and obtains the probability that  $X$  is smaller or equal to a given value  $x$ . For multivariate continuous distributions, the concept of a *CDF* can be extended to the joint cumulative distribution function

$$\Pr[X_1 \leq x_1, \dots, X_d \leq x_d] = F(x_1, \dots, x_d), \quad (1.3)$$

for  $d$  random variables  $X_1, \dots, X_d$  and gives the probability that each  $X_1, \dots, X_d$  is smaller or equal to  $x_1, \dots, x_d$ . While simulation studies can use *PDFs* and *CDFs* to define a random variable, the underlying distribution of a random variable in an observed sample is unknown and needs to be estimated. Instead, empirical samples are typically described by their mean, (co)variance, skewness, and kurtosis (Blanca et al., 2013). In this chapter, I give an overview of measures and tests used to assess the distribution of an empirical sample. These measures are necessary for simulation studies, as they provide guidelines for which distributions a statistical method is robust and therefore connect simulation studies with empirical practice.

### 1.1.1 Univariate Measures

Univariate measures assess the characteristics of marginal distributions, which is especially useful in cases where the assumption of normality is made for single random variables such as the errors in a linear regression (e.g. Gelman & Hill, 2007).

Skewness and kurtosis are the most commonly used indicators of non-normality, and both are standardized central moments of the distribution.

### Univariate Skewness

Population skewness is defined as

$$\gamma_1 = \text{E} \left[ \left( \frac{X - \mu}{\sigma} \right)^3 \right], \quad (1.4)$$

for a random variable  $X$  with mean  $\mu$ , standard deviation  $\sigma$ , and  $\text{E}$  is the expected value. Skewness is generally used as an indicator of asymmetry that can take values from negative to positive infinity. Symmetric distributions (such as the normal distribution) have skewness  $\gamma_1 = 0$  and non-zero values indicate that the distribution is asymmetric. In particular, a positive skewness results if the distribution 'leans' to the left, has longer right tails, and/or a higher density in the right tail, whereas negative skewness is associated with a right-leaning distribution and longer or fatter left tails. For example, reaction time data often have positive skewness, as responses cannot fall below a threshold due to response times of the motor system and very slow responses tend to occur less often (Palmer, Horowitz, Torralba, & Wolfe, 2011). Accuracy data of simple cognitive tasks tend to be negatively skewed, because most participants respond to most tasks correctly (Wang, Zhang, McArdle, & Salthouse, 2008). Figure 1.1 displays *PDFs* of (standardized) generalized normal distributions (Log-Normal3, Asquith, 2017) with shape parameters  $\kappa = 0.71$ ,  $\kappa = 0.44$ , and  $\kappa = 0$ , resulting in skewness  $\gamma_1 = 3$ ,  $\gamma_1 = 1.5$ , and a standard normal distribution with  $\gamma_1 = 0$ . Skewness is not always easy to interpret because it depends on both characteristics of the tails and center of the distribution. It is a common misconception to state that a skewness of  $\gamma_1 = 0$  implies that a distribution is symmetric (e.g., Blanca et al., 2013). Distributions can be left-leaning and have a longer left tail, resulting in skewness  $\gamma_1 = 0$  and an asymmetric distribution (see e.g., Meijer, 2000).

Sample skewness is usually estimated by Fisher's  $G_1$  estimate, defined as

$$G_1 = \frac{\sqrt{N(N-1)}}{N-2} \frac{m_3}{m_2^{3/2}}, \quad (1.5)$$

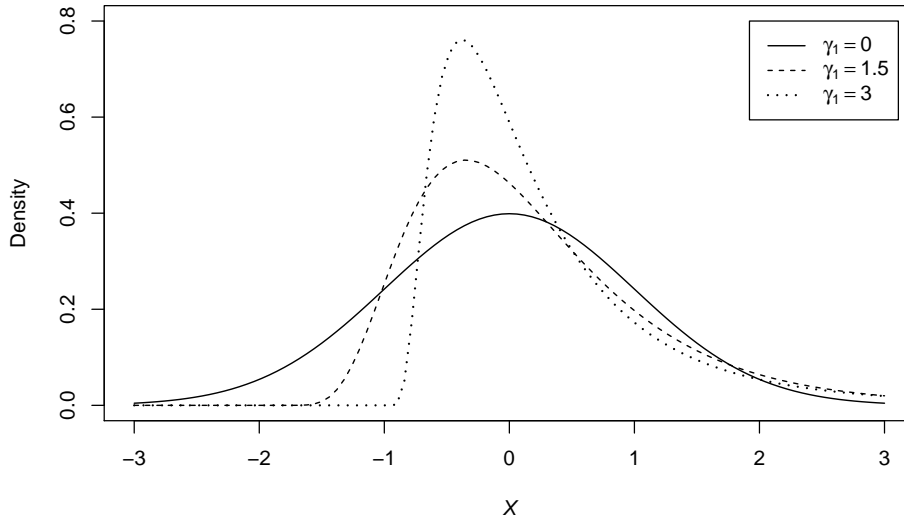


Figure 1.1: Probability density functions of (standardized) generalized normal distributions (Log-Normal3) with shape parameters  $\kappa = 0$ ,  $\kappa = 0.44$ , and  $\kappa = 0.71$ . The resulting distributions have skewness  $\gamma_1 = 0$ ,  $\gamma_1 = 1.5$ , and  $\gamma_1 = 3$ , respectively.

where

$$m_r = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^r. \quad (1.6)$$

In general,  $G_1$  is a consistent but not unbiased estimate of  $\gamma_1$  and can deviate from the population skewness considerably even in samples with  $N = 100$  (Joanes & Gill, 1998).

Cain et al. (in press) investigated the empirical skewness of 1,567 variables from 194 psychological studies and found that 66% exhibited skewnesses significantly different from 0, which increased to 82% in studies with larger sample sizes ( $N > 106$ ). The range of observed skewnesses was  $[-10.87, 25.54]$  with  $-1.17$  and  $2.77$  as the 5th and 95th percentile, indicating that the absolute skewness is typically smaller than three. For studies with very small sample sizes ( $N \leq 30$ ), Blanca et al. (2013) reported skewness estimates from 693 studies in the range  $[-2.49, 2.33]$ , which is considerably less extreme but potentially underestimates the population skewness because  $G_1$  was used as an estimator (Joanes & Gill, 1998). Overall, skewness is a property of distributions commonly encountered in samples of observed random variables.

## Univariate Kurtosis

Kurtosis is defined as

$$\gamma_2 = \text{E} \left[ \left( \frac{X - \mu}{\sigma} \right)^4 \right], \quad (1.7)$$

and can vary between 1 and positive infinity, where normal distributions have  $\gamma_2 = 3$ .<sup>1</sup> It is typically interpreted as the probability density of the peak and tails of the distribution, relative to the shoulders (at  $\mu \pm \sigma$ ). Values of kurtosis  $\gamma_2 < 3$  are associated with platykurtic distributions that are less peaked, have flatter tails, and higher density in the shoulders, whereas distributions with  $\gamma_2 > 3$  are leptokurtic and have a higher peak and longer/fatter tails on both sides (DeCarlo, 1997). Figure 1.2 shows the PDFs of platykurtic (standardized) exponential power distributions (Nadarajah, 2005) with kurtosis  $\gamma_2 = 1.85$ ,  $\gamma_2 = 2.2$ , and a standard normal distribution ( $\gamma_2 = 3$ ) on the left. The right panel of Figure 1.2 displays leptokurtic Student  $t$  distributions with  $df = 4.1$ ,  $df = 6$ , and  $df = \infty$  resulting in kurtosis  $\gamma_2 = 60$ ,  $\gamma_2 = 6$ , and a standard normal distribution. Observed variables with extreme outliers are often leptokurtic, such as reaction times (Palmer et al., 2011) or income (Cain et al., in press). A typical example for a platykurtic distribution is age, as there is in general a comparable number of individuals across the age range (Cain et al., in press).

Sample kurtosis is often estimated by Fisher's  $G_2$  estimate,

$$G_2 = \frac{N - 1}{(N - 2)(N - 3)} \left[ (N + 1) \left( \frac{m_4}{m_2} - 3 \right) + 6 \right] + 3, \quad (1.8)$$

with  $m_2$  and  $m_4$  as in Equation 1.6. In general,  $G_2$  is consistent but not unbiased and tends to underestimate the kurtosis in smaller samples ( $N \leq 100$ ), especially if the population kurtosis is large (Joanes & Gill, 1998).

In empirical samples, Cain et al. (in press) reported a range from 1.80 to 1,096.48 for kurtosis with 1.72 and 12.48 as the 5th and 95th percentiles, indicating that few observed variables exhibit extreme values. However, the kurtosis of a majority of distributions again deviated from the normal distribution (54%). In very small sam-

<sup>1</sup>Excess kurtosis is defined as  $\gamma_{2,ex} = \gamma_2 - 3$  (so that normal distributions have  $\gamma_{2,ex} = 0$ ) and is sometimes used as an alternative definition of kurtosis. To avoid confusion, I only use kurtosis as defined in Equation 1.7.

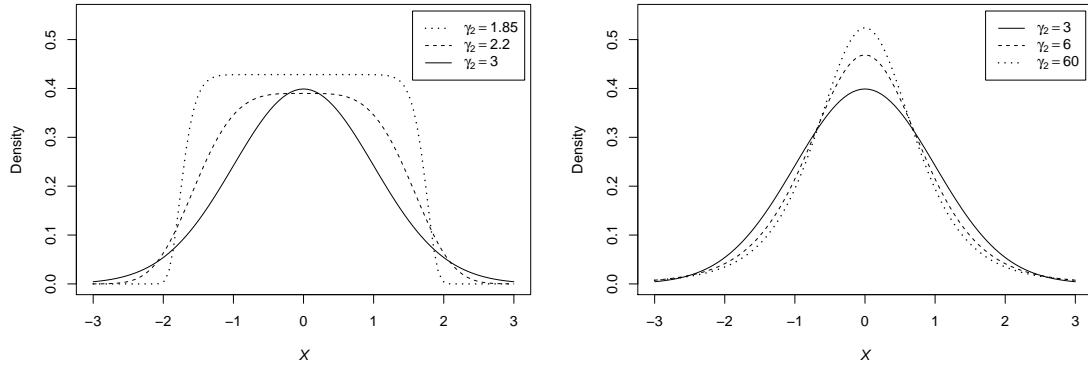


Figure 1.2: Probability density functions of distributions with varying kurtosis. The left panel displays (standardized) exponential power distributions with shape parameters  $\beta = 13.45$ ,  $\beta = 3.93$ , and  $\beta = 2$ , resulting in distributions with kurtosis  $\gamma_2 = 1.85$ ,  $\gamma_2 = 2.2$ , and a standard normal distribution ( $\gamma_2 = 3$ ). The right panel displays (standardized)  $t$  distributions with  $df = \infty$ ,  $df = 6$ , and  $df = 4.1$ . The corresponding distributions have kurtosis  $\gamma_2 = 3$ ,  $\gamma_2 = 6$ , and  $\gamma_2 = 60$ , respectively.

ples ( $N \leq 30$ ), the range of observed kurtosis values was  $[1.08, 10.41]$  and therefore more narrow (Blanca et al., 2013), but this could be due to  $G_2$ 's bias to underestimate the population kurtosis. In sum, most observed variables display kurtosises different from  $\gamma_2 = 3$  and both leptokurtic as well as platykurtic distributions are not uncommon.

## Central Moments

Standardized central moments can be understood as the generalization of skewness and kurtosis. The  $k$ th central moment of a distribution is defined as

$$\mu_k = \text{E}[(X - \mu)^k], \quad (1.9)$$

where  $\text{E}$  is again the expected value and  $\mu$  is the population mean. The central moment can be standardized to obtain

$$\tilde{\mu}_k = \frac{\mu_k}{\sigma^k}, \quad (1.10)$$

with standard deviation  $\sigma$ . If  $k \in \{3, 4\}$ , this is equivalent to Equation 1.4 for skewness and Equation 1.7 for kurtosis, respectively. The standardized moments

with  $k > 4$  can be used to describe a distribution further<sup>2</sup> but they are virtually never used in empirical practice, likely as a result of being difficult to interpret. However, central moments draw attention to the fact that univariate distributions can differ despite equal mean, variance, skewness, and kurtosis.

Figure 1.3 displays two random variables with skewness  $\gamma_1 = 0$  and kurtosis  $\gamma_2 = 3$ . The left panel shows the PDF

$$f_{mix}(x) = \frac{1}{2}\Gamma(x, 2.30, 0.36) + \frac{1}{2}\Gamma(-x, 2.30, 0.36), \quad (1.11)$$

which is a mixture distribution where  $\Gamma(x, k, \theta)$  is a gamma distribution with shape parameter  $k = 2.30$  and scale parameter  $\theta = 0.36$ . The shape and scale parameter were chosen to obtain  $\sigma = 1$  and  $\gamma_2 = 3$ . The distribution is symmetric, so  $\gamma_1 = 0$ . The right panel displays a discrete probability mass function with three unique values,  $m_1$ ,  $m_2$ , and  $m_3$ . These values with corresponding probabilities  $p_1, p_2, p_3$  were chosen to satisfy

$$\begin{aligned} p_1 m_1 + p_2 m_2 + p_3 m_3 &= 0 \\ p_1 m_1^2 + p_2 m_2^2 + p_3 m_3^2 &= 1 \\ p_1 m_1^3 + p_2 m_2^3 + p_3 m_3^3 &= 0 \\ p_1 m_1^4 + p_2 m_2^4 + p_3 m_3^4 &= 3 \\ p_1 + p_2 + p_3 &= 1, \end{aligned} \quad (1.12)$$

which guarantees the desired skewness and kurtosis. In the solution displayed in Figure 1.3,  $m_1 = -3.43, m_2 = -0.64, m_3 = 1.28$  with  $p_1 = .014, p_2 = .632, p_3 = .354$ , respectively. Both random variables are indistinguishable from a standard normal distribution based on the first four moments but differ regarding moments of higher order. The distributions are clearly not normal, thereby illustrating the shortcomings of relying on a few moments to characterize a distribution appropriately.

---

<sup>2</sup>However, even an infinite sequence of all moments is in general insufficient to define a unique distribution, which is known as the problem of moments (e.g. Joe, 1997).

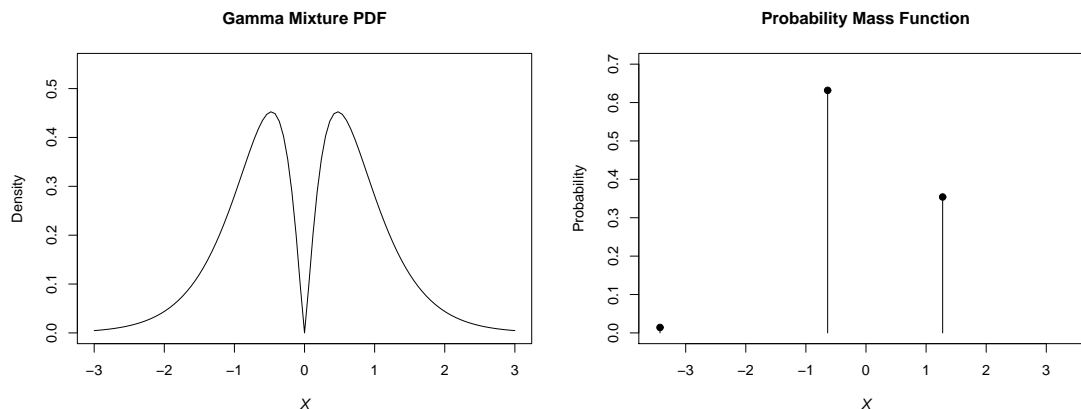


Figure 1.3: Two random variables with skewness  $\gamma_1 = 0$  and kurtosis  $\gamma_2 = 3$ , equal to the skewness and kurtosis of a normal distribution. The left panel displays the probability density function of a standardized mixture of gamma distributions with shape parameter  $k = 2.30$ . The right panel shows the probability mass function of a discrete distribution with values  $m_1 = -3.43$ ,  $m_2 = -0.64$ , and  $m_3 = 1.28$ .

### 1.1.2 Multivariate Measures

Multivariate measures are used to assess characteristics of the multivariate distribution, which is necessary if the assumption of normality applies to the joint distribution of observed variables as in structural equation models (SEM, Bollen, 1989). Univariate measures are also regularly (and mistakenly) used to investigate multivariate normality, despite the fact that a multivariate distribution can be non-normal while exhibiting normal marginals (Dutta & Genton, 2014). Figure 1.4 shows a bivariate distribution with normal marginals, in which the density of quadrant II and IV is redistributed to quadrant I and III, according to the example of Dutta and Genton (2014). If one would only check the marginals of distributions as in Figure 1.4, the distribution would appear perfectly normal despite the obvious deviation from multivariate normality. Consequently, distributional aspects of marginal distributions may fall short to allow for conclusions regarding the underlying multivariate distribution. Instead, measures that attempt to capture properties of the multivariate distribution itself are required.

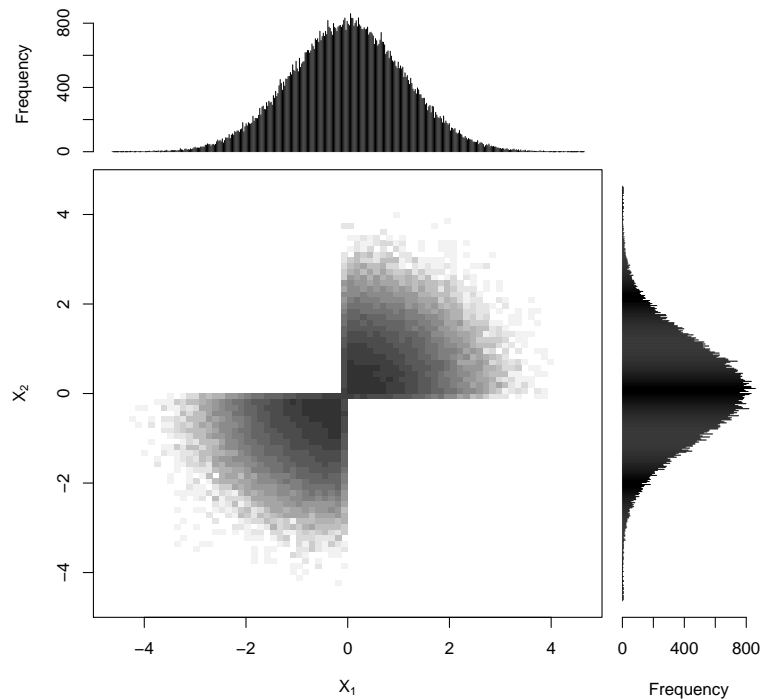


Figure 1.4: Heatmap of a bivariate sample ( $N = 100,000$ ) of a non-normal distribution with normal marginals, according to Dutta and Genton (2014).

### Multivariate Skewness and Kurtosis

Multivariate skewness and kurtosis are the multivariate extension of their respective univariate counterpart (Mardia, 1970). They assess similar characteristics as univariate skewness and kurtosis, but are based on the joint distribution and take the covariance between random variables into account. Let  $\mathbf{X} = (X_1, \dots, X_d)$  be a  $d \times 1$  vector of  $d$  random variables with biased sample covariance matrix  $S$  defined as

$$S = \frac{1}{N} \sum_{i=1}^N (\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})'. \quad (1.13)$$

Then, multivariate sample skewness is defined as

$$b_{1,d} = \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N [(\mathbf{X}_i - \bar{\mathbf{X}})' S^{-1} (\mathbf{X}_j - \bar{\mathbf{X}})]^3. \quad (1.14)$$

Multivariate normal distributions have a multivariate skewness of  $b_{1,d} = 0$  and higher values indicate a stronger deviation from normality. In empirical samples, Cain et



al. (in press) reported values in the range from 0 to 1,263 with a median of 3.08 and mean 32.94. A majority of data sets demonstrated multivariate skewness that significantly differed from zero (58%).

Multivariate sample kurtosis is defined as

$$b_{2,d} = \frac{1}{N} \sum_{i=1}^N [(\mathbf{X}_i - \bar{\mathbf{X}})' S^{-1} (\mathbf{X}_i - \bar{\mathbf{X}})]^2, \quad (1.15)$$

where  $S$  and  $\mathbf{X}$  are as above. Normal distributions have a multivariate kurtosis of  $d(d+2)$ , smaller values indicate a platykurtic distribution, and larger values indicate a leptokurtic distribution. Cain et al. (in press) observed empirical multivariate kurtosis in the range from 1.99 to 1,476 with median 18.90 and mean 78.70. Again, 57% of kurtosis values were significantly different from the corresponding value of a normal distribution. Overall, either skewness or kurtosis deviated in 68% of all cases and in 94% of cases with larger sample sizes ( $N > 106$ ), indicating that only a small portion of empirical data sets is normally distributed.

### Tail Dependence

Tail dependence is a measure of dependence for bivariate distributions (Joe, 1997). For a pair of random variables  $X_1, X_2$ , upper tail dependence is based on the conditional probability that  $X_1$  exceeds its quantile  $q$ , given that  $X_2$  is larger than its own quantile  $q$ . More specifically, upper tail dependence  $td_u$  is the limit of this probability if  $q \rightarrow 1$ , so

$$td_u = \lim_{q \rightarrow 1} P(X_1 > F_1^{-1}[q] \mid X_2 > F_2^{-1}[q]), \quad (1.16)$$

where  $F_1^{-1}, F_2^{-1}$  are the inverse *CDFs* of  $X_1, X_2$ , respectively. Similarly, lower tail dependence  $td_l$  is the limit of the conditional probability that  $X_1$  is smaller than quantile  $q$ , given that  $X_2$  is below  $q$ , for  $q \rightarrow 0$ :

$$td_l = \lim_{q \rightarrow 0} P(X_1 < F_1^{-1}[q] \mid X_2 < F_2^{-1}[q]). \quad (1.17)$$

Tail dependence is a measure for random variables that cannot be applied to a sample of a distribution, because it is defined on the limit of quantiles. In contrast

to multivariate skewness and kurtosis, which are based on the entire density of a distribution, tail dependence is only influenced by the most extreme outcomes. Nevertheless, tail dependence plays an important role in economic models because market prices are better modeled by distributions with tail dependence as extreme prices for one good tend to result in extreme prices for another good (e.g. Hartmann, Straetmans, & De Vries, 2004). Gaussian distributions, the distributions assumed by most statistical tests, always have zero tail dependence unless they are perfectly correlated (Joe, 1997). As I will summarize in Section 1.2.2, distributions with non-zero tail dependence seem to have a stronger impact on the robustness of SEM (Foldnes & Grønneberg, 2015).

### 1.1.3 Tests of normality

The assumption of a normal distribution can also be assessed by statistical tests. For a univariate distribution, the Kolmogorov–Smirnov test (Kolmogorov, 1933; Smirnov, 1948) compares the empirical *CDF* to the *CDF* of a completely specified reference distribution, such as a normal distribution with mean  $\mu$  and variance  $\sigma^2$ . The associated test statistic is based on the largest difference between the *CDFs* and defined as

$$D_n = \sup_x |F(x) - F_n(x)|, \quad (1.18)$$

where  $F$  is the *CDF* of the reference distribution,  $F_n$  is the *CDF* of the observed random variable, and  $\sup$  the supremum. Importantly, the test statistic is only valid if the reference distribution does not contain parameters, such as mean and variance, that are estimated from the sample (Lilliefors, 1967). The Shapiro–Wilk test (Shapiro & Wilk, 1965) can be used to test the hypothesis that the sample was drawn from any (univariate) normal distribution, i.e. when the mean and variance of the distribution are unspecified. The test is useful when the assumption of normality is made for a single random variable, such as for the errors in a linear regression or the observed variables in each group of a *t*-test, but falls short when the assumption pertains to a multivariate distribution.

Multivariate normality is often tested with Mardia’s tests for normality (Mardia, 1970), which are based on the multivariate skewness and kurtosis described in the

previous section (for an overview of other tests of multivariate normality, see e.g. Henze, 2002). The test statistic  $b_{1,d}^*$  for multivariate skewness  $b_{1,d}$ , as defined in Equation 1.14, is

$$b_{1,d}^* = \frac{N}{6} b_{1,d}, \quad (1.19)$$

and follows a  $\chi^2$  distribution with  $df = d(d+1)(d+2)/6$ . The respective test statistic  $b_{2,d}^*$  for multivariate kurtosis  $b_{2,d}$  (Equation 1.15) is

$$b_{2,d}^* = \sqrt{N} \frac{b_{2,d}(N+1) - d(d+2)(N-1)}{(N+1)\sqrt{8d(d+2)}} \quad (1.20)$$

and follows a standard normal distribution. One limitation of tests for normality that are based on  $b_{1,d}$  and  $b_{2,d}$  is that a distribution can have zero multivariate skewness and  $d(d+2)$  multivariate kurtosis but still have a non-normal distribution (Horswell & Looney, 1992). Therefore, a non-significant result in Mardia's tests does not imply that the distribution is in fact normal.

All statistical tests of normality share the disadvantage that the power of these tests to detect deviations from a normal distribution is greater for larger samples (e.g. Razali & Wah, 2011), whereas the effect of non-normality is usually greater if the sample size is small (e.g. Harwell et al., 1992). Especially in smaller samples, a non-significant deviation from a normal distribution might therefore not indicate that the inferential method is robust, which is why non-normality is often assessed by the measures presented in this chapter and graphical examination (Tabachnick & Fidell, 2012).

## 1.2 Effects of Non-Normality

The assumption of multivariate normality applies to different methods in different ways. In linear regression models, the residuals of the analysis are required to be normally distributed, whereas other popular methods that are based on the empirical covariance matrix (e.g. structural equation models) incorporate the assumption that the observed variables themselves have a multivariate normal distribution (Bollen, 1989; Gelman & Hill, 2007). In this section, I give an overview of the consequences of non-normality for linear regressions, ANOVAs,  $t$ -tests, and SEM.

### 1.2.1 Effects on the General Linear Model

ANOVAs and  $t$ -tests are based on the assumption that distributions in each group are normal (Loveland, 2011), which can be investigated by varying the corresponding univariate distribution. A number of Monte Carlo studies have investigated the effect of non-normality (Bradley, 1973; Cain et al., in press; Clinch & Keselman, 1982; Glass, Peckham, & Sanders, 1972; Harwell, 2003; Harwell et al., 1992; Levine & Dunlap, 1982; Schmider, Ziegler, Danay, Beyer, & Bühner, 2010; Yanagihara & Yuan, 2005). For  $t$ -tests, the most relevant factors that influence robustness are the number of observations, the skewness of observed variables, and whether the test is one-tailed or two-tailed. For example, Cain et al. (in press) showed that the empirical Type I error rate  $p_{emp}$  of a one-sample, two-tailed  $t$ -test increases to  $p_{emp} = .177$  for  $\gamma_1 = 6.32$  and  $N = 18$  (for a nominal  $\alpha = .05$ ). If the test is one-tailed and tests the 'shorter' tail of the distribution (in this case the lower tail), the error increases even further ( $p_{emp} = .216$ ), whereas a test for the 'longer' tail results in artificially small Type I errors ( $p_{emp} = .005$ ). Larger sample sizes lead to higher robustness ( $p_{emp} = .123$ ,  $N = 48$ , two-tailed), but Type I errors are still substantial even at large sample sizes if the distribution is very skewed ( $p_{emp} = .090$ ,  $N = 105$ ,  $\gamma_1 = 6.32$ , two-tailed). As expected, distributions with lower skewness also lead to smaller Type I errors ( $p_{emp} = .064$ ,  $N = 105$ ,  $\gamma_1 = 2.77$ , two-tailed). Furthermore, robustness of  $t$ -tests is lower if a smaller nominal  $\alpha$  is chosen, at least relative to the nominal  $\alpha$  (Bradley, 1973). That is, the ratio of  $p_{emp}$  and  $\alpha$  can be very large even for high  $N$  (e.g.  $p_{emp} = .006$  for  $\alpha = .001$ ,  $\gamma_1 = 3.18$ , and  $N = 1024$ ; Bradley, 1973). Most results for  $t$ -tests similarly hold for planned comparisons in single factor ANOVAs, which are also based on the  $t$  distribution (Yanagihara & Yuan, 2005)

In general, ANOVAs display higher robustness than  $t$ -tests for non-normality with regards to the  $\alpha$  error, at least if normality is the only assumption that is violated (Glass et al., 1972; Cain et al., in press). Clinch and Keselman (1982) showed that in a single factor ANOVA with four groups, Type I error rates were slightly decreased for a  $\chi^2_{(2)}$  distribution if the sample size and variances were equal across groups ( $p_{emp} = .038$ , overall  $N = 48$ ,  $\alpha = .05$ ). However, non-normality has small to moderate effects when combined with unequal variances, unequal group sizes, or

both. In particular, if both group size and variance were unequal, Type I error rates increased to  $p_{emp} = .224$ , compared to  $p_{emp} = .207$  for a normal distribution. Importantly, alternatives to the ANOVA  $F$  test that do not assume equal variances like the Welch test (Welch, 1951) or the procedure by Brown and Forsythe (1974) were affected by non-normality when combined with unequal variances and group sizes (Welch:  $p_{emp, \chi^2_{(2)}} = .127$ , compared to  $p_{emp, normal} = .064$ , Brown:  $p_{emp, \chi^2_{(2)}} = .103$ , compared to  $p_{emp, normal} = .072$ ). In addition to the effect of non-normality on Type I errors, Levine and Dunlap (1982) demonstrated that the effect on Type II errors is even more pronounced. For example, in a condition with four groups and overall  $N = 64$ , the Type II error was  $\beta = .329$  for log-normal distributions, compared to the expected  $\beta = .081$  for normal distributions. Again, larger sample sizes lead to higher robustness, but the effect was still substantial for  $N = 128$  ( $\beta_{log-normal} = .285$ , compared to  $\beta_{normal} = .074$  for normal).

Linear regressions have the assumption that the error terms in the model are normally distributed (e.g. Gelman & Hill, 2007). Since ANOVAs and  $t$ -tests are only special cases of a linear regression, the results presented so far apply to linear regressions, too. However, in multilevel regression analyses, the random parts of the model are also assumed to be multivariate normal (e.g. Hox, 2010). In contrast to the corresponding assumption in simple linear regressions, this assumption pertains to a multivariate distribution, which could have effects on the robustness beyond the prespecified marginals. Additionally, the normality assumption could be especially problematic for multilevel models because the number of observations at higher levels is typically smaller than sample sizes in simple linear regressions (Browne & Draper, 2000). Maas and Hox (2004) investigated the effects of  $\chi^2_{(1)}$  distributed random parts ( $\gamma_1 = 2.83, \gamma_2 = 15$ ), as compared to normal random parts, for different group sizes (5, 30, 50) and number of groups (30, 50, or 100). Full maximum likelihood (ML) estimation underestimated the standard errors of both fixed and random effects. For example, the true value of the random slope variance was only covered in 64% of all simulated samples by the 95% confidence interval. Maas and Hox (2004) also employed robust Huber/White standard errors (Huber, 1967; White, 1982), which performed considerably better but still underestimated the correct standard errors (e.g. 95% confidence intervals covered the true random slope variance in 85% of all

samples). However, the simulation study by Maas and Hox (2004) did not include conditions that varied the multivariate distribution independently of the prespecified marginals.

### 1.2.2 Effects on SEM

In contrast to most models discussed so far, the broad class of SEM with ML, weighted least squares, or generalized least squares estimation requires that the observed variables have a multivariate normal distribution (Bollen, 1989; Browne, 1974). Curran, West, and Finch (1996) investigated the effect of non-normality by comparing marginal distributions with  $\gamma_1 = 2$ ,  $\gamma_2 = 10$  (moderately non-normal), as well as  $\gamma_1 = 3$ ,  $\gamma_2 = 24$  (severely non-normal) to a multivariate normal distribution using ML estimation. Both non-normality conditions severely increased the ratio at which a correctly specified confirmatory factor model was rejected (moderately non-normal: rejection rate  $rr = 23.5\%$ , severely non-normal:  $rr = 38.5\%$ , normal:  $rr = 5.6\%$ ). Importantly, this effect did not diminish in large sample sizes of  $N = 1,000$  (moderately non-normal:  $rr_{N=1,000} = 24\%$ , normal:  $rr_{N=1,000} = 7\%$ ) and was even more pronounced in the severely non-normal condition ( $rr_{N=1,000} = 48\%$ ). Curran et al. (1996) also examined the asymptotic distribution free estimator (Browne, 1984), which does not assume any particular distribution, as well as the Satorra-Bentler  $\chi^2$  (SB, Satorra & Bentler, 1994), which corrects for the observed multivariate kurtosis, as alternatives to ML. The asymptotic distribution free estimator did not perform well in conditions with  $N \leq 500$  (moderately non-normal:  $rr = 21.6\%$ , severely non-normal:  $rr = 33.8\%$ , normal:  $rr = 24.3\%$ ), indicating that larger samples are required for this estimator. SB improved rejection rates when data were non-normal (moderately non-normal:  $rr_{N \geq 200} = 7.5\%$ , severely non-normal:  $rr_{N \geq 200} = 7.8\%$ , normal:  $rr_{N \geq 200} = 7\%$ ), but was still biased in smaller samples (moderately non-normal:  $rr_{N=100} = 8.5\%$ , severely non-normal:  $rr_{N=100} = 13\%$ , normal:  $rr_{N=100} = 7.5\%$ ). Foldnes and Olsson (2015) further examined the performance of the SB correction for both correct and misspecified models. They found that with increasing kurtosis, SB led to higher rejection rates for correct models and lower rejection rates for misspecified models. Furthermore, SEM robustness is similarly affected for non-normal Likert scales (Muthén & Kaplan, 1985, 1992) and

symmetric distributions (Hu, Bentler, & Kano, 1992), indicating that kurtosis has a more detrimental effect than skewness. For other popular estimation methods, Olsson, Foss, Troye, and Howell (2000) demonstrated that generalized least squares estimation led to comparable results as ML, whereas weighted least squares was even less robust if the underlying marginal distributions were leptokurtic.

Only few Monte Carlo studies examined effects of the multivariate distribution on SEM beyond the prespecified skewness and kurtosis of the marginal distribution (Foldnes & Grønneberg, 2015; Mair, Satorra, & Bentler, 2012). Mair et al. (2012) presented an approach to generate sample from multivariate non-normal distributions that can be used in SEM, but only applied it with the asymptotic distribution free estimator to validate their generation routine. Foldnes and Grønneberg (2015) investigated the effects of two multivariate non-normal distributions with different tail dependencies  $td_{u1} = 0$ ,  $td_{u2} = 0.93$  but equal multivariate kurtosis  $\beta_{2,2} = 156.4$  for a confirmatory factor model with two latent and four observed variables. Non-zero tail dependence led to parameter biases (bias of latent covariance  $\hat{\Phi}_{td_u=0.93} = 0.043$ ,  $\hat{\Phi}_{td_u=0} = -0.003$ , for  $N = 500$  and  $\Phi = 0$ ) and inflated standard errors ( $SE_{td_u=0.93} = 0.171$ ,  $SE_{td_u=0} = 0.088$ , for  $N = 500$ ), compared to the non-normal distribution with zero tail dependence. Furthermore, the kurtosis of the marginal distributions was higher in the condition with zero tail dependence, so that the effect of tail dependence might be even larger if univariate kurtosis was controlled for. The model was only estimated using standard ML without the SB correction. However, SB corrections are based on the multivariate kurtosis which was equal in both conditions, so results would likely be similar for SB. Overall, SEM appear to be less robust with regards to non-normality, as rejection rates are typically more inflated compared to the general linear model for similar  $N$ . Furthermore, aspects of the multivariate distribution that are not captured by the marginal distributions seem to have a strong impact on the performance of ML estimation.

### 1.3 Methods that Generate Non-Normal Data

Sampling data from a univariate non-normal distribution is not particularly challenging since routines for various distributions are implemented in most software

packages. These routines can be used to investigate robustness of linear regressions and ANOVAs, as the assumption of normality pertains only to single random variables. Similarly, sampling from a multivariate distribution in which all random variables are independent from each other is also not challenging, because routines for single random variables can be used separately for each random variable.

The multivariate case with prespecified dependence (e.g. covariance) among the random variables is only straightforward in the case of a joint normal distribution. In such cases, matrix decomposition can be used on the target covariance matrix  $\Sigma_X$  to obtain an upper triangular matrix  $\mathbf{U}$  with  $\Sigma_X = \mathbf{U}^T \mathbf{U}$ , provided that  $\Sigma_X$  is not singular. The matrix  $\mathbf{U}$  can be multiplied with a sample from an independent joint normal distribution which guarantees to desired covariance. However, the task of generating samples that both comply with certain non-normal distributions and a prespecified covariance matrix is considerably more difficult. A number of studies targeted this issue (e.g. Bradley & Fleisher, 1994; Cook & Johnson, 1981; Foldnes & Olsson, 2016; Headrick, 2002; Headrick & Mugdadi, 2006; Headrick & Sawilowsky, 1999; Koran, Headrick, & Kuo, 2015; Mair et al., 2012; Mattson, 1997; Ruscio & Kaczetow, 2008). In this section, I provide an overview of three popular approaches: power constants (Fleishman, 1978; Vale & Maurelli, 1983), NORTA (NORmal To Anything, Cario & Nelson, 1997), and copulas (Joe, 1997).

### 1.3.1 Power Constants and its Extensions

The power constants approach generates univariate non-normal variables  $X$  with prespecified  $\gamma_1$  and  $\gamma_2$  as

$$X = a + bZ + cZ^2 + dZ^3, \quad (1.21)$$

where  $Z \sim \mathcal{N}(0, 1)$ . The constants  $a$ ,  $b$ ,  $c$ , and  $d$  are obtained by solving four equations provided by Fleishman (1978) to guarantee that  $X$  has the desired skewness and kurtosis. The Vale-Maurelli approach (VM, Vale & Maurelli, 1983) extends the power constants to the multivariate case in three steps and generates samples from a population that also comply with a prespecified covariance matrix  $\Sigma_X$ . First, VM uses the Fleishman equations for each random variable to obtain power constants



associated with the desired skewness and kurtosis. If one were to simulate multivariate normal data according to  $\Sigma_X$  and use Equation 1.21 to create non-normal  $\mathbf{X} = (X_1, \dots, X_d)$ , the covariance of  $\mathbf{X}$  would, in general, be different from  $\Sigma_X$ . This is due to the fact that the function associated with Equation 1.21 (the function that maps  $Z$  to  $X$ ) is non-linear if  $c \neq 0$  or  $d \neq 0$ .<sup>3</sup> In a second step, VM calculates an intermediate correlation matrix  $\Sigma_Z$  that counteracts the distortion caused by the non-normality transformation. Third, samples from normal distributions are drawn according to  $\Sigma_Z$  and transformed by the Fleishman equations to have the desired (univariate) skewness and kurtosis, as well as covariance.

VM is very popular especially in robustness studies (e.g. Curran et al., 1996; Fouladi, 2000; Hu et al., 1992; Muthén & Kaplan, 1985, 1992; Savalei, 2010) and implemented in most SEM software packages like Mplus (Muthén & Muthén, 2010), lavaan (Rosseel, 2012), EQS (Bentler, 2006), and Lisrel (Jöreskog & Sorbom, 2006). The ability to specify skewness and kurtosis in advance allows simulation studies to investigate the range, in which statistical tests can be used for non-normal distributions. However, as demonstrated in the previous sections, skewness and kurtosis are insufficient to fully describe a distribution and some methods might be influenced by other characteristics of the distribution. VM only generates a very specific distribution for given  $\gamma_1$  and  $\gamma_2$ , while other distributions with the same  $\gamma_1$  and  $\gamma_2$  could have a different impact on the robustness of a statistical test. For example, all distributions generated by VM are based on the transformation of normal variables and thus have tail dependence  $td_u = td_l = 0$  (Foldnes & Grønneberg, 2015). Because tail dependence negatively impacts the robustness of SEM, simulation studies that only use VM might draw overly optimistic conclusions concerning the validity of statistical tests in practice. Furthermore, VM is unable to generate certain univariate distributions like the family of  $\chi^2$  distributions and is also limited in the degree of non-normality that can be generated. For example, kurtosis has the lower bound  $\gamma_2 = 1.85$  for symmetric distributions (Headrick & Sawilowsky, 2000). If distributions are asymmetric, this lower bound increases further to e.g.  $\gamma_2 = 4.11$  for  $\gamma_1 = \pm 1.20$ . Headrick (2002) improved VM by generating non-normal  $X$  according

---

<sup>3</sup>However, if  $c = 0$ ,  $d = 0$ , and  $Z \sim \mathcal{N}(0, 1)$ , the resulting  $X$  would have normal distribution as well with  $X \sim \mathcal{N}(a, b^2)$ .

to

$$X = a + bZ + cZ^2 + dZ^3 + eZ^4 + fZ^5, \quad (1.22)$$

instead of Equation 1.21, thereby decreasing the lower bounds of kurtosis to  $\gamma_2 = 1.61$  for symmetric distributions and  $\gamma_2 = 3.91$  for  $\gamma_1 = \pm 1.20$ . However, this extension neither allows any control over the multivariate distribution.

### 1.3.2 NORTA

NORTA (Cario & Nelson, 1997) is an algorithm that allows full specification of the marginal distribution as well as the correlation matrix<sup>4</sup> of random variables. Similarly to VM, NORTA is based on joint normal random variables that are transformed by non-linear functions to comply with the desired marginal distributions. However, instead of a function with power constants, NORTA is based on the inverse *CDF* of the desired marginal distribution. For  $d$  non-normal random variables  $X_i$  ( $1 \leq i \leq d$ ) with desired *CDF*  $F_i$ ,

$$X_i = F_i^{-1}(\Phi(Z_i)), \quad (1.23)$$

where  $Z_i \sim \mathcal{N}(0, 1)$  and  $\Phi$  is the *CDF* of a standard normal distribution. Note that  $Z_i \sim \mathcal{N}(0, 1)$  implies that

$$\Phi(Z_i) \sim \mathcal{U}(0, 1), \quad (1.24)$$

where  $\mathcal{U}(0, 1)$  is a uniform distribution with support  $[0, 1]$ . Applying the inverse *CDF*  $F_i^{-1}$  to the uniform random variable  $\Phi(Z_i)$  ensures that  $X_i$  is distributed according to  $F_i$ . Similarly to VM, the non-normality transformation in Equation 1.23 again affects the covariance among  $Z_i$ , so that  $\Sigma_Z \neq \Sigma_X$  (unless all target distributions  $F_i$  are normal). The problem then is to select an intermediate correlation matrix  $\Sigma_Z$  that gives the desired covariance  $\Sigma_X$ , after the non-normality transformation is applied.

Each element of  $\Sigma_X$  represents the desired correlation between two random variables  $X_i, X_j$  ( $1 \leq i, j \leq d, i \neq j$ ) and is denoted as  $\rho_X(i, j)$ . Importantly,  $\rho_X(i, j)$

---

<sup>4</sup>Note that by specifying a correlation matrix and all marginal distributions, the target covariance matrix is also predefined, because it only depends on the correlation matrix and the variances of the random variables.

only depends on the corresponding elements  $\rho_Z(i, j)$  in  $\Sigma_Z$ , because

$$\rho_X(i, j) = \text{Corr}(X_i, X_j) = \text{Corr}(F_i^{-1}(\Phi(Z_i)), F_j^{-1}(\Phi(Z_j))). \quad (1.25)$$

The correlation  $\rho_X(i, j)$  is defined as

$$\rho_X(i, j) = \frac{E(X_i X_j) - E(X_i)E(X_j)}{\sqrt{\text{Var}(X_i)\text{Var}(X_j)}}. \quad (1.26)$$

NORTA can only adjust  $E(X_i X_j)$  because  $\text{Var}(X_i)$ ,  $\text{Var}(X_j)$  as well as  $E(X_i)$ ,  $E(X_j)$  are already predefined by the corresponding distributions  $F_i$ ,  $F_j$ . For bivariate normal density  $\phi_{\rho_Z(i, j)}$  with correlation  $\rho_Z(i, j)$ , the expected value is

$$E[X_i X_j] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} F_i^{-1}(\Phi(Z_i)) F_j^{-1}(\Phi(Z_j)) \phi_{\rho_Z(i, j)}(Z_i, Z_j) dZ_i dZ_j. \quad (1.27)$$

The goal of NORTA is to find  $\rho_Z(i, j)$  in the equation above, so that  $E[X_i X_j]$  has the desired value. While there is in general no closed form expression of Equation 1.27, Cario and Nelson (1997) show that the function that maps  $\rho_Z(i, j)$  to  $E[X_i X_j]$  is non-decreasing and continuous, thus enabling an efficient numerical search for  $\rho_Z(i, j)$ .

NORTA has the advantage that any inverse *CDF* can be used to generate non-normal target distributions, thereby allowing full control over the marginals. If the *CDF* has defined higher order central moments, NORTA generates samples that also comply with moments beyond skewness and kurtosis, in contrast to VM. Despite this clear advantage and popularity in other fields (e.g. Clemen & Reilly, 1999; Henderson, Chiera, & Cooke, 2000; Lurie & Goldberg, 1998), NORTA has to my knowledge never been used in robustness studies of SEM. One reason for this could be that robustness studies are usually not aimed at investigating a specific type of distribution. Instead, the focus often is on a range of distributions that need to be indicated by a set of measures for non-normality, as introduced in Chapter 1.1. For example, the information that a *t* distribution with  $df = 5$  leads to a robust statistical test is not as useful as claiming that the statistical test is robust for distributions with kurtosis  $\gamma_2 = 6$ . One could obviously choose a *CDF* that is associated with the desired skewness and kurtosis, but the problem then is to vary the *CDF* in a

way to obtain a continuous range for skewness, kurtosis, and (ideally) higher order central moments. Furthermore, NORTA shares the same disadvantage as VM in its lack of control over the multivariate distribution. For example, since the underlying variables are joint normal, NORTA distributions also have zero tail dependence, leading to distributions that - while being beneficial for the robustness of statistical tests - are unrealistic in practice (Foldnes & Grønneberg, 2015). Finally, NORTA is limited in the degree of dependence that can be generated in the correlation matrix  $\Sigma_X$ . Specifically, the matrix  $\Sigma_Z$  obtained by the algorithm might not be positive semi-definite (and therefore not a correlation matrix), even though a multivariate distribution with the given correlation and marginals exists (Ghosh & Henderson, 2002). Ghosh and Henderson (2003) showed that this problem becomes more likely as the number of random variables increases and suggested a modification of NORTA that partially solves the issue.

### 1.3.3 Copulas

Copulas can be understood as a mathematical reformulation of a multivariate distribution (for an overview, see Joe, 1997). For *CDFs*  $F_1, \dots, F_d$ , the multivariate *CDF*  $F$  can be written as

$$F(x_1, \dots, x_d) = C(F_1[x_1], \dots, F_d[x_d]). \quad (1.28)$$

That is, there exists a function  $C : [0, 1] \times \dots \times [0, 1] \rightarrow [0, 1]$ , called copula, that expresses the multivariate distribution in terms of the marginals and  $C$  is unique if the marginals are continuous, which is known as Sklar's theorem (Sklar, 1959). Note that for any distribution,  $F_X(X) \sim U(0, 1)$  if  $X$  is distributed according to  $F_X$ .

There are different families of copulas that can be used to sample from multivariate distributions. One family consists of Gaussian copulas, which capture the dependence among random variables in the same way as a multivariate normal distribution but can have arbitrary marginals (Clemen & Reilly, 1999). Specifically, a

multivariate distributions has a Gaussian copula if

$$\mathbf{Z} = (\Phi^{-1}(F_1(X_1)), \dots, \Phi^{-1}(F_d(X_d))), \quad (1.29)$$

has a multivariate normal distribution, where  $\Phi^{-1}$  is the inverse of a standard normal *CDF*. As can be seen by comparing Equations 1.23 and 1.29, the multivariate distributions generated by NORTA always have Gaussian copulas. Similarly, Foldnes and Grønneberg (2015) showed that VM also leads to Gaussian copulas.

In general, copulas do not correspond to a particular correlation matrix, because correlations also depend on the marginals of the distribution. Instead, copulas capture dependence among random variables by a set of parameters of the function  $C$ . For example, the Clayton copula (Clayton, 1978) is defined as

$$C(U_1, \dots, U_d) = \Psi(\Psi^{-1}(U_1) + \dots + \Psi^{-1}(U_d)) \quad (1.30)$$

where  $U_i = F_i^{-1}(X_i)$ ,  $1 \leq i \leq d$ , and

$$\Psi(t) = (1 + t)^{-\frac{1}{\theta}}, \quad (1.31)$$

with dependence parameter  $\theta$  and  $\theta > 0$ . Figure 1.5 shows the contour plots of bivariate distributions based on a Clayton copula with  $\theta = 1$  or  $\theta = 2.5$  (Yan, 2007). The marginal distributions were set to be either both standard normal or standard normal for  $X_1$  and exponential with rate  $\lambda = 0.5$  for  $X_2$ . The resulting correlations are displayed in Table 1.1 and vary depending on both  $\theta$  and the selected marginal distributions. Since robustness studies often need to prespecify a correlation matrix, copulas are difficult to use. Mair et al. (2012) suggested to generate random variables  $\mathbf{X} = (X_1, \dots, X_d)$  with non-normal distribution according to

$$\mathbf{X} = \mathbf{Y} S_Y^{-\frac{1}{2}} \Sigma_X^{\frac{1}{2}} \quad (1.32)$$

where  $\mathbf{Y} = (Y_1, \dots, Y_d)$  is a random variable based on a copula,  $\Sigma_X$  is the desired correlation matrix, and  $S_Y$  is the covariance matrix of  $\mathbf{Y}$ . If  $\mathbf{Y}$  is mean-centered

and scaled, we have

$$\frac{1}{N} \mathbf{X}' \mathbf{X} = \Sigma_X^{\frac{1}{2}} S_Y^{-\frac{1}{2}} \left( \frac{1}{N} \mathbf{Y}' \mathbf{Y} \right) S_Y^{-\frac{1}{2}} \Sigma_X^{\frac{1}{2}} \quad (1.33)$$

so that

$$\begin{aligned} \frac{1}{N} \mathbf{X}' \mathbf{X} &= \Sigma_X^{\frac{1}{2}} S_Y^{-\frac{1}{2}} S_Y S_Y^{-\frac{1}{2}} \Sigma_X^{\frac{1}{2}} \\ &= \Sigma_X^{\frac{1}{2}} \Sigma_X^{\frac{1}{2}} \\ &= \Sigma_X, \end{aligned} \quad (1.34)$$

and  $\mathbf{X}$  has the desired covariance matrix. While the approach by Mair et al. (2012) offers some control over the multivariate distribution, the transformation in Equation 1.32 changes the marginals of  $\mathbf{X}$  depending on  $\Sigma_X$ . Therefore, the approach does not allow to specify univariate distributions (or at least skewness and kurtosis) of  $\mathbf{X}$  in advance.

Table 1.1: Correlation of two random variables with Clayton copula and different marginal distributions

Marginals	$\theta = 1$	$\theta = 2.5$
$X_1, X_2 \sim \mathcal{N}(0, 1)$	.50	.74
$X_1 \sim \mathcal{N}(0, 1), X_2 \sim \text{Exp}(0.5)$	.36	.57

*Note.*  $\theta$  = dependence parameter of the Clayton copula.  $\text{Exp}(\lambda)$  = Exponential distribution with rate  $\lambda$ .

The algorithm presented in the next sections of my thesis attempts to solve this issue. Other algorithms offer no control over the multivariate distribution<sup>5</sup> (VM and NORTA), the marginal distribution (Mair et al., 2012), or the correlation matrix (copulas). In contrast, my algorithm allows manipulations of all three parameters simultaneously. The algorithm is developed and applied in three papers, which I will summarize in the next section.

<sup>5</sup>That is, a given correlation matrix and either marginal distribution (NORTA) or skewness and kurtosis (VM) fully determine the distribution generated by both NORTA and VM, despite the fact that other distributions with the same properties exist.

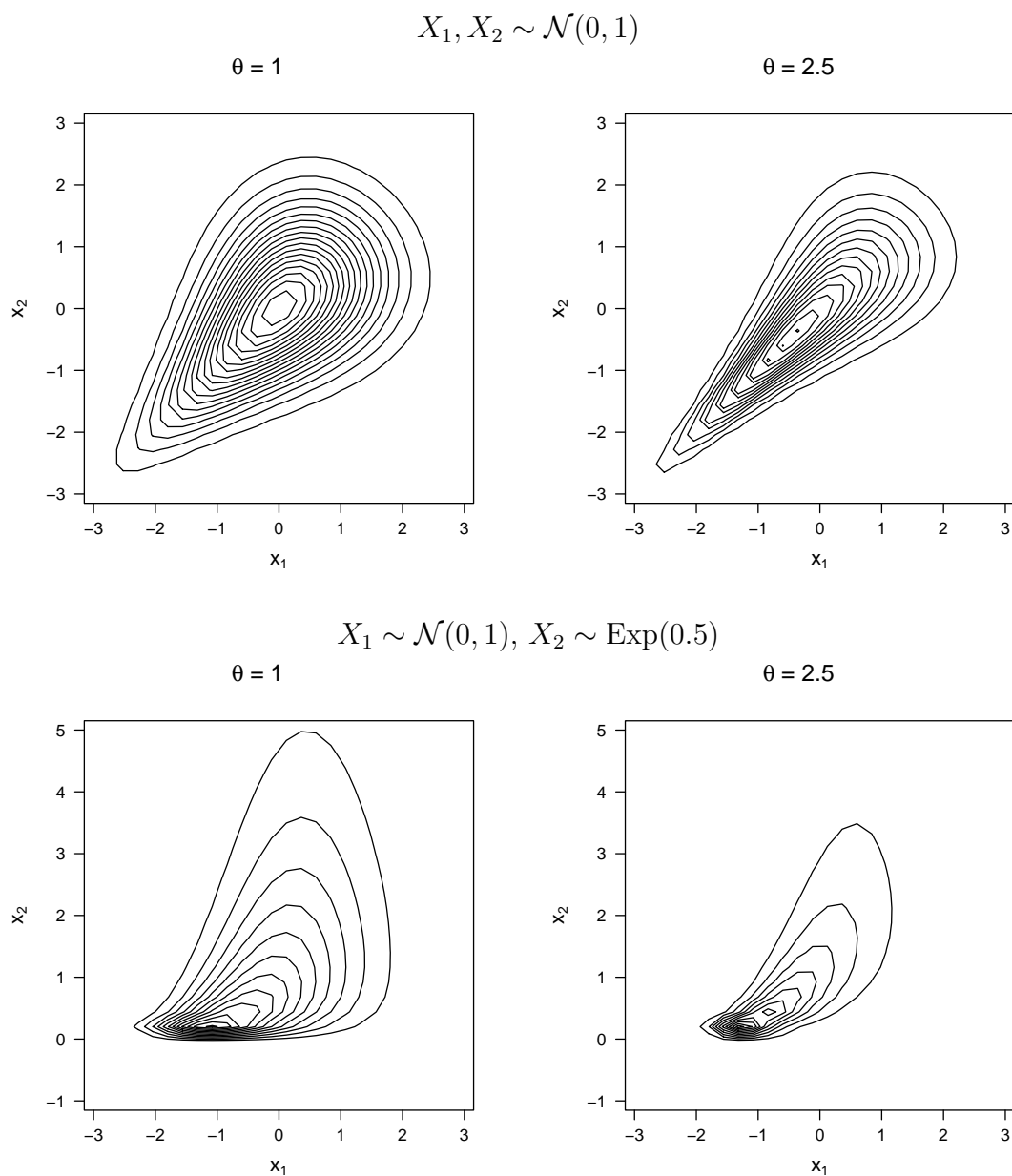


Figure 1.5: Contour plots of bivariate distributions with Clayton copula and dependence parameter  $\theta = 1.5$  (left) or  $\theta = 2.5$  (right). The upper panels display Clayton copulas when both marginals are standard normal. In the lower panels,  $X_1$  also has standard normal distribution and  $X_2$  has exponential distribution with rate  $\lambda = 0.5$

# Summary of Manuscripts

The following sections provide summaries of the three manuscripts on which this thesis is based. For the sake of brevity, I will focus on the main results of each manuscript, as a more technical description of the algorithms and simulation conditions can be found in the original manuscripts appended to this dissertation. Furthermore, I will discuss how each manuscript contributes to the central goal of this dissertation: (1) to develop an algorithm that allows for a more flexible manipulation of the underlying (multivariate) distribution of generated samples, and (2) to clarify the impact of distributional characteristics beyond central moments on the robustness of statistical methods.

## 2.1 A method for multivariate non-normality

Auerswald, M., & Moshagen, M. (2015). Generating correlated, non-normally distributed data using a non-linear structural model. *Psychometrika*, *80*, 920-937.

In this article, we developed the linking functions algorithm that creates non-normally distributed random variables with a prespecified covariance matrix. The basic idea stems from a structural model with normally distributed latent factors and errors. Users provide the algorithm with a latent covariance matrix and loadings for each observed variable, which together define the target covariance matrix  $\Sigma_T$ . Additionally, a set of so-called linking functions needs to be specified. The algorithm introduces non-normality in the observed variables by applying these linking functions to the latent factors, the errors, or both. Specifically, an observed variable  $M$  is defined as

$$M = b \cdot g(L) + c \cdot h(\xi), \quad (2.1)$$



where  $L \sim \mathcal{N}(0, 1)$  is a latent factor,  $\xi \sim \mathcal{N}(0, 1)$  is a unique error,  $g, h$  are linking functions, and  $b, c$  are correction factors estimated by the algorithm. Similarly to VM and NORTA, non-linear functions  $g$  and  $h$  lead to non-normal  $M$  but distort the covariance matrix that would be expected based on the loadings and latent covariance matrix.

This distortion is corrected in two steps. First, the algorithm applies the inverse standard normal *CDF* to a large vector of quantiles (e.g.  $[10^{-7}, 2 \cdot 10^{-7}, \dots, 1 - 10^{-7}]$ ) to create a vector  $z$ . This vector contains values that would be expected when drawing a very large sample from a standard normal distribution and is used to estimate the effects of the non-normality transformation, i.e.

$$\text{cor}(z, g(z)) \approx \text{cor}(L, g(L)), \quad (2.2)$$

because  $L \sim \mathcal{N}(0, 1)$ . The first correction estimates  $b$  and  $c$  from Equation 2.1 based on the correlation  $\text{cor}(z, g(z))$  and has the goal that the correlation between  $M$  and  $L$  is equal to the prespecified (standardized) loading  $\lambda$ . If  $L, M$ , and  $g(L)$  were standardized

$$\begin{aligned} \lambda &= \text{cor}(M, L) \\ &= \text{cor}(M, g(L)) \cdot \text{cor}(g(L), L), \end{aligned} \quad (2.3)$$

because  $\xi$  and  $L$  are independent. The algorithm requires that  $|\text{cor}(g(L), L)| \geq |\lambda|$  and  $\text{cor}(g(L), L) \neq 0$ , in which case

$$b^* = \frac{\lambda}{\text{cor}(z, g(z))}, \quad (2.4)$$

if  $g(L)$  is standardized. For unstandardized  $g(L)$ , the standard deviation of  $g(L)$  needs to be estimated and addressed as

$$b = \frac{\lambda}{\text{cor}(z, g(z))\sigma(g(z))}. \quad (2.5)$$

At this point, the loadings for each observed variable are correctly specified. However, there is remaining deviation in the resulting covariance matrix because the first correction only addresses dependencies between the observed variables and the latent factors, but not among the observed variables themselves. This remaining

deviation depends on the similarity between linking functions (and therefore the resulting marginal distributions) of different observed variables. For example, if two observed variables have the same linking functions  $g$  and  $h$ , the resulting distributions will also be similar. The first correction then erroneously increases the expected correlation among observed variables with similar distributions beyond the desired correlation specified in  $\Sigma_T$ . The algorithm accounts for this deviation by correlating the unique errors in Equation 2.1 accordingly.

One advantage of the linking functions approach as compared to VM is the flexibility of transformation functions that can be used to create different marginal distributions. While VM only uses power functions as in Equation 1.21, the linking functions approach can use any function as long as  $|\text{cor}(g(L), L)| \geq |\lambda|$  holds. However, it is often desirable in robustness studies to control the degree of non-normality in the underlying distributions, which depends on the choice of functions. In general, the degree of non-normality can be controlled by defining a linking function  $g_\alpha$  as

$$g_\alpha = \alpha \cdot g + (1 - \alpha) \cdot id, \quad (2.6)$$

for  $\alpha \in [0, 1]$ , the identity function  $id$ , and a linking function  $g$ . The reason is that only non-linear linking functions lead to non-normal distributions. If both  $g$  and  $h$  in Equation 2.1 are linear functions (such as  $id$ ), the resulting variable  $M$  would be the sum of two normally distributed variables and also be normal. Therefore, the function defined in Equation 2.6 (and applied to both the latent factor and unique error) leads to normal distributions if  $\alpha = 0$ , so that  $g_\alpha = id$ . Increasing  $\alpha$ , up to  $\alpha = 1$ , would result in increasing non-normality in the marginal distribution because  $g_\alpha$  is the weighted sum of  $id$  and  $g$ . The function  $g_\alpha$  can also be used to approximate one prespecified central moment if an appropriate function  $g$  is chosen. For example, if  $g$  results in kurtosis  $\gamma_2 = 30$ , any value for kurtosis between 3 (the corresponding value of a normal distribution and  $id$ ) and 30 can be chosen. The algorithm applies a bisection search to obtain a value for  $\alpha$  that matches the desired moment.

The main contribution of this paper to NOTAMO is the ability to manipulate the multivariate distribution, depending on whether non-normality is introduced through non-linear functions for the unique errors, the latent factors, or both. Figure 2.1 shows the bivariate distribution for an exponential linking function, one factor

with standardized loadings of  $\lambda = .7$ , and prespecified kurtosis  $\gamma_2 = 15$ . If non-normality is introduced by non-linear functions for the latent factors, while the functions of the errors are linear, outliers for  $X_1$  are more likely to occur given that  $X_2$  is also an outlier. In contrast, if only the functions of the unique errors are non-linear, outliers for  $X_1$  are far less likely given that  $X_2$  is an outlier. If both errors and latent factors are non-normal due to the non-linear exponential linking function, outliers of  $X_1$  and  $X_2$  are more or less independent. Importantly, the kurtosis of the resulting marginal distributions is 15 and the correlation of  $X_1$  and  $X_2$  is  $.49 (= \lambda^2)$  in all three cases. However, the multivariate distribution is different depending on the way in which non-normality is introduced. Therefore, the linking functions approach allows the manipulation of the multivariate distribution beyond the prespecified central moment and correlation matrix. Note that the distributions displayed in Figure 2.1 do not differ with regard to tail dependence. While tail dependence is also associated with the probability that  $X_1$  is an outlier given that  $X_2$  is an outlier, upper (lower) tail dependence is the limit of this probability for quantile  $q \rightarrow 1$  ( $q \rightarrow 0$ ).

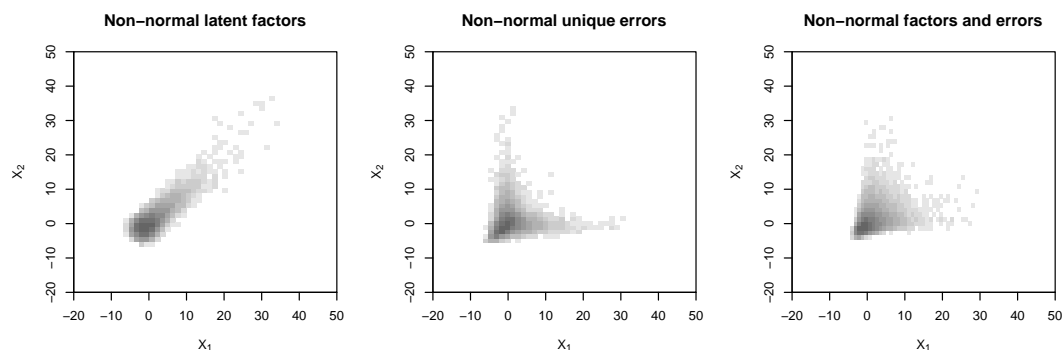


Figure 2.1: Heatmap of samples from bivariate distributions generated with the linking functions approach. The left panel shows the bivariate distribution given that the linking functions of the latent factors are non-linear; the center panel displays non-linear functions for the unique errors. In the right panel, both latent factors and unique errors have non-linear linking functions. All distributions are specified to have univariate kurtosis  $\gamma_2 = 15$  and a correlation of  $.49$ .

The article includes a robustness study to illustrate the relevance of the multivariate distribution, comparing non-normal latent factors, non-normal unique errors, and samples generated with VM for a correctly specified confirmatory factor model.

The marginal distributions had the same kurtosis  $\gamma_2 = 15$  in all three conditions. The empirical rejection rates for ML estimation and the SB correction are displayed in Table 2.1. As can be seen, the robustness of both test statistics varied depending on the multivariate distribution. While non-normal unique errors barely increased the empirical rejection rate beyond the nominal  $\alpha$ , non-normal latent factors led to inflated rejection rates for both  $T_{ML}$  and  $T_{SB}$ . The VM algorithm seems to produce conditions between non-normal errors and latent factors, but is in contrast to the linking functions approach unable to control the multivariate distribution. These results also indicate that univariate kurtosis does not properly assess the non-normality that is critical for the statistical test in SEM. This finding is especially relevant for empirical practice, as some data conditions lead to severely inflated rejection rates, whereas other conditions with the same kurtosis are more or less robust. In the former case, this might lead to overly optimistic conclusion regarding the data set at hand. In the latter case, practitioners might not be able to publish their results, despite the fact that the violation of normality barely affected the validity of the statistical test.

Table 2.1: Empirical rejection rates (in %) under different conditions of non-normality.

Non-normality	$T_{ML}$	$T_{SB}$
Non-normal errors	5.3	6.5
Non-normal latent factors	23.6	13.0
VM	13.9	7.4

*Note.*  $T_{ML}$  = Maximum likelihood estimation,  $T_{SB}$  = Satorra-Bentler correction, VM = Vale-Maurelli procedure. Nominal  $\alpha = 5\%$ .

In sum, the first manuscript presented an algorithm that allowed to prespecify a correlation matrix and one central moment of the marginal distributions while simultaneously manipulating the multivariate distribution. We show that the multivariate distribution needs to be considered in robustness studies, which is not possible in other data generation algorithms. For both VM and NORTA, the multivariate distribution is a direct function of the prespecified univariate non-normality

and correlation matrix. However, the linking function approach also has three drawbacks. First, it only allows the specification of one central moment. Many robustness studies aim at investigating (at least) skewness and kurtosis simultaneously, which cannot be achieved with the linking functions approach. Second, the vector  $z$  that is used to estimate distortions in the correlation matrix results in rather slow processing times.<sup>1</sup> Third, the algorithm is only implemented in MATLAB, which is proprietary software and not openly available. All three disadvantages are addressed in the next article.

## 2.2 Combining moment- and distribution-based methods

Auerswald, M., & Moshagen, M. (2017). *Sampling from arbitrary non-normal distributions with given covariance and central moments*. Manuscript submitted for publication.

The second manuscript introduces the NOTAMO algorithm, which is an extension and combination of the linking functions approach and NORTA. In contrast to NORTA, the algorithm allows the specification of standardized central moments in advance. This has the advantage that simulation studies can vary the degree of non-normality continuously, instead of relying on only one specific non-normal distribution. Furthermore, NOTAMO also allows the manipulation of the multivariate distribution while keeping the marginal distributions and correlation matrix constant. Compared to the linking functions approach, NOTAMO has the advantage that more than one central moment can be prespecified. The algorithm is also faster<sup>2</sup> and implemented in the popular open source programming language R (R Core Team, 2017).

The basic idea of NOTAMO is to select appropriate inverse *CDFs* for NORTA that comply with the prespecified central moments. NORTA creates non-normal  $X$

---

<sup>1</sup>The usage of a smaller vector  $z$  is also not advised, because it would cause a lower accuracy with regards to the correlation matrix.

<sup>2</sup>For example, on an Intel Core i7-4790K, estimating the appropriate covariance matrix of 12 observed variables takes 47 *min* with the linking functions approach but only 18 *s* with NOTAMO.

as

$$X = F^{-1}(\Phi(Z)), \quad (2.7)$$

where  $Z \sim \mathcal{N}(0, 1)$  and  $F^{-1}$  is the inverse *CDF* of the desired marginal distribution. Instead of one inverse *CDF*, NOTAMO requires a set of  $m$  inverse *CDFs*  $F_1^{-1}, \dots, F_m^{-1}$  for each marginal distribution. NOTAMO attempts to find parameters  $a_1, \dots, a_m$  such that the weighted sum

$$F^{-1} = \sum_{j=1}^m a_j F_j^{-1} \quad (2.8)$$

is an inverse *CDF* that matches the desired central moments when applied in Equation 2.7 and

$$\sum_{j=1}^m a_j = 1, \quad a_j \geq 0. \quad (2.9)$$

That is, NOTAMO implements a search for weights in a mixture of quantile distributions, so that the resulting inverse *CDF* complies with the prespecified moments. The central moments are estimated by a vector  $z$  that contains values that would be expected when drawing a large sample from a normal distribution, similar to  $z$  in the linking functions approach. For example,  $z = \Phi^{-1}([10^{-7}, 2 \cdot 10^{-7}, \dots, 1 - 10^{-7}])$  can be used to approximate the expected moments associated with  $F^{-1}$ , because

$$F^{-1}(\Phi[z]) \approx F^{-1}(\Phi[Z]), \quad (2.10)$$

for  $Z \sim \mathcal{N}(0, 1)$ . For a set of  $m$  inverse *CDFs* and  $k$  prespecified moments, this results in a system of  $k + 1$  non-linear equations (one for each moment and Equation 2.9) and  $m$  unknowns, the parameters  $a_1, \dots, a_m$  in Equation 2.8. NOTAMO attempts to solve the system of non-linear equations using algorithms implemented in the packages `nloptr` (Johnson, 2014) and `rootsolve` (Soetaert, 2009) in R. The algorithm is not guaranteed to converge, in part because a solution might not exist for a given set of moments and inverse *CDFs*. However, we conducted a simulation study in the manuscript, demonstrating that NOTAMO is applicable to a wide range of non-normality conditions and reproduces the target central moments with high accuracy. Once the weight parameters in Equation 2.8 are determined, the resulting inverse *CDFs* and desired correlation matrix are passed to the NORTA algorithm,

which is also implemented in R (Su, 2014).

NOTAMO combines the advantages of algorithms that prespecify a set of moments, such as VM for skewness and kurtosis, and algorithms that prespecify a particular distribution, such as NORTA. Specifying a set of central moments in advance allows robustness studies to continuously vary the degree of non-normality as indicated by measures (e.g., skewness and kurtosis) that can also be applied to an empirical sample. NORTA, on the other hand, allows to investigate different distributions with the same skewness and kurtosis. NOTAMO can prespecify central moments and, at the same time, create different distributions with the same first central moments, depending on the inverse *CDFs* that were provided to the algorithm. For example, Figure 2.2 displays two (standardized) marginal distributions with the same skewness  $\gamma_1 = 0$  and kurtosis  $\gamma_2 = 2$ , but different quantile mixtures. In the left panel, NOTAMO estimated the weights for the inverse *CDF* of a standard normal distribution and a uniform distribution, the latter with support  $[0, 1]$ . The right panel displays the quantile mixture based on a standard normal distribution, as well as a binomial distribution with one trial and success probability  $p = .5$ . The resulting marginal distributions clearly vary depending on the set of inverse *CDFs* from which the weights of the distribution are estimated.

NOTAMO also allows manipulations of the multivariate distribution, independently of prespecified central moments. Similarly to the linking functions approach, the basic idea is to define each random variable of interest  $X_i$  ( $1 \leq i \leq d$ , for  $d$  random variables) as the sum of two random variables  $L_i$  and  $E_i$ , so that

$$X_i = L_i + E_i, \quad 1 \leq i \leq d. \quad (2.11)$$

The random variables  $L_1, \dots, L_d$  are correlated, whereas  $E_1, \dots, E_d$  are independent and thus uncorrelated. Furthermore, all  $L_i$  are required to be independent from all  $E_i$  ( $1 \leq i \leq d$ ). The NOTAMO algorithm can be used to generate either non-normal  $L_i$  or  $E_i$ , whereas the other set of random variables is normally distributed. If the  $L_i$  are non-normal, the resulting multivariate distribution will be similar to a linking functions distribution, in which only the latent factors have non-linear functions (see Figure 2.3). Non-normal  $E_i$  lead to distributions similar to non-linear linking functions for the unique errors, because the non-normal variables are independent.

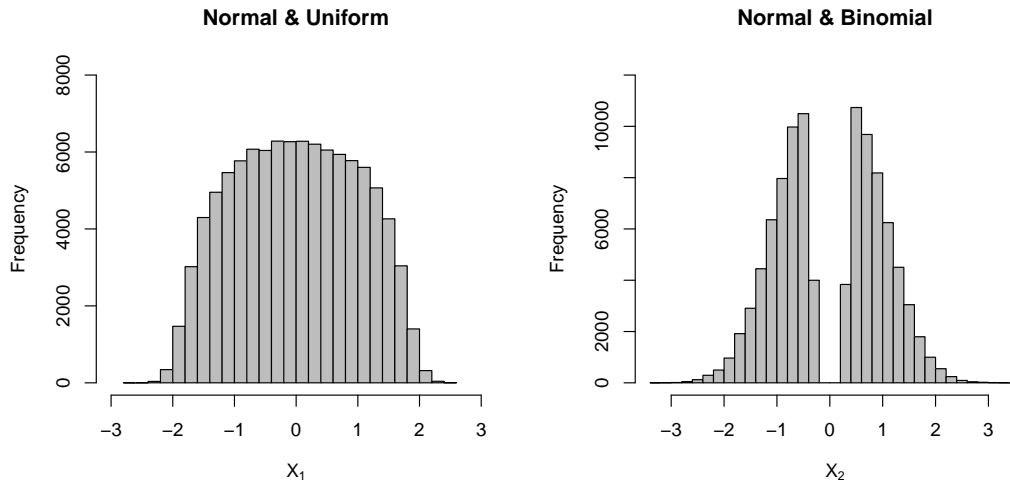


Figure 2.2: Histograms of two marginal distributions generated using NOTAMO with sample size  $N = 100,000$ . Both variables have the same mean, variance, skewness, and kurtosis. The left panel shows the result of a quantile mixture of a standard normal and a uniform distribution  $\mathcal{U}(0, 1)$ . In the right panel, the marginal distribution is estimated based on a standard normal and a binomial distribution with  $\mathcal{B}(n = 1, p = .5)$

One strength of the suggested procedure is that the resulting marginal distributions will be exactly the same if either  $L_i$  or  $E_i$  is generated with NOTAMO. For any distribution  $\mathcal{D}$ , this follows from the fact that

$$X_1 = L_1 + E_1, \text{ where } L_1 \sim \mathcal{D}, E_1 \sim \mathcal{N}(0, 1), \quad (2.12)$$

and

$$X_2 = L_2 + E_2, \text{ where } L_2 \sim \mathcal{N}(0, 1), E_2 \sim \mathcal{D}, \quad (2.13)$$

lead to the same distribution for  $X_1$  and  $X_2$  because  $E_1, E_2$  and  $L_1, L_2$  are independent.

Overall, the second manuscript showed that the combination of the linking functions approach and NORTA yields a powerful algorithm for robustness studies. NOTAMO can be applied in a variety of research contexts and is especially useful if the assumption of normality pertains to more than one random variable. The next manuscript investigates the effects of multivariate non-normality on factor extraction criteria in exploratory factor analysis (EFA).



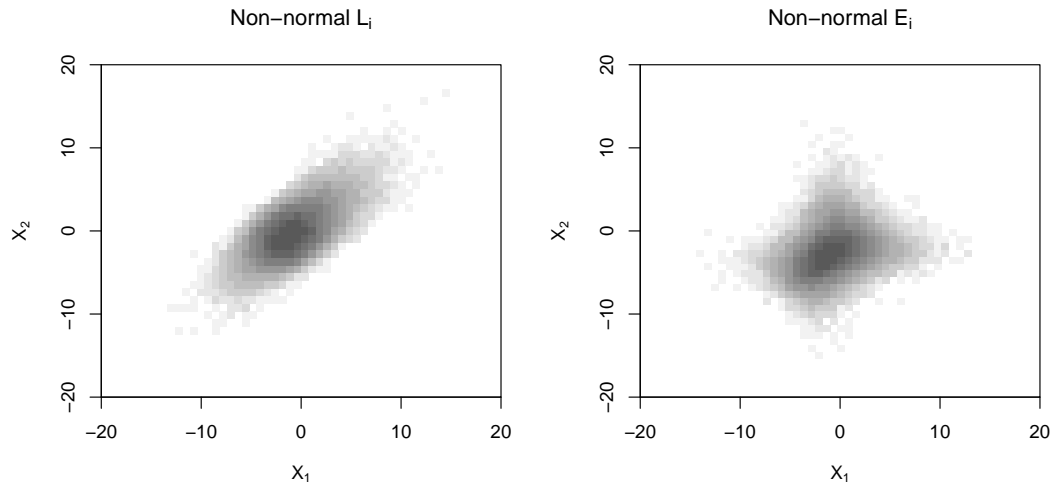


Figure 2.3: Heatmap of samples with  $N = 1,000,000$  from bivariate distributions generated with NOTAMO. In the left panel, the correlated random variables  $L_i$  (see Equation 2.11) are non-normal, the right panel displays non-normal  $E_i$ . The resulting distributions have a correlation of .45 in both cases. The marginal distributions of  $X_1$  and  $X_2$  have prespecified skewness  $\gamma_2 = 0.5$  and kurtosis  $\gamma_2 = 5.5$ .

### 2.3 Non-normality and exploratory factor analysis

Auerswald, M., & Moshagen, M. (2017). *How to determine the number of factors to retain in exploratory factor analysis? A comparison of extraction methods under realistic conditions*. Manuscript submitted for publication.

The first two manuscripts developed the NOTAMO algorithm that allows to perform robustness studies in which the correlation matrix, univariate distributions, their central moments, and the multivariate distribution can be specified in advance. The third paper consists of an exemplary application of the linking functions algorithm in the context of EFA.

EFA is a popular tool to investigate latent factors underlying a large number of observed variables. The model assumes a number of potentially correlated common factors that explain (co)variations among the observed variables, as well as one unique error for each observed variable (Thurstone, 1947). The latent factors are supposed to reflect the underlying psychological variables of interest, whereas the unique errors are assumed to represent item-specific measurement error. A central

problem in EFA is to decide how many factors need to be extracted, because both under- and overextractions (extracting too few or too many factors, respectively) can lead to considerable error, for example in the estimation of factor loadings (Wood, Tataryn, & Gorsuch, 1996).

The decision on the number of factors is typically based on so-called extraction criteria, which are decision heuristics based on the observed covariance matrix. The most prominent are the Kaiser criterion (Kaiser, 1960), Cattell's scree test (Cattell, 1966), and parallel analysis (Horn, 1965). Parallel analysis is often considered as the method of choice, because it displays the highest ratio of correctly retrieved factors (referred to as accuracy) in Monte Carlo studies (e.g. Peres-Neto, Jackson, & Somers, 2005; Zwick & Velicer, 1986). However, four new extraction criteria have been suggested recently that displayed a higher accuracy in some simulation conditions: (1) the empirical Kaiser criterion (Braeken & van Assen, in press), (2) revised parallel analysis (Green, Levy, Thompson, Lu, & Lo, 2012), (3) comparison data (Ruscio & Roche, 2012), and (4) the Hull method (Lorenzo-Seva, Timmerman, & Kiers, 2011). The purpose of the manuscript was to compare these new extraction criteria to parallel analysis under conditions commonly encountered in empirical research, including non-normally distributed observed variables.

Of the five criteria under investigation, only the empirical Kaiser criterion and the Hull method explicitly assume multivariate normality in the observed variables, while the other three criteria do not assume any particular distribution. However, all criteria are based on the sample covariance matrix and sampling errors for covariances are larger in leptokurtic distributions (DeCarlo, 1997). Hence, the higher sampling variations of (co)variances might lead to lower accuracy of all five criteria. We assumed that this would in turn be due to a tendency to overextract, because additional latent factors might account for the additional variability. Previous Monte Carlo studies showed that the accuracy of parallel analysis is more or less independent of the underlying distribution (Dinno, 2009; Garrido, Abad, & Ponsoda, 2013; Glorfeld, 1995; Peres-Neto et al., 2005). However, these studies only manipulated the marginal distributions and only considered traditional parallel analysis.

The simulation study involved six orthogonally manipulated independent variables to represent a wide range of plausible data conditions, one of which was the

underlying distribution of the observed variables. We considered a multivariate normal distribution and two types of non-normal distributions generated with the linking functions approach. These non-normal distributions applied four non-linear functions either to the latent factors or the unique errors, while the other functions were linear (see Figure 2.1). We prespecified the kurtosis to be  $\gamma_2 = 12$  and used the following linking functions:

- $f_1(x) = x^5 + x^3$
- $f_2(x) = e^{2x}$
- $f_3(x) = \begin{cases} \sqrt{x}, & \text{for } x > 0 \\ -x^2, & \text{for } x \leq 0 \end{cases}$
- $f_4(x) = \begin{cases} -50, & \text{for } x < -3 \\ -1, & \text{for } -3 \leq x < 0 \\ 1, & \text{for } 0 \leq x < 3 \\ 50, & \text{for } x \geq 3 \end{cases}$

Table 2.2 shows the results for accuracy and bias of all five extraction criteria. Bias was calculated as the number of suggested factors minus the true number of factors in the population. Thus, positive values indicate overextraction, negative values indicate underextraction, and zero indicates no bias. As can be seen, non-normality did not lead to lower accuracy for any extraction criterion except for comparison data. In line with previous simulation studies, traditional parallel analysis was barely affected by the underlying distribution. The other three criteria under consideration even displayed higher accuracy if the observed variables were non-normal, especially if non-normality was introduced through the unique errors. One explanation for this unexpected advantage in leptokurtotic conditions might be that revised parallel analysis, the Hull method, and the empirical Kaiser criterion generally underestimated the number of factors. However, non-normal distribution increased the number of suggested factors on average, hence counteracting the general tendency to underextract and increasing accuracy overall.

Traditional parallel analysis displayed a high accuracy overall and identified the correct number of factors most often for every distribution under consideration. In the manuscript, we also explore the performance of so-called combination rules that use the suggested number of factors of multiple criteria in conjunction with each other and improve overall accuracy even further. However, parallel analysis is the best single criterion and can be used if data are not normally distributed, at least if the kurtosis of observed variables is not larger than  $\gamma_2 = 12$ . The results also underline that manipulations of the multivariate distribution do not necessarily lead to implications that are different from studies that only considered the marginal distribution. The accuracy of traditional parallel analysis was unaffected in both cases (Dinno, 2009; Garrido et al., 2013; Glorfeld, 1995; Peres-Neto et al., 2005).

Table 2.2: Average accuracy and bias of extraction criteria under different distributional conditions

Average accuracy (in %)					
Distribution	PA-T	PA-R	Hull	CD	EKC
Normal	92	73	84	82	82
Lat-NN	91	77	85	74	84
Err-NN	94	82	89	78	88
Average bias (with standard deviation)					
Distribution	PA-T	PA-R	Hull	CD	EKC
Normal	-0.10 (0.51)	-0.49 (1.13)	-0.50 (1.22)	-0.12 (0.79)	-0.37 (0.92)
Lat-NN	-0.10 (0.53)	-0.27 (0.96)	-0.43 (1.15)	0.03 (0.80)	-0.32 (0.85)
Err-NN	-0.07 (0.42)	-0.22 (0.87)	-0.33 (1.03)	0.08 (0.67)	-0.25 (0.76)

*Note.* Bias is calculated as the difference between extracted factors and underlying factors. PA-T = traditional parallel analysis, PA-R = revised parallel analysis, Hull = Hull method, CD = comparison data, EKC = Empirical Kaiser Criterion, Lat-NN = non-normal latent variables and normal errors, Err-NN = non-normal error variables and normal latent variables.

In sum, this last manuscript demonstrated that there is considerable variability which methods become less accurate as the result of multivariate non-normality. Most extraction criteria for EFA were not negatively affected by non-normality and displayed comparable or even higher accuracy in non-normal conditions. If the decision on the number of factors should be based on a single criterion, we recommend parallel analysis which displayed the highest accuracy overall and was unaffected by non-normal distributions.

## General Discussion & Outlook

The goal of my thesis was the development of the NOTAMO algorithm that allows sampling from multivariate non-normal distributions for robustness studies. The algorithm can specify the univariate distribution, the associated central moments, and the correlation matrix in advance, while simultaneously manipulating the multivariate distribution. The first manuscript introduced the linking functions approach and, thereby, one idea on how a multivariate distribution can be manipulated. The linking functions approach creates non-normal random variables by adding two other random variables, conceptualized as latent factors and unique errors of a structural model. The algorithm applies (potentially) non-linear functions to these random variables and the resulting multivariate distribution varies, depending on which functions are chosen as non-linear. We also demonstrated that this variation in the multivariate distribution has a large effect on the robustness of model tests in SEM.

In the second manuscript, we combined the linking functions approach with NORTA. The resulting algorithm, NOTAMO, is the main result of my thesis. It allows the manipulation of the multivariate distribution, similar to the linking functions approach, but can prespecify any number of central moments at the same time. NOTAMO also has the advantage that the extent of non-normality can be varied continuously, which is useful for Monte Carlo studies that aim to provide guidelines for empirical research.

Finally, the third manuscript examined the effect of, among others, the multivariate distribution on factor extraction criteria in EFA. We investigated two types of non-normal distributions that led to highly different results for SEM, but found no comparable effect for the extraction criteria. Moreover, most extraction criteria were not negatively affected by any type of non-normality that we investigated and accuracies were comparable or even higher if the underlying distribution was

not multivariate normal. Parallel analysis displayed the highest accuracy and is a suitable extraction criterion even if observed variables are moderately leptokurtic.

Overall, the NOTAMO algorithm is primarily beneficial for Monte Carlo studies that investigate the robustness of statistical tests. ANOVAs and  $t$  tests assume that the observations in each group are normally distributed, whereas linear regressions assume a normally distributed error (e.g. Tabachnick & Fidell, 2012). In these cases, the assumption is made for a single random variable and routines from any modern software package can be used to explore the effects of non-normality. However, NOTAMO can be advantageous when a distribution with specific standardized central moments (e.g., skewness and kurtosis) is desired. The ability to continuously vary a univariate distribution based on a mixture of quantile distributions might simplify the search for an appropriate distribution that complies with given central moments. Additionally, NOTAMO can be used to clarify whether skewness and kurtosis are indeed crucial in assessing the degree to which a non-normal distribution affects robustness. For example, a simulation study could employ NOTAMO to investigate whether two non-normal distributions with the same skewness and kurtosis result in the same robustness of basic methods such as widely used linear regressions or  $t$  tests. This would in turn benefit empirical research that routinely has to consider non-normality because a majority of observed variables is not normal (Cain et al., in press).

Other statistical tests and procedures incorporate a multivariate normality assumption. For example, the observed variables in SEM are assumed to be normal when ML, weighted least squares, or generalized least squares is used for estimation (Bollen, 1989; Browne, 1974). Multilevel models assume that the random intercepts and slopes of the model have a multivariate normal distribution (Hox, 2010). In ANOVAs with repeated measures or MANOVAs, the assumption also pertains to the observed variables (Tabachnick & Fidell, 2012). Additionally, the treatment of missing data, for example with full information ML, assumes multivariate normality (Tabachnick & Fidell, 2012). Robustness studies for all of these methods require an algorithm that generates samples from non-normal multivariate distributions. In these cases, at least three parameters become relevant: (1) the marginal distributions, (2) the multivariate distribution, and (3) the correlation matrix. In contrast

to previously used algorithms, NOTAMO allows simultaneous manipulations of all three parameters. NORTA (Cario & Nelson, 1997) and VM (Vale & Maurelli, 1983) offer control over the marginal distributions and the correlation matrix. However, these two parameters fully determine the resulting multivariate distribution of both algorithms, even though other multivariate distributions that would also comply with the prespecified univariate distribution and correlation matrix exist. Moreover, as demonstrated in the first manuscript, the multivariate distribution can have different effects regarding robustness, despite similar marginals and the same correlation matrix. Copulas allow to specify the marginals and the multivariate distribution in advance, the latter via dependence parameters of the copula. However, the corresponding correlation matrix depends on both and thus cannot be prespecified. The approach by Mair et al. (2012) allows manipulations of the multivariate distribution and the correlation matrix, but the resulting marginals cannot be specified in advance. As I summarized in the introduction, the univariate distribution with skewness and kurtosis is often of main interest in robustness studies. In contrast, NOTAMO manipulates the marginals, the multivariate distribution, and the correlation matrix. This allows robustness studies to consider a wider range of data conditions and thereby empirical research a more accurate assessment of the validity of the statistical tests on which they rely.

### 3.1 Limitations

Despite the advantages presented so far, NOTAMO also has some limitations. First, we could not define conditions under which the algorithm is guaranteed to converge. If users prespecify a central moment that is more extreme than the corresponding central moment in any of the distributions that constitute the quantile mixture, no combination of said distributions can be expected to reproduce the desired central moment. However, even if the desired central moments are within the range of supplemented distributions, the non-linear root finding and non-linear optimization algorithms implemented in NOTAMO might not find an appropriate combination of inverse *CDFs*. Because of that, the choice of distributions that constitute a suitable quantile mixture is not perfectly flexible. However, the second manuscript contains



a number of inverse *CDFs* that could be used to generate distributions in a wide range regarding skewness and kurtosis. Furthermore, the algorithm also checks the accuracy of a solution and prints a warning if the desired moments could not be reproduced.

Another disadvantage of NOTAMO as well as the other non-normality methods presented in my thesis is that the algorithms use non-linear functions to create non-normal distributions. For example, VM applies lower order polynomial functions to normal random variables to obtain a non-normal distribution. Similarly, NOTAMO and NORTA utilize inverse *CDFs* to create non-normality and those inverse *CDFs* are also non-linear. Therefore, these algorithms cannot be applied to investigate the robustness of non-linear models, for example in latent growth curve models (Duncan, Duncan, & Strycker, 2006). The non-linear functions lead to a higher dependence among the random variables for the non-normal as compared to normal distributions. E.g., in the bivariate case with NORTA, two prespecified  $\chi^2$  marginals with  $df = 1$ , and correlation of  $r_X = .70$ , the underlying normal variables need to be correlated as  $r_Z = .75$  to counteract the decrease in correlation introduced by the non-linear transformation. Therefore, the (non-linear) dependence between these two variables would be higher and non-linear models would account for this dependence with non-linear functions. The resulting conditions would not be comparable for non-linear models, because they differ regarding both higher dependencies and non-normal distributions. Despite the importance of non-linear models, NOTAMO can still be used to explore the robustness of various linear models that assume normality.

Finally, NOTAMO lacks a measure of multivariate non-normality that specifies the multivariate distribution, or some aspect of it, in advance. That is, NOTAMO creates different distributions with the same marginals and underlying correlation matrix, but does so by either having normal correlated variables and non-normal uncorrelated variables, or vice versa. However, it would be beneficial to allow continuous variations of the multivariate distribution as well, for example with prespecified multivariate skewness and kurtosis.

## 3.2 Future Research Questions

First and foremost, a number of Monte Carlo studies could more thoroughly investigate the robustness of statistical tests that incorporate a normality assumption. For example, this includes widely used  $t$  tests, ANOVAs, and linear regressions, as well as SEM, linear mixed models, MANOVAs, or the treatment of missing data via full information ML (Bollen, 1989; Hox, 2010; Loveland, 2011; Tabachnick & Fidell, 2012). As I outlined in the introduction, this would be directly relevant for empirical research, because most observed variables are not normally distributed (Cain et al., in press; Micceri, 1989). For example, Blanca et al. (2013) reported that out of 693 distributions of various psychological variables, 39.9% were considered as slightly non-normal, 34.5% as moderately non-normal, 10.4% as highly non-normal, and a further 9.6% as extremely non-normal.

In conjunction with these robustness studies, the most crucial step would be to obtain a measure that can be applied to samples of a distribution and that captures the effect of non-normality on robustness. For example, the first manuscript showed that  $\alpha$  errors in SEM were severely inflated when the non-normality was based on the latent factors. In contrast, non-normality had only small or moderate effects on empirical rejection rates if the multivariate distribution was manipulated by non-normal unique errors. The goal would be to define a measure that can predict this effect, based on a sample of the underlying distribution. Kurtosis was unable to capture this difference, since kurtosis was equivalent across conditions.

One might wonder whether tail dependence could be this relevant measure of non-normality. For example, Foldnes and Grønneberg (2015) also showed that standard errors were inflated and model parameters of a SEM were biased when the non-normal distribution had tail dependence, compared to a non-normal distribution with similar multivariate kurtosis but no tail dependence. Indeed, the difference between the two types of distributions we investigated seems insofar related, as the non-normal latent factors also lead to a multivariate distribution in which the probability of an outlier in one variable increases the probability of an outlier in a second random variable. However, tail dependence seems unsuitable for two reasons. First, it is only defined for bivariate distributions. Since most multivariate analyses

that assume some sort of normality are usually used with more than two variables, a measure that considers the multivariate distributions seems more appropriate. Second, and more importantly, tail dependence is the limit of the probability of an outlier, given an outlier in the second variable, and, therefore, not defined for samples of a distribution. This is rather critical when the results of robustness studies should be used in empirical research. Investigators working with data sets as diverse as reaction times, income, neurological data, accuracies from cognitive tests, and age all need to make a decision on how to treat non-normality (Cain et al., in press; Palmer et al., 2011; Wang et al., 2008). If the measure of non-normality can only be applied to a theoretical distribution but not be inferred from a sample, these investigators would be unable to incorporate the information from Monte Carlo studies in their empirical analysis. Only a measure of non-normality that is both relevant for robustness and applicable for samples might be able to bridge the gap between simulation studies and empirical research.

## Conclusion

In my thesis, I suggested a new algorithm called NOTAMO that samples from non-normal distributions with prespecified central moments and correlation, while simultaneously manipulating the largely overlooked multivariate distribution. Through Monte Carlo simulation studies, I have shown that the multivariate distribution can have an impact on the robustness of statistical tests beyond the marginals and should be considered when observed variables are non-normal. NOTAMO can be used to further investigate the robustness of statistical tests commonly used in various fields of empirical research. Similarly, NOTAMO could be the first step in obtaining a measure based on the multivariate distribution that captures whether the degree of non-normality is relevant in statistical applications.

## References

- Asquith, W. H. (2017). *lmomco—L-moments, censored L-moments, trimmed L-moments, L-comoments, and many distributions* [Computer software manual]. Retrieved from <https://CRAN.R-project.org/package=lmomco> (R package version 2.2.7)
- Bentler, P. (2006). *EQS 6 structural equations program manual*. Encino, CA: Multivariate Software.
- Blanca, M. J., Arnau, J., López-Montiel, D., Bono, R., & Bendayan, R. (2013). Skewness and kurtosis in real data samples. *Methodology, 9*, 78–84. doi: 10.1027/1614-2241/a000057
- Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.
- Bradley, D. R., & Fleisher, C. L. (1994). Generating multivariate data from nonnormal distributions: Mihal and Barrett revisited. *Behavior Research Methods, Instruments, & Computers, 26*, 156–166. doi: 10.3758/BF03204610
- Bradley, J. V. (1973). The central limit effect for a variety of populations and the influence of population moments. *Journal of Quality Technology, 5*, 171–177.
- Braeken, J., & van Assen, M. A. (in press). An empirical Kaiser criterion. *Psychological Methods*. doi: 10.1037/met0000074
- Brown, M. B., & Forsythe, A. B. (1974). The small sample behavior of some statistics which test the equality of several means. *Technometrics, 16*, 129–132. doi: 10.2307/1267501
- Browne, M. W. (1974). Generalized least squares estimators in the analysis of covariance structures. *South African Statistical Journal, 8*, 1–24.
- Browne, M. W. (1984). Asymptotically distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology, 37*, 62–83. doi: 10.1111/j.2044-8317.1984.tb00789.x

- Browne, W. J., & Draper, D. (2000). Implementation and performance issues in the Bayesian and likelihood fitting of multilevel models. *Computational Statistics*, *15*, 391–420. doi: 10.1007/s001800000041
- Cain, M. K., Zhang, Z., & Yuan, K.-H. (in press). Univariate and multivariate skewness and kurtosis for measuring nonnormality: Prevalence, influence and estimation. *Behavior Research Methods*. doi: 10.3758/s13428-016-0814-1
- Cario, M. C., & Nelson, B. L. (1997). *Modeling and generating random vectors with arbitrary marginal distributions and correlation matrix* (Tech. Rep.). Citeseer.
- Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, *1*, 245–276. doi: 10.1207/s15327906mbr0102\_10
- Clayton, D. G. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*, *65*, 141–151. doi: 10.2307/2335289
- Clemen, R. T., & Reilly, T. (1999). Correlations and copulas for decision and risk analysis. *Management Science*, *45*, 208–224. doi: 10.1287/mnsc.45.2.208
- Clinch, J. J., & Keselman, H. J. (1982). Parametric alternatives to the analysis of variance. *Journal of Educational Statistics*, *7*, 207–214. doi: 10.2307/1164645
- Cook, R. D., & Johnson, M. E. (1981). A family of distributions for modelling non-elliptically symmetric multivariate data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 210–218.
- Curran, P. J., West, S. G., & Finch, J. F. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological Methods*, *1*, 16–29. doi: 10.1037/1082-989X.1.1.16
- DeCarlo, L. T. (1997). On the meaning and use of kurtosis. *Psychological Methods*, *2*, 292–307. doi: 10.1037/1082-989X.2.3.292
- Dinno, A. (2009). Exploring the sensitivity of Horn’s parallel analysis to the distributional form of random data. *Multivariate Behavioral Research*, *44*, 362–388. doi: 10.1080/00273170902938969
- Duncan, T. E., Duncan, S. C., & Strycker, L. A. (2006). *An introduction to latent variable growth curve modeling: Concepts, issues, and application*. New York: Psychology Press.
- Dutta, S., & Genton, M. G. (2014). A non-Gaussian multivariate distribution with

- all lower-dimensional Gaussians and related families. *Journal of Multivariate Analysis*, *132*, 82–93. doi: 10.1016/j.jmva.2014.07.007
- Fleishman, A. I. (1978). A method for simulating non-normal distributions. *Psychometrika*, *43*, 521–532. doi: 10.1007/BF02293811
- Foldnes, N., & Grønneberg, S. (2015). How general is the Vale–Maurelli simulation approach? *Psychometrika*, *80*, 1066–1083. doi: 10.1007/s11336-014-9414-0
- Foldnes, N., & Olsson, U. H. (2015). Correcting too much or too little? the performance of three chi-square corrections. *Multivariate Behavioral Research*, *50*, 533–543. doi: 10.1080/00273171.2015.1036964
- Foldnes, N., & Olsson, U. H. (2016). A simple simulation technique for nonnormal data with prespecified skewness, kurtosis, and covariance matrix. *Multivariate Behavioral Research*, 1–13. doi: 10.1080/00273171.2015.1133274
- Fouladi, R. T. (2000). Performance of modified test statistics in covariance and correlation structure analysis under conditions of multivariate nonnormality. *Structural Equation Modeling*, *7*, 356–410. doi: 10.1207/S15328007SEM0703\_2
- Garrido, L. E., Abad, F. J., & Ponsoda, V. (2013). A new look at Horn’s parallel analysis with ordinal variables. *Psychological Methods*, *18*, 454–474. doi: 10.1037/a0030005
- Gelman, A., & Hill, J. (2007). *Data analysis using regression and multi-level/hierarchical models*. Cambridge: Cambridge University Press.
- Ghosh, S., & Henderson, S. G. (2002). Chessboard distributions and random vectors with specified marginals and covariance matrix. *Operations Research*, *50*, 820–834. doi: 10.1287/opre.50.5.820.364
- Ghosh, S., & Henderson, S. G. (2003). Behavior of the NORTA method for correlated random vector generation as the dimension increases. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, *13*, 276–294.
- Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Review of Educational Research*, *42*, 237–288. doi: 10.3102/00346543042003237
- Glorfeld, L. W. (1995). An improvement on Horn’s parallel analysis methodology for selecting the correct number of factors to retain. *Educational and Psychological*

- Measurement*, 55, 377–393. doi: 10.1177/0013164495055003002
- Green, S. B., Levy, R., Thompson, M. S., Lu, M., & Lo, W.-J. (2012). A proposed solution to the problem with using completely random data to assess the number of factors with parallel analysis. *Educational and Psychological Measurement*, 72, 357–374. doi: 10.1177/0013164411422252
- Hartmann, P., Straetmans, S., & De Vries, C. G. (2004). Asset market linkages in crisis periods. *Review of Economics and Statistics*, 86, 313–326. doi: 10.1162/003465304323023831
- Harwell, M. R. (2003). Summarizing Monte Carlo results in methodological research: the single-factor, fixed-effects ANCOVA case. *Journal of Educational and Behavioral statistics*, 28, 45–70.
- Harwell, M. R., Rubinstein, E. N., Hayes, W. S., & Olds, C. C. (1992). Summarizing Monte Carlo results in methodological research: The one-and two-factor fixed effects ANOVA cases. *Journal of Educational Statistics*, 17, 315–339. doi: 10.3102/10769986017004315
- Headrick, T. C. (2002). Fast fifth-order polynomial transforms for generating univariate and multivariate nonnormal distributions. *Computational Statistics & Data Analysis*, 40, 685–711. doi: 10.1016/S0167-9473(02)00072-5
- Headrick, T. C., & Mugdadi, A. (2006). On simulating multivariate non-normal distributions from the generalized Lambda distribution. *Computational Statistics & Data Analysis*, 50, 3343–3353. doi: 10.1016/j.csda.2005.06.010
- Headrick, T. C., & Sawilowsky, S. S. (1999). Simulating correlated multivariate nonnormal distributions: Extending the Fleishman power method. *Psychometrika*, 64, 25–35. doi: 10.1007/BF02294317
- Headrick, T. C., & Sawilowsky, S. S. (2000). Weighted simplex procedures for determining boundary points and constants for the univariate and multivariate power methods. *Journal of Educational and Behavioral Statistics*, 25, 417–436.
- Henderson, S. G., Chiera, B. A., & Cooke, R. M. (2000). Generating” dependent” quasi-random numbers. In *Proceedings of the 2000 Winter Simulation Conference* (pp. 527–536). IEEE: Piscataway, N.J.
- Henze, N. (2002). Invariant tests for multivariate normality: A critical review.



- Statistical Papers*, 43, 467–506. doi: 10.1007/s00362-002-0119-6
- Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30, 179–185. doi: 10.1007/BF02289447
- Horswell, R. L., & Looney, S. W. (1992). A comparison of tests for multivariate normality that are based on measures of multivariate skewness and kurtosis. *Journal of Statistical Computation and Simulation*, 42, 21–38. doi: 10.1080/00949659208811407
- Hox, J. J. (2010). *Multilevel analysis: Techniques and applications*. New York: Routledge.
- Hu, L.-t., Bentler, P. M., & Kano, Y. (1992). Can test statistics in covariance structure analysis be trusted? *Psychological Bulletin*, 112, 351–362. doi: 10.1037/0033-2909.112.2.351
- Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the Fifth Berkeley symposium on mathematical statistics and probability* (pp. 221–233). University of California Press, Berkeley.
- Joanes, D., & Gill, C. (1998). Comparing measures of sample skewness and kurtosis. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 47, 183–189. doi: 10.1111/1467-9884.00122
- Joe, H. (1997). *Multivariate models and multivariate dependence concepts*. Boca Raton, FL: Chapman & Hall/CRC.
- Johnson, S. G. (2014). The NLOpt nonlinear-optimization package [Computer software manual]. (R package 1.0.4)
- Jöreskog, K., & Sorbom, D. (2006). *Lisrel version 8.8*. Lincolnwood, IL: Scientific Software International.
- Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, 20, 141–151. doi: 10.1177/001316446002000116
- Keselman, H., Huberty, C. J., Lix, L. M., Olejnik, S., Cribbie, R. A., Donahue, B., & Kowalchuk, R. K. (1998). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analyses. *Review of Educational Research*, 68, 350–386. doi: 10.1037/1082-989X.13.2.110

- Kolmogorov, A. N. (1933). Sulla determinazione empirica di una legge di distribuzione. *Giornale dell'Istituto Italiano degli Attuari*, *4*, 83–91.
- Koran, J., Headrick, T. C., & Kuo, T. C. (2015). Simulating univariate and multivariate nonnormal distributions through the method of percentiles. *Multivariate Behavioral Research*, *50*, 216–232. doi: 10.1080/00273171.2014.963194
- Levine, D. W., & Dunlap, W. P. (1982). Power of the F test with skewed data: Should one transform or not? *Psychological Bulletin*, *92*, 272–280. doi: 10.1037/0033-2909.92.1.272
- Lilliefors, H. W. (1967). On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association*, *62*, 399–402. doi: 10.1080/01621459.1967.10482916
- Lorenzo-Seva, U., Timmerman, M. E., & Kiers, H. A. (2011). The Hull method for selecting the number of common factors. *Multivariate Behavioral Research*, *46*, 340–364. doi: 10.1080/00273171.2011.564527
- Loveland, J. L. (2011). *Mathematical justification of introductory hypothesis tests and development of reference materials* (Tech. Rep.). Utah State University.
- Lurie, P. M., & Goldberg, M. S. (1998). An approximate method for sampling correlated random variables from partially-specified distributions. *Management Science*, *44*, 203–218. doi: 10.1287/mnsc.44.2.203
- Maas, C. J., & Hox, J. J. (2004). Robustness issues in multilevel regression analysis. *Statistica Neerlandica*, *58*, 127–137. doi: 10.1046/j.0039-0402.2003.00252.x
- Mair, P., Satorra, A., & Bentler, P. M. (2012). Generating nonnormal multivariate data using copulas: Applications to SEM. *Multivariate Behavioral Research*, *47*, 547–565. doi: 10.1080/00273171.2012.692629
- Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika*, 519–530. doi: 10.2307/2334770
- Mattson, S. (1997). How to generate non-normal data for simulation of structural equation models. *Multivariate Behavioral Research*, *32*, 355–373. doi: 10.1207/s15327906mbr3204\_3
- Meijer, E. (2000). An asymmetric distribution with zero skewness. Available at SSRN: <http://dx.doi.org/10.2139/ssrn.2531847>.
- Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures.

- Psychological Bulletin*, 105, 156–166.
- Muthén, B., & Kaplan, D. (1985). A comparison of some methodologies for the factor analysis of non-normal Likert variables. *British Journal of Mathematical and Statistical Psychology*, 38, 171–189. doi: 10.1111/j.2044-8317.1985.tb00832.x
- Muthén, B., & Kaplan, D. (1992). A comparison of some methodologies for the factor analysis of non-normal Likert variables: A note on the size of the model. *British Journal of Mathematical and Statistical Psychology*, 45, 19–30. doi: 10.1111/j.2044-8317.1992.tb00975.x
- Muthén, B., & Muthén, L. (2010). *Mplus software (version 6.1)*. Los Angeles, CA: Muthén & Muthén.
- Nadarajah, S. (2005). A generalized normal distribution. *Journal of Applied Statistics*, 32, 685–694. doi: 10.1080/02664760500079464
- Olsson, U. H., Foss, T., Troye, S. V., & Howell, R. D. (2000). The performance of ML, GLS, and WLS estimation in structural equation modeling under conditions of misspecification and nonnormality. *Structural Equation Modeling*, 7, 557–595. doi: 10.1207/S15328007SEM0704\_3
- Palmer, E. M., Horowitz, T. S., Torralba, A., & Wolfe, J. M. (2011). What are the shapes of response time distributions in visual search? *Journal of Experimental Psychology: Human Perception and Performance*, 37, 58–71. doi: 10.1037/a0020747
- Peres-Neto, P. R., Jackson, D. A., & Somers, K. M. (2005). How many principal components? Stopping rules for determining the number of non-trivial axes revisited. *Computational Statistics & Data Analysis*, 49, 974–997. doi: 10.1016/j.csda.2004.06.015
- R Core Team. (2017). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from <https://www.R-project.org/>
- Razali, N. M., & Wah, Y. B. (2011). Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. *Journal of Statistical Modeling and Analytics*, 2, 21–33.
- Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, 48, 1–36. Retrieved from

- <http://www.jstatsoft.org/v48/i02/> doi: 10.18637/jss.v048.i02
- Ruscio, J., & Kacetow, W. (2008). Simulating multivariate nonnormal data using an iterative algorithm. *Multivariate Behavioral Research*, *43*, 355–381. doi: 10.1080/00273170802285693
- Ruscio, J., & Roche, B. (2012). Determining the number of factors to retain in an exploratory factor analysis using comparison data of known factorial structure. *Psychological Assessment*, *24*, 282–292. doi: 10.1037/a0025697
- Satorra, A., & Bentler, P. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. Eye & C. C. Clogg (Eds.), *Latent variable analysis: Applications for developmental research*. Thousand Oaks, CA: Sage.
- Savalei, V. (2010). Expected versus observed information in SEM with incomplete normal and nonnormal data. *Psychological Methods*, *15*, 352–367. doi: 10.1037/a0020143
- Schmider, E., Ziegler, M., Danay, E., Beyer, L., & Bühner, M. (2010). Is it really robust? *Methodology*, *6*, 147–151. doi: 10.1027/1614-2241/a000016
- Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, *52*, 591–611. doi: 10.1093/biomet/52.3-4.591
- Sklar, M. (1959). Fonctions de répartition à  $n$  dimensions et leurs marges [Distribution functions of  $n$  dimensions and their origins]. *Publications de l'Institut de Statistique de l'Université de Paris*, *8*, 229–231.
- Smirnov, N. (1948). Table for estimating the goodness of fit of empirical distributions. *The Annals of Mathematical Statistics*, *19*, 279–281.
- Soetaert, K. (2009). rootsolve: Nonlinear root finding, equilibrium and steady-state analysis of ordinary differential equations [Computer software manual]. (R package 1.6)
- Su, P. (2014). NORTARA: Generation of multivariate data with arbitrary marginals [Computer software manual]. Retrieved from <https://CRAN.R-project.org/package=NORTARA> (R package version 1.0.0)
- Tabachnick, B. G., & Fidell, L. S. (2012). *Using multivariate statistics* (6th ed.). London: Pearson Education.

- Thurstone, L. L. (1947). *Multiple factor analysis*. Chicago: University of Chicago Press.
- Vale, C. D., & Maurelli, V. A. (1983). Simulating multivariate nonnormal distributions. *Psychometrika*, *48*, 465–471. doi: 10.1007/BF02293687
- Wang, L., Zhang, Z., McArdle, J. J., & Salthouse, T. A. (2008). Investigating ceiling effects in longitudinal data analysis. *Multivariate Behavioral Research*, *43*, 476–496. doi: 10.1080/00273170802285941
- Welch, B. (1951). On the comparison of several mean values: An alternative approach. *Biometrika*, *38*, 330–336. doi: 10.2307/2332579
- White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, *50*, 1–25. doi: 10.2307/1912526
- Wood, J. M., Tataryn, D. J., & Gorsuch, R. L. (1996). Effects of under- and overextraction on principal axis factor analysis with varimax rotation. *Psychological Methods*, *1*, 354–365. doi: 10.1037/1082-989X.1.4.354
- Yan, J. (2007). Enjoy the joy of copulas: With a package copula. *Journal of Statistical Software*, *21*, 1–21. doi: 10.1.1.610.7783
- Yanagihara, H., & Yuan, K.-H. (2005). Four improved statistics for contrasting means by correcting skewness and kurtosis. *British Journal of Mathematical and Statistical Psychology*, *58*, 209–237. doi: 10.1348/000711005X64060
- Zwick, W. R., & Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin*, *99*, 432–442. doi: 10.1037/0033-2909.99.3.432

## Co-author statement

I hereby confirm that the following articles were primarily conceived and written by Dipl.-Psych. Max Auerswald, School of Social Sciences, University of Mannheim.

Auerswald, M., & Moshagen, M. (2015). Generating correlated, non-normally distributed data using a non-linear structural model. *Psychometrika*, 80, 920-937.

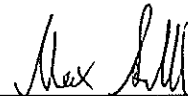
Auerswald, M., & Moshagen, M. (2017). *Sampling from arbitrary non-normal distributions with given covariance and central moments*. Manuscript submitted for publication.

Auerswald, M., & Moshagen, M. (2017). *How to determine the number of factors to retain in exploratory factor analysis? A comparison of extraction methods under realistic conditions*. Manuscript submitted for publication.

89069 UIm  
Universität UIm  
Institut für Psychologie und Pädagogik  
Psychologische Forschungsmethoden  
Prof. Dr. Morten Moshagen  
May 2017, Prof. Dr. Morten Moshagen

## Statement of Originality

I hereby declare that I am the sole author of this thesis and have used no other sources than those cited in this work.

A handwritten signature in black ink, appearing to read 'Max Auerswald', is positioned above a horizontal line.

Dipl.-Psych. Max Auerswald

Kassel, 03.06.2017

## GENERATING CORRELATED, NON-NORMALLY DISTRIBUTED DATA USING A NON-LINEAR STRUCTURAL MODEL

MAX AUERSWALD

UNIVERSITY OF MANNHEIM

UNIVERSITY OF KASSEL

MORTEN MOSHAGEN

UNIVERSITY OF KASSEL

An approach to generate non-normality in multivariate data based on a structural model with normally distributed latent variables is presented. The key idea is to create non-normality in the manifest variables by applying non-linear linking functions to the latent part, the error part, or both. The algorithm corrects the covariance matrix for the applied function by approximating the deviance using an approximated normal variable. We show that the root mean square error (RMSE) for the covariance matrix converges to zero as sample size increases and closely approximates the RMSE as obtained when generating normally distributed variables. Our algorithm creates non-normality affecting every moment, is computationally undemanding, easy to apply, and particularly useful for simulation studies in structural equation modeling.

Key words: Non-normal multivariate data, Structural equation modeling, Simulation.

Monte Carlo simulations are an important tool in determining the robustness and the validity of statistical methods. A crucial step in Monte Carlo studies is the creation of data that violate certain assumptions of the statistical test; the most prominent one being the assumption of normally distributed data. In the univariate case, creating data following a specific distribution is straightforward. The multivariate case is more challenging since the generated data are usually required to follow a prespecified covariance matrix. Any non-normalization process potentially distorts the covariation and needs to be counteracted.

A number of approaches directed towards this issue have been proposed (e.g., Bradley & Fleisher, 1994; Burr, 1942; Cario & Nelson, 1998; Cook & Johnson, 1981; Headrick, 2002; Headrick & Mugdadi, 2006; Johnson, 1949; Mair, Satorra, & Bentler, 2012; Mattson, 1997; Nagahara, 2004; Ramberg & Schmeiser, 1974; Ruscio & Kaczetow, 2008; Tadikamalla, 1980) with the power constant approach using polynomial transformations being the most popular one (Vale & Maurelli, 1983). This approach relies on a technique suggested by Fleishman (1978), which determines a non-normally distributed variable  $X$  by

$$X = a + bZ + cZ^2 + dZ^3, \quad (1)$$

where  $Z$  is a standard normally distributed random variable and the power constants  $a$ ,  $b$ ,  $c$ , and  $d$  are obtained by solving a system of four equations provided by Fleishman (1978) for the first four moments. Vale and Maurelli (1983) extended this approach for multivariate data using matrix

**Electronic supplementary material** The online version of this article (doi:10.1007/s11336-015-9468-7) contains supplementary material, which is available to authorized users.

Correspondence should be made to Max Auerswald, Institute of Psychology, University of Kassel, Holländische Straße 36-38, 34127 Kassel, Germany. Email: auerswald@uni-kassel.de



decomposition. Their procedure comprises two main steps: First, an intermediate correlation matrix anticipating the effect of non-normalization is calculated. Second, normally distributed variables are generated through matrix decomposition according to the intermediate correlation matrix and are finally non-normalized. The non-normally distributed variables have the desired covariances in the population distribution. The algorithm was further extended by the fifth-order polynomial allowing for the specification of the first six moments (Headrick, 2002; Headrick & Sawilowsky, 1999) and using other distributions instead of normally distributed variable  $Z$ , e.g., uniform or triangular distributions (Hodis, Headrick, & Sheng, 2012).

The power constants approach contributed significantly to the field and is still very popular. However, it also has some drawbacks. Defining any finite number of central moments is not sufficient to define a distribution (Ruscio & Kaczetow, 2008). Changes in higher moments lead to an infinite number of distributions with the same finite amount of first moments, and even infinite moments may not result in a unique probability distribution (Devroye, 1986, p. 684). As the power constants approach only allows for specifying the first four moments (or six using recent extensions), it is not possible to generate certain distribution families, such as the  $\chi^2$  or the lognormal distribution. Moreover, the power constants approach has an additional boundary condition if it is desired to specify a valid probability density function (PDF) or cumulative density function (CDF) for the generated data (Headrick & Kowalchuk, 2007). For  $\mu = 0$ ,  $\sigma = 1$ , and normally distributed variable  $Z$ , the boundary for the fourth moment is then  $3 < \mu_4 < 46.2$  for symmetric distributions. With increasing asymmetry in the distribution, the boundary of the fourth moment becomes gradually narrower as

$$\frac{\mu_3^2}{\mu_4 - 3} < \frac{9}{14} \quad (2)$$

needs to hold (Headrick & Kowalchuk, 2007). Both boundaries can be extended using higher order polynomials (Headrick, 2002), but it is still not possible to generate arbitrary marginal distributions. In addition, higher order polynomials may also lead to non-unique solutions, such that more than one set of power constants fits to a single set of first six moments (Headrick & Kowalchuk, 2007). This is related to the so-called ‘classical problem of moments’, i.e., finding a unique distribution given a set of moments (Devroye, 1986; Headrick, 2010, p. 26). Note that for fourth-order polynomials, the choice of power constants within the class of distributions following Equation 2 will always be unique, if it exists (Devroye, 1986, p. 685).

The problem of generating different distributions becomes more severe when considering the obtained multivariate distribution. Foldnes and Grønneberg (in press) examined the tail dependence of a generalized Vale and Maurelli method. Briefly, tail dependence is a measure of bivariate dependency based on the probability of drawing an extreme value from one distribution, conditional on having an extreme value in the other (Joe, 1997). They showed that the Vale and Maurelli procedure (based on normally distributed variables) has no tail dependence, a property the transformed variables share with the multivariate normal distribution. Furthermore, choosing a distribution with non-zero tail dependence led to higher bias and standard errors in the estimation of a population covariance in a simple confirmatory factor model, compared to samples simulated with the Vale and Maurelli procedure. Foldnes and Grønneberg (in press) concluded “that the truly multivariate aspects of data generation using the VM [Vale and Maurelli] approach is exactly equal to the Normal model” (Conclusion section, para. 2).

As an alternative approach, Yuan and Bentler (1999) suggested to generate correlated, non-normal variables as the product of two random variables with prespecified skewness and kurtosis. Let  $\Sigma$  be the desired covariance matrix of dimension  $p$  and  $I_p$  the  $p$ -dimensional identity matrix.

For a non-negative random variable  $r$  and  $Z \sim N(0, I_p)$ , independent of  $r$ ,

$$X = r\Sigma^{\frac{1}{2}}Z, \quad (3)$$

defines an elliptical distribution (Fang, Kotz, & Ng, 1990). Yuan and Bentler extended this procedure to

$$X = rAZ, \quad (4)$$

with  $AA' = \Sigma$ , allowing for non-symmetrical distribution  $Z$  and unrestricted distributions of  $r$ , which in general results in non-elliptical distribution  $X$ . If the resulting marginal distributions have the same kurtosis, the obtained distributions are said to be pseudo-elliptical. If the kurtosis is equal to that of a normal distribution, distributions are called pseudo-normal. While those types of distributions are defined by this specific generation scheme, Yuan and Bentler could examine asymptotic robustness properties of various test statistics used in structural equation modeling (SEM). However, it is unclear whether observed data usually fall into this family of distributions, albeit larger than the family of elliptical distribution.

A third approach not relying on moments and without univariate distributional constraints was proposed by [Ruscio and Kaczetow \(2008\)](#). This procedure starts generating correlated normal data following a given covariance matrix and uncorrelated non-normal data following a given distribution. The normally distributed data are then replaced by the latter while sorting both distributions by each variable. Consequently, the rank order for all individuals and each variable stays the same, thereby ensuring a correlation in the simulated data. In an iterative process, the covariance matrix for the normally distributed variables is modified to minimize the difference between the target and simulated covariance matrix. The approach allows for a high flexibility in marginal distributions not limited to the first four or six moments and therefore addresses one of the main objections raised against the power constants method. However, the minimization idea has two consequences: First, the sampling fluctuation is too small for small sample sizes, as achieving the minimum difference in sample and target covariance is unrealistic for a natural sampling procedure. For example, if two normally distributed variables correlate to  $\rho = .2$  in the population, it is unlikely to obtain a correlation very close to  $.2$  in every random sample with  $n = 50$ . Second, for a given set of distributions, the algorithm only converges to the minimum difference and not necessarily to zero. Moreover, the algorithm is computationally demanding for larger sample sizes due to the sorting embedded in a trial-and-error process that needs to be performed for every generated data set.

The purpose of the present article is to introduce an algorithm mainly relying on functions instead of moments, which allows for high flexibility in the data generation process as well as specifying gradual deviations from a normal distribution. In addition, by avoiding a trial-and-error process, the algorithm is computationally undemanding, which is important for robustness studies involving large sample sizes. The algorithm is based on a latent structural model, so that the distribution of a manifest variable is determined by the sum of two random variables (latent and error distributions), in turn making the algorithm especially suited for robustness studies in SEM.

## 1. Algorithm Description

The goal is to simulate correlated non-normally distributed manifest variables potentially deviating in more than the first six moments from a normal distribution. The basic idea is to apply arbitrary (non-linear) linking functions to normally distributed latent and/or error variables and

to correct for the discrepancy in the covariance matrix caused by the transformation. The discrepancy is assessed using estimates of a normally distributed variable and the covariance matrix of every linking function applied to this variable. Given a desired covariance matrix (expressed as a structural equation model) and a set of linking functions, the algorithm first estimates the deviation due to the non-normality transformation using the estimate of a normal variable. The loadings in the model are then corrected for this deviation in both variance and covariance. Finally, the remaining deviance among the manifest variables is estimated and the error terms are correlated to counteract this deviance. We first describe the data generating process using a priori specified linking function. Thereafter, we show how to control for the degree of non-normality (for example in terms of a specific moment) by systematically varying the linking functions.

### 1.1. Data Generation Using Prespecified Linking Functions

The starting point of the algorithm is to define the desired correlation matrix among the to-be generated variables using a structural equation model with latent and manifest variables as well as error terms. The desired manifest covariance matrix,  $\Sigma_T$ , is thus a function of the number of manifest and latent variables, the (standardized) factor loadings of the manifest variables, and the correlations among the latent variables and the error terms. Let  $L = (L_1, \dots, L_m)$  be the set of latent variables with covariance matrix  $\Sigma_L$  and  $M = (M_1, \dots, M_n)$  the manifest variables with  $\xi = (\xi_1, \dots, \xi_n)$  normally distributed error terms, correlated according to  $\Sigma_\xi$ . The error terms are later used to correct for remaining deviances due to the non-normalization process which changes the underlying error covariance matrix to  $\Sigma_E$ . Note that this correction only addresses the effects of non-normalization, but retains the effects of the correlations among the errors on the manifest variables. The loading of the  $j$ -th manifest variable on the  $i$ -th latent variable is denoted as  $k_j$ . For simplicity, we require that any manifest variable only loads on a single latent variable (later in this article we show how to implement more complex loading structures). Generation of  $L_i$  given  $\Sigma_L$  proceeds by any suitable matrix decomposition such as Cholesky decomposition. Without loss of generality, we assume  $\mu(M_j) = 0$  and  $\sigma(M_j) = 1$ . The manifest variable  $M_j$  is then given by

$$M_j = b_j g_j(L_i) + c_j h_j(\xi_j), \quad (5)$$

where  $g_j$  and  $h_j$  are arbitrary linking functions for the latent and error variables, respectively. The distribution of  $M_j$  thus depends on the applied linking functions.  $M_j$  is normally distributed when both  $g_j$  and  $h_j$  are linear, while non-linear functions  $g_j$  or  $h_j$  result in a non-normal distribution for  $M_j$ . The scalars  $b_j$  and  $c_j$  are required to correct for this transformation. Every  $b_j$  needs to be determined such that

$$r(L_i, M_j) = k_j \quad (6)$$

holds, where  $r$  is the Pearson product-moment correlation.

To calculate  $b_j$ , the procedure applies the inverse standard normal cumulative distribution function to an accuracy vector of  $p$  values to create a variable  $Z$ . This accuracy vector  $P$  contains values starting at  $10^{-a}$  up to  $(1 - 10^{-a})$  in steps of  $10^{-a}$  for every component of the vector and  $a \in \mathbb{N}$ . With increasing  $a$ ,  $Z$  more closely approximates the characteristics of a standard normal distribution due to a more refined vector of  $p$  values.  $Z$  serves as an estimate of a normally distributed variable, so

$$r(Z, g(Z)) = \hat{r}(L_i, g(L_i)) \quad (7)$$

and

$$\sigma(g(Z)) = \hat{\sigma}(g(L_i)). \quad (8)$$

The algorithm requires that

$$|r(Z, g(Z))| \geq |k_j|. \quad (9)$$

If a transformation function reduces the correlation of  $Z$  and  $g(Z)$  below the desired loading,  $|r(L_i, M_j)|$  would also be reduced below  $|k_j|$ . Given the standardized loadings, Equation 9 and  $k_j = r(g(L_i), M_j)$  ensure that  $|b_j|$  is smaller than one. For monotone functions and typically used loadings, the restriction holds, as demonstrated in the first simulation study below. Every  $b_j$  is calculated as a correction factor for the deviation in the standard deviation and the loadings as

$$b_j = \frac{k_j}{r(Z, g_j(Z))} \frac{1}{\sigma(g_j(Z))}. \quad (10)$$

The same logic applies for  $c_j$ , which is calculated as

$$c_j = \left( 1 - \frac{k_j^2}{r(Z, g_j(Z))^2} \right)^{1/2} \frac{1}{\sigma(h_j(Z))}. \quad (11)$$

The CDF of the manifest variables,  $F(M_j)$ , follows the distribution of the sums of two random variables, the transformed latent variable, and the transformed error. For the latent part (the CDF of  $b_j g_j(L_i)$ ), the support  $A_j$  of  $L_i$  is partitioned into disjunctive, convex subsets  $A_j = A_{j1} \cup \dots \cup A_{jp} \cup A_{j(p+1)} \cup \dots \cup A_{jq}$ , where the function  $g_j$  is strictly monotone on  $A_{j1}, \dots, A_{jp}$ , and subsets  $A_{j(p+1)}, \dots, A_{jq}$ , where  $g_j$  is constant. Using the distribution function method, the CDF of the latent part is

$$\begin{aligned} F_y(y) &= P(b_j \cdot g_j(L_i) \leq y) \\ &= \sum_{k=1}^p F_{L_i} \left( g_j^{-1} \left( \frac{y}{b_j} \right) \right) - F_{L_i}(\inf(A_{jk})) \mid \left( \frac{y}{b_j} \in g_j|_{A_{jk}}(A_{jk}) \right) \\ &\quad + \sum_{k=1}^p (F_{L_i}(\sup(A_{jk})) - F_{L_i}(\inf(A_{jk}))) \mid \left( \frac{y}{b_j} \geq g_j|_{A_{jk}}(\sup(A_{jk})) \right) \\ &\quad + \sum_{l=p+1}^q (F_{L_i}(\sup(A_{jl})) - F_{L_i}(\inf(A_{jl}))) \mid \left( \frac{y}{b_j} \geq g_j|_{A_{jl}}(L_i) \right), \end{aligned} \quad (12)$$

where  $L_i$  is distributed standard normal. The error part is the CDF of  $c_j h_j(\xi_j)$ . Following the same logic, the support  $B_j$  of  $\xi_j$  is split into subsets  $B_{j1}, \dots, B_{jr}$ , where  $h_j$  is strictly monotone on every subset and subsets  $B_{j(r+1)}, \dots, B_{js}$ , where  $h_j$  is constant. The CDF for the error part is

$$\begin{aligned} F_y(y) &= P(c_j \cdot h_j(\xi_j) \leq y) \\ &= \sum_{k=1}^r F_{\xi_j} \left( h_j^{-1} \left( \frac{y}{c_j} \right) \right) - F_{\xi_j}(\inf(B_{jk})) \mid \left( \frac{y}{c_j} \in h_j|_{B_{jk}}(B_{jk}) \right) \\ &\quad + \sum_{k=1}^r (F_{\xi_j}(\sup(B_{jk})) - F_{\xi_j}(\inf(B_{jk}))) \mid \left( \frac{y}{c_j} \geq h_j|_{B_{jk}}(\sup(B_{jk})) \right) \\ &\quad + \sum_{l=r+1}^s (F_{\xi_j}(\sup(B_{jl})) - F_{\xi_j}(\inf(B_{jl}))) \mid \left( \frac{y}{c_j} \geq h_j|_{B_{jl}}(\xi_j) \right) \end{aligned} \quad (13)$$

At this point, the latent model is correctly specified. However, the covariance matrix of the manifest variables still deviates from the desired covariance matrix, since the adjustments by  $b_j$  and  $c_j$  only correct the correlation to the latent variables. The remaining deviance depends on the similarity of the distributions for two manifest variables. The covariance among them may either be too low for dissimilar distributions or too high for similar distributions due to the correction process for  $b_j$  and  $c_j$ . For example, if two variables share the same linking function, the correlation among them exceeds the correlation given by  $\Sigma_T$ , because both loadings increase due to the correction process. In order to counteract this deviation, the error terms are correlated accordingly.

The algorithm uses the covariance of the variables  $b_j g_j(Z)$  to estimate the correlations between different functions. Let  $\mathbf{D}$  be the deviation matrix with entries  $d_{j_1 j_2}$  for different manifest variables  $M_{j_1}$ ,  $M_{j_2}$  and corresponding latent variables  $L_{i_1}$ ,  $L_{i_2}$ . For ease of notation, we define

$$M'_j := b_j g_j(Z) \quad (14)$$

as the vector of the estimated transformed variables. Then,  $M'_j(k)$  is the  $k$ -th element of the vector and  $P(k)$  is the  $k$ -th element of the vector of  $p$  values. The estimated deviation is

$$d_{j_1 j_2} = (\Sigma_T)_{j_1 j_2} - \begin{cases} \text{cov}(M'_{j_1}, M'_{j_2}) & \text{if } i_1 = i_2 \\ \text{cov}_{\text{adj}}(M'_{j_1}, M'_{j_2}) & \text{if } i_1 \neq i_2. \end{cases} \quad (15)$$

The case  $i_1 = i_2$  applies if both functions refer to manifest variables with loadings on the same latent variable. Otherwise, the covariance is adjusted for the correlation of the latent variables (denoted as  $r_L$ ) by

$$\text{cov}_{\text{adj}}(M'_{j_1}, M'_{j_2}) = \sum_{k=1}^{10^a-2} \sum_{k'=1}^{10^a-2} M'_{j_1}(k) M'_{j_2}(k') \int_{P(k)}^{P(k+1)} \int_{P(k')}^{P(k'+1)} \phi_{r_L}(v, w) \, dv \, dw. \quad (16)$$

The estimation uses the bivariate density  $\phi_{i_1, i_2}$  of two normal variables correlated according to the prespecified correlation of  $L_{i_1}$  and  $L_{i_2}$  (Cario & Nelson, 1998). In this case, every pair of characteristic elements in  $M'_{j_1}$ ,  $M'_{j_2}$  is weighted by the corresponding probability of  $\phi$ , multiplied, and summed. If error correlations are prespecified, the deviance  $d_{j_1 j_2}$  is adjusted by the respective value of  $(\Sigma_{\xi})_{j_1 j_2}$ . In order to improve the performance of the algorithm, a smaller accuracy vector is typically used for this estimation.

The diagonal of  $\mathbf{D}$  contains the variances of the transformed corrected errors. However, since the errors are also transformed by a linking function, correlating the errors according to the off-diagonal elements of  $\mathbf{D}$  would still result in an incorrect covariance matrix. The discrepancy due to the error functions is estimated using the covariance of  $c_j h_j(Z)$ . In line with the notation for  $M'$ , we define  $E'_j = c_j h_j(Z)$  and  $r_E$  as the required correlation of error variables. Then, the correlation of two error variables needs to be set such that

$$-d_{j_1 j_2} = \sum_{k=1}^{10^a-2} \sum_{k'=1}^{10^a-2} E'_{j_1}(k) E'_{j_2}(k') \int_{P(k)}^{P(k+1)} \int_{P(k')}^{P(k'+1)} \phi_{r_E}(v, w) \, dv \, dw. \quad (17)$$

In this case, the correlation of the bivariate normal variables  $r_E$  is unknown. However, the resulting error correlation matrix  $\Sigma_E$  needs to be positive definite, so  $-1 < r_E < 1$  holds. Any preimplemented general equation solving routine can be employed and the solution is approximated

computationally fast. Most importantly, the estimation time is independent of desired sample size or number of samples.

Finally, the error variables are generated according to  $\Sigma_E$  by matrix decomposition. The correction requires the existence of a positive definite, symmetric matrix  $\Sigma_E$ . Certain (rather extreme) conditions involving the combination of (1) a large number of manifest variables with high loadings on a single latent variable, (2) severe deviations from normality, and (3) highly different distributions of manifest variables loading on the same latent variable may result in a non-positive definite matrix  $\Sigma_E$ . However, as can be seen in the simulation example presented below, under conditions typically encountered in SEM,  $\Sigma_E$  is positive definite and can be decomposed.

In summary, the algorithm consists of the following seven steps:

1. Apply the inverse standard normal cumulative distribution function to the accuracy vector of  $p$  values to obtain an estimate of a standard normal distribution  $Z$ .
2. Apply the linking functions to the approximation of a normal distribution to estimate the deviation of covariance between manifest and latent variables.
3. Change the loadings and links to error variables by the estimated values in step (2). At this point, the correlations to the latent model and the variances are correctly specified. However, there is remaining deviance due to the degree of similarity of the distributions of the manifest variables that needs to be counteracted using the covariance of the error variables.
4. Use the variable from step (1) and the correction factors from step (2) to estimate the residual covariance among the manifest variables.
5. Estimate the error covariance  $\Sigma_E$  that counteracts the deviation given their decrease in correlation due to  $h_j$  and the prespecified error covariance  $\Sigma_\xi$ .
6. Generate standard normal variables according to  $\Sigma_L$  and standard normal errors according to  $\Sigma_E$ .
7. Apply the respective linking functions to the generated variables in step (6) and add them, weighted by the correction factors in step (2).

### 1.2. Determining Linking Functions to Control the Degree of Non-normality

An important requirement for simulation algorithms in the context of robustness studies is the ability to control the degree of non-normality. Only if a wide range of normality violations is covered, reliable conclusions regarding the intervals in which a statistical method is robust can be drawn. In the algorithm description above, the linking functions that control the degree of non-normality have been treated as input arguments of the algorithm. In this section, a method to choose and systematically vary the linking functions is introduced.

The general idea of this continuous variation stems from the notion that (a) any non-linear function results in non-normally distributed data and (b) any linear function results in normally distributed data. Assume  $g$  is a non-linear linking function. Let  $id_x$  be the identity function and  $\alpha \in [0, 1]$ . Then

$$g_\alpha := \alpha g + (1 - \alpha)id_x, \tag{18}$$

is the weighted sum of the non-linear function  $g$  and the (linear) identity function. If  $\alpha = 0$ ,  $g_\alpha$  is linear. By increasing the value for  $\alpha$ ,  $g_\alpha$  becomes increasingly non-linear, up to the point where  $\alpha = 1$ , such that  $g_\alpha = g$ . This general process is useful for the transition of any non-normal distribution to a normal distribution (as demonstrated in the second simulation study below). Since many robustness studies rely on moments as a proxy for the degree of non-normality, a linking function search that matches a prespecified moment can also be employed. This problem can be solved either analytically or numerically. To obtain an analytic solution, the moment generating function (MGF) of the manifest variable  $M_j$  can be solved for a parameter (e.g.,  $\alpha$ ) in the respective

linking functions. Since the MGF of the sum of two random variables is the sum of the MGFs of the respective variables, the MGF of the manifest variable is determined by the known CDFs of the latent and error part. The resulting MGF's  $k$ -th derivative yields the solution for the  $k$ -th moment and  $\alpha$ . However, as of now, no closed form solution exists for general functions  $g$ .

Alternatively,  $\alpha$  can be estimated numerically, provided that the desired  $k$ -th moment is on an interval with boundaries of the  $k$ -th moment of a normal distribution and the  $k$ -th moment of the resulting variable generated by the linking functions approach with linking function  $g$ . Let the  $k$ -th moment of the desired random variable  $M$  be  $\mu_k(M)$ ,  $N$  a normally distributed random variable, and  $G$  the resulting random variable using the algorithm and linking function  $g$ , estimated by  $g(Z)$ . Provided that

$$\mu_k(M) \in [\min(\mu_k(N), \mu_k(G)), \max(\mu_k(N), \mu_k(G))] \quad (19)$$

holds, an implementation of the bisection method can find the parameter  $\alpha$  for  $g_\alpha$  by bisecting the  $[0, 1]$  interval and iterating over the interval that still contains the corresponding  $k$ -th moment.

In every iteration, the algorithm determines the moment of the resulting variable, which is the sum of transformed latent and error variables, by calculating the sum of the involved cumulants and transforming the result back to moments. Note that this statement also holds for different linking functions for latent and error variables. As long as the desired moment is on the interval for  $\alpha \in [0, 1]$ , the algorithm can determine  $\alpha$  such that the resulting variable has the desired moment. As illustrated below, depending on the chosen linking function and depending on whether it transforms the latent or the error part, the algorithm creates different distributions, all incorporating the same moment.

## 2. Simulation Studies

We conducted two simulation studies to evaluate the performance of the proposed linking functions approach. In the first simulation study, we comparatively evaluated the root mean square error (RMSE) of various approaches of generating non-normal data. In the second simulation study, we examined how well the proposed linking functions approach approximates prespecified central moments of the distributions. In addition, we also investigated the performance of test statistics in SEM using non-normal data generated by either the linking functions approach or the Vale–Maurelli approach.

### 2.1. Simulation Study 1

The aims of the first study were to determine the RMSE of the correlation matrices and to obtain an impression of the generated univariate and bivariate distributions. The structural model used in this study comprised four latent variables measured by three manifest indicators each (see Figure 1). All standardized loadings were .7 and the covariance matrix among the latent variables was

$$\Sigma_L = \begin{pmatrix} 1 & & & \\ 0.3 & 1 & & \\ 0.4 & 0.1 & 1 & \\ -0.2 & -0.1 & 0.2 & 1 \end{pmatrix}.$$

This setup thus implied a desired target correlation matrix  $\Sigma_T$  among the manifest variables of





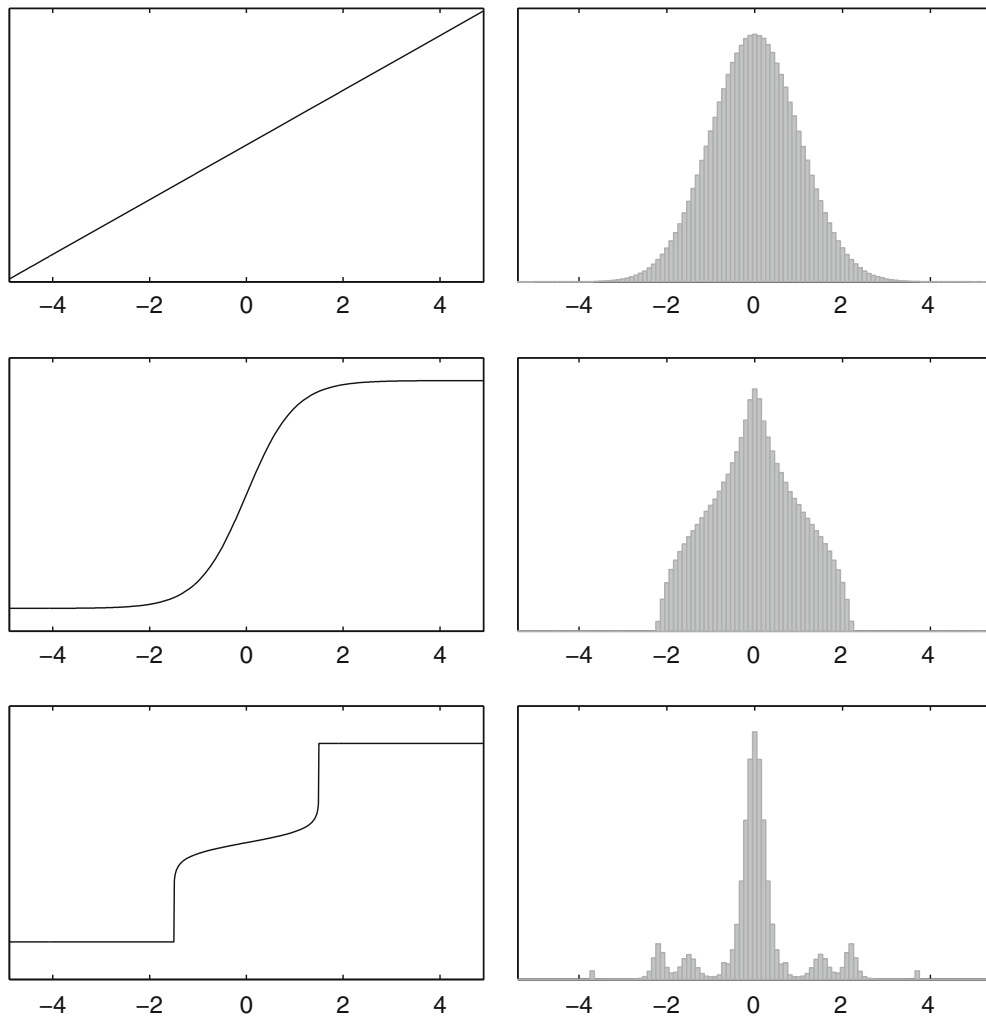


FIGURE 2.

The *left panels* show the linking functions  $g_1, g_2, g_3$  (Equations 20–22), the *right panels* show the histograms of the manifest variables using the function on the left to transform both latent and error variables.

Note that the linear function  $g_1$  yields normally distributed  $M_j$ , whereas the remaining linking functions are associated with different degrees of non-normality. Every linking function was used twice for the 12 loadings and twice for the respective errors (see Figure 1), so in this case  $g_j = h_j$  (an additional simulation study not reported here with linking functions  $g_j \neq h_j$  yielded similar results with regard to the RMSE). We used an accuracy value of  $a = 7$  for estimating the loading and variance correction factors, as this captures the extreme values of a normal distribution while still being computed comparatively fast. The convergence behavior of the algorithm was checked with eight independent sample sizes of  $n = (50; 100; 250; 500; 1000; 10,000; 100,000)$ . We used 2000 replications. The algorithm was implemented in the MATLAB computing language. The scripts, along with a detailed simulation example, are provided in the supplementary online material.

We compared the convergence behavior of the algorithm with two other approaches to generate non-normal data. In the product-based approach by [Yuan and Bentler \(1999\)](#), non-normal

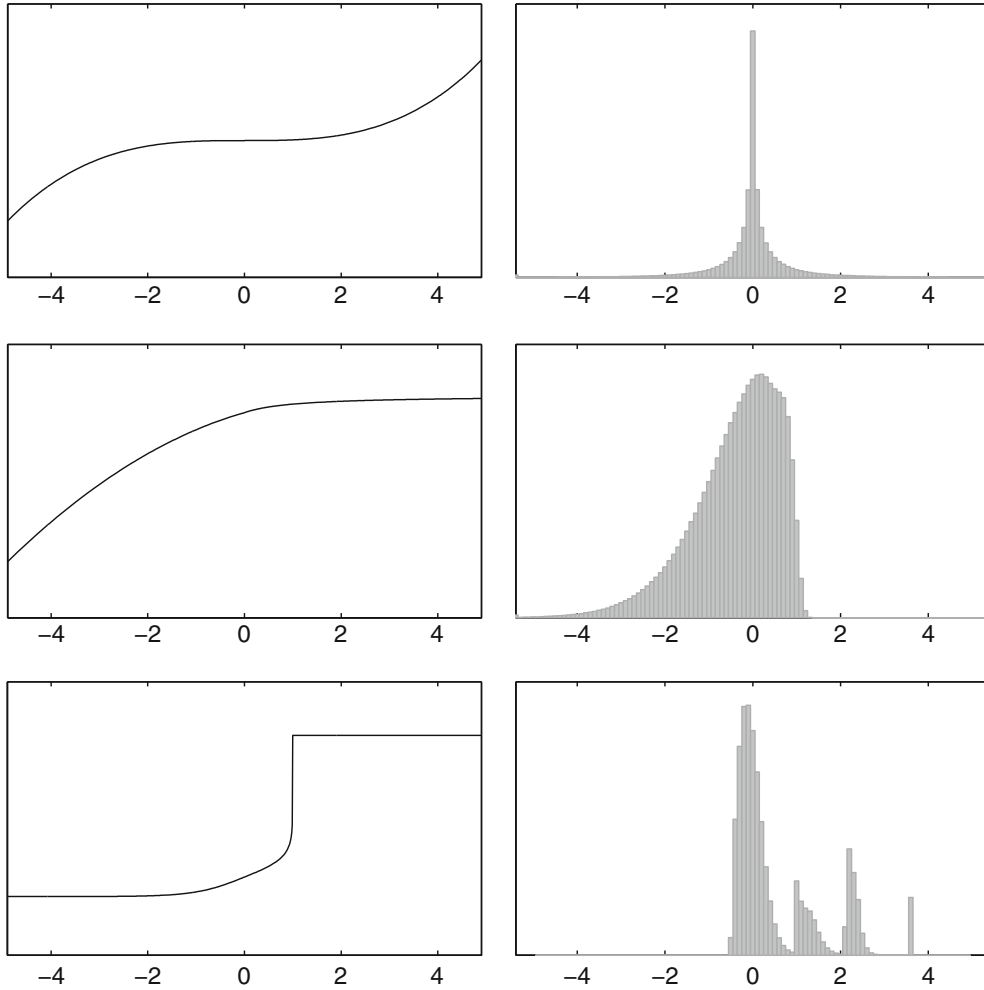


FIGURE 3.

The *left panels* show the linking functions  $g_4, g_5, g_6$  (Equations 23–25), the *right panels* show the histograms of the manifest variables using the function on the left to transform both latent and error variables.

variables are specified as the product of two random variables. The distributions generated by the linking functions, however, are always the sum of two random variables, emulating a structural equation model. Except for special cases, the distributions therefore differ by design. In Equation 3, we chose a standardized central  $\chi^2$  distribution with  $df = 1$  for both  $Z$  and  $r$ . Using the transpose of the Cholesky decomposition of  $\Sigma_T$ , the setup results in a pseudo-elliptical distribution with covariance matrix  $\Sigma_T$  as specified above. Unlike the product-based approach, the procedure by [Ruscio and Kacetow \(2008\)](#) allows to specify exactly the same distributions  $g_1$ – $g_6$  as used for our algorithm. However, the method by Ruscio and Kacetow did not converge for 12 variables and 4 latent variables. We generated data only for 3 variables, corresponding to  $g_1, g_2,$  and  $g_3$ , to give a comparison in the cases where the method converges.

Figure 4 shows the RMSE of the correlation matrices as a function of sample size for the considered approaches. For comparison purposes, the RMSE resulting from the generation of normally distributed data (according to  $\Sigma_T$ ) by means of Cholesky decomposition is also shown. Ideally, the RMSE should converge to zero for a method while having sampling errors comparable

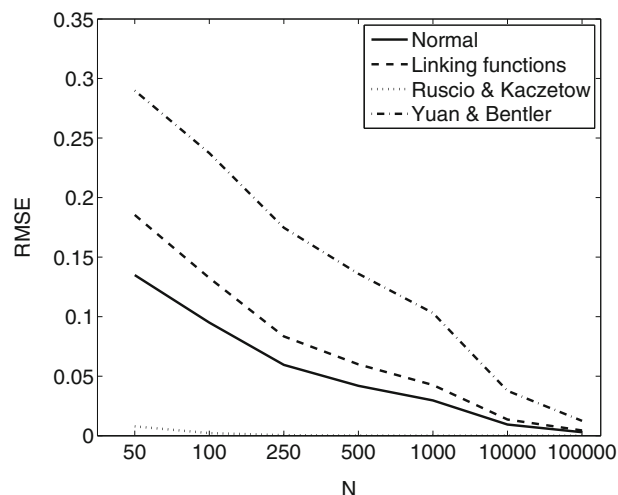


FIGURE 4.

Root mean square errors of the manifest correlation matrices as a function of sample size. The *solid line* depicts the RMSE using normally distributed variables and Cholesky decomposition. The *dashed line* depicts the RMSE using the linking function approach with functions  $g_1$ – $g_6$  (Equations 20–25). The *dotted line* depicts the method by Ruscio and Kaczetow (2008) using the distributions resulting from  $g_1$ ,  $g_2$ , and  $g_3$ . The *dash-dotted line* depicts a pseudo-elliptical distribution using the method by Yuan and Bentler (1999).

to normal distribution sampling. The RMSE indicates convergence of the simulated correlation matrix to the target correlation matrix for all three algorithms. However, the sampling errors for the linking functions approach match the ones of the normal distribution sampling best. While the method by Ruscio and Kaczetow has sampling errors close to zero regardless of sample size, the method by Yuan and Bentler suffers comparatively large sampling errors. In contrast, the linking functions approach is able to reproduce a target correlation matrix with an appropriately high degree of accuracy even with small sample sizes.

We also ran a simulation with  $n = 5,000,000$  to get a clearer picture of the generated distributions and their central moments. The six histograms for the different functions are depicted in the right panels of Figures 2 and 3. The first eight central moments of these six distributions are presented in Table 1. The moments show that the applied linking functions lead to high distributional diversity. Also note that the function  $g_6$  yields a distribution exceeding the boundary conditions of the Vale and Maurelli procedure for normally distributed variable  $Z$  (Headrick & Sawilowsky, 1999). Function  $g_2$  can be generated by the Vale and Maurelli procedure (Headrick & Kowalchuk, 2007), but has an unknown PDF (see Equation 2).

## 2.2. Simulation Study 2

The first goal of the second simulation study was to examine the performance of the linking functions approach regarding the approximation of prespecified central moments of the univariate distributions. Given the thereby generated data exhibiting a prespecified level of the fourth moment, the second purpose was to examine the behavior of test statistics in SEM under such conditions.

This simulation was based on a structural model with two latent variables, correlated with  $r = .3$ . Each latent variable was measured by five manifest indicators with loadings .7, .6, .5, .4, and .3. We specified the fourth moment for all univariate distributions as  $\mu_4 = 15$  and used the following linking functions as a starting point for the algorithm:

TABLE 1.  
Central moments of the six functions from the first simulation example.

Central moment	$g_1$	$g_2$	$g_3$	$g_4$	$g_5$	$g_6$
$\mu_1$	0.00	0.00	0.00	0.00	0.00	0.00
$\mu_2$	1.00	1.00	1.00	1.00	1.00	1.00
$\mu_3$	0.00	0.00	0.00	0.02	-1.07	1.30
$\mu_4$	3.00	2.31	5.06	33.11	4.48	3.67
$\mu_5$	-0.01	0.00	-0.03	-1.77	-13.18	8.81
$\mu_6$	15.07	6.86	37.84	5033.70	53.61	24.22
$\mu_7$	-0.08	0.02	-0.36	-6298.23	-230.82	67.89
$\mu_8$	105.26	22.97	379.87	1,694,352.40	1120.82	197.17

$\mu_n = n$ -th central moment of the resulting manifest variable, applying the functions  $g_j$  and  $h_j$  for latent and error variable (see Equations 20–25). In the first simulation example,  $g_j = h_j$ .

TABLE 2.  
Average fourth central moment of every distribution in the second simulation example.

Design	$h_1$	$h_2$	$h_3$	$h_4$	$h_5$
Non-normal latent	14.19	14.39	14.45	13.72	15.07
Non-normal error	14.84	16.24	15.89	16.45	15.13

Fourth central moment of the resulting manifest variables in the second simulation example, applying the functions from Equations 26 to 30 either as latent linking functions with linear error functions (non-normal latent) or vice versa (non-normal error).

$$\mathbf{H1} \quad h_1(x) = h_6(x) = x^5 \tag{26}$$

$$\mathbf{H2} \quad h_2(x) = h_7(x) = \exp(x) \tag{27}$$

$$\mathbf{H3} \quad h_3(x) = h_8(x) = \begin{cases} -x^4 & \text{if } x < 0 \\ x^{\frac{1}{2}} & \text{if } x \geq 0 \end{cases} \tag{28}$$

$$\mathbf{H4} \quad h_4(x) = h_9(x) = x^5 + 10\sin(x) \tag{29}$$

$$\mathbf{H5} \quad h_5(x) = h_{10}(x) = \begin{cases} -50 & \text{if } x \leq -3 \\ -1 & \text{if } -3 < x \leq 0 \\ 1 & \text{if } 0 < x \leq 3 \\ 50 & \text{if } x > 3 \end{cases} \tag{30}$$

All functions meet the required assumption of being able to create more extreme values for the fourth moment than desired (Equation 19), as can easily be shown by applying the algorithm to the functions without prespecified fourth moments. The functions were used in two designs. One design implemented the non-linear functions  $h_1$ – $h_5$  as latent linking functions and used linear functions for the error terms, while the second design used linear latent linking functions and non-linear linking functions  $h_1$ – $h_5$  for the errors. The simulation used 10,000 samples of sample size  $n = 500$  for each design and an accuracy of  $a = 7$ .

Table 2 shows the average resulting fourth central moments of the five distributions. The central moments approximate the desired  $\mu_4 = 15$  for all variables and in both designs. Note, however, that the resulting marginal distributions differ depending on the choice of linking functions. Figure 5 exemplarily shows the histograms related to functions  $h_1$  and  $h_5$ , generated from

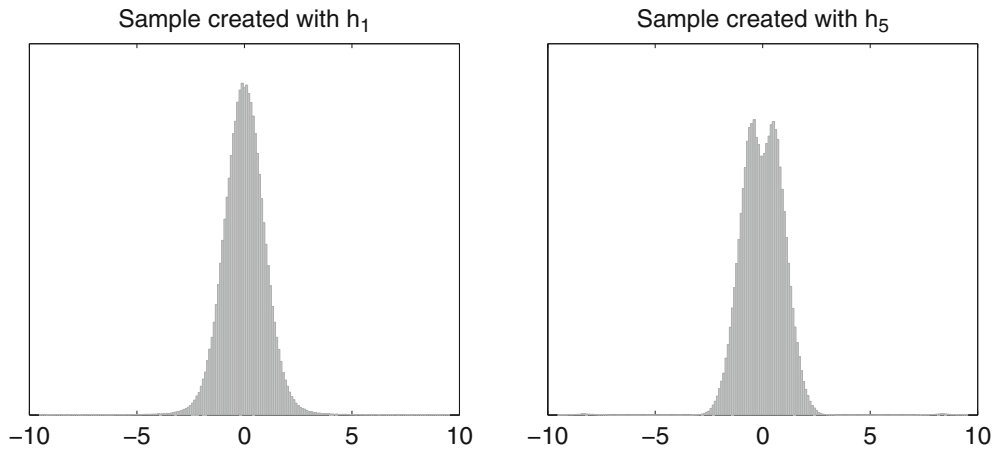


FIGURE 5.

Samples created with the functions  $h_1$  and  $h_5$  as the error transformation functions, estimated to create distributions with identical fourth moments  $\mu_4 = 15$ .

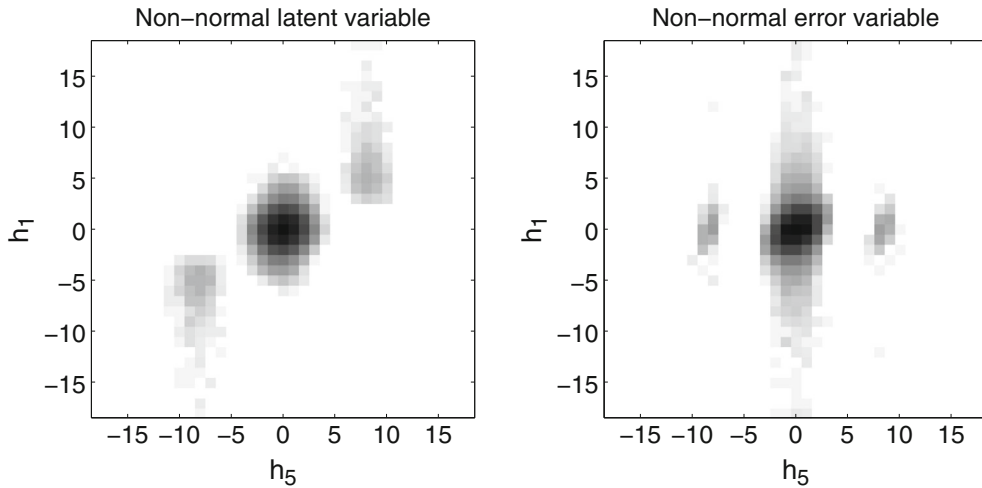


FIGURE 6.

Heat map of distributions  $h_1$  and  $h_5$ . The *left panel* shows the bivariate distribution given that the latent variables are transformed; the *right panel* shows transformed errors. Both distributions are specified to have  $\mu_4 = 15$  and a Pearson correlation of .21.

a sample size  $N = 500,000$  for the non-normal error distributions. Both distributions are symmetrical (and therefore equivalent on every odd central moment) and by design equivalent on the fourth moment. With the power constants approach (Vale & Maurelli, 1983), we could not differentiate between these two distributions.

Figure 6 shows a heat map of the bivariate distribution related to functions  $h_1$  and  $h_5$  for both designs. Compared to non-normal error terms, using the same non-normality transformation on the latent variables creates a different dependency pattern, which in turn could lead to different robustness behavior. The algorithm is thus able to generate different multivariate distributions sharing the same (say, fourth) moment, allowing stricter assessments of robustness.

TABLE 3.  
Kolmogorov–Smirnov distances and empirical rejection rates under different conditions of non-normality.

Design	$T_{ML}$		$T_{SB}$	
	KS-distance	RR (%)	KS-distance	RR (%)
Non-normal error	.03	5.3	.05	6.5
Non-normal latent	.29	23.6	.14	13.0
Vale–Maurelli	.15	13.9	.08	7.4

Kolmogorov–Smirnov (KS) distances and empirical rejection rates (RR) of the normal theory likelihood-ratio test statistic ( $T_{ML}$ ) and the mean-scaled test statistic ( $T_{SB}$ ) under data generated by the linking functions approach using non-normal latent linking functions with linear error functions (non-normal latent) or linear latent functions with non-linear error functions (non-normal error), or data generated via the approach by Vale and Maurelli.

To illustrate the relevance of being able to generate different multivariate distributions based on univariate distributions sharing the same value for a certain moment, we estimated the (correctly specified) confirmatory factor model described above in each of the first 1000 generated samples of the non-normal latent and the non-normal error condition, respectively. For comparison purposes, we also generated 1000 data sets using the Vale–Maurelli approach with the first four moments specified to be equal to the ones generated by the linking functions approach (i.e., the fourth moment was 15 for all indicator variables, while the third moment ranged from  $-1.39$  to  $1.51$  depending on the indicator variable). The factor model was estimated with MPlus (version 7.11) using normal theory maximum likelihood. We considered both the likelihood-ratio test statistic  $T_{ML}$  and the Satorra–Bentler mean-scaled test statistic  $T_{SB}$  (Satorra & Bentler, 2012).

Table 3 shows the Kolmogorov–Smirnov distances to the asymptotic  $\chi^2(34)$  distribution as well as the empirical rejection rates using a nominal  $\alpha$ -error level of .05. It is evident that the behavior of both test statistics varied as a function of the data generating approach, despite the fact that univariate kurtosis was always the same. In particular,  $T_{ML}$  closely followed the theoretical  $\chi^2$  distribution in the non-normal error condition, showed a moderate bias in the Vale–Maurelli condition,<sup>1</sup> and exhibited a substantial bias in the non-normal latent condition.  $T_{SB}$  followed a similar pattern, but performed generally somewhat better compared to  $T_{ML}$ . This study thus indicates that the behavior of test statistics in SEM is not well described by considering univariate kurtosis in isolation, but depends on characteristics of the underlying multivariate distribution, which can be successfully manipulated using the proposed linking functions approach.

### 3. Discussion and Conclusion

In the present paper, we presented a method to generate correlated, non-normally distributed multivariate data. The method is based on a structural model involving manifest and latent variables as well as error terms. Non-normality is introduced by specifying arbitrary linking functions to the latent part, the error part, or both. We further showed how to determine the linking functions such that the degree of non-normality can be systematically varied (for example in terms of a specific moment). Although we exemplified the proposed procedure drawing on monotone linking functions, the algorithm does not require monotony. However, using monotone linking functions is recommended, as this is associated with larger correlations between latent and transformed latent

<sup>1</sup>Note that the empirical rejection rate of  $T_{ML}$  for the Vale–Maurelli data sets appears only moderately inflated when compared to those observed in similar simulation studies (e.g., Curran, West, & Finch, 1996; Savalei, 2010). This discrepancy is due to the use of lower loadings and lower factor intercorrelations in the present study.

variables, thereby meeting the restriction in Equation 9 even for high absolute loadings. Depending on the applied linking functions, the algorithm generates distributions varying considerably in any central moment and is associated with reasonably small sampling errors, while not using a computationally demanding trial-and-error process. The simulation examples demonstrated the flexibility and good convergence behavior.

Using the proposed algorithm depends on expressing the target covariance matrix by specifying a congeneric structural model. Cross-loadings are an important specification in SEM research, for example by allowing to evaluate the effects of misspecified factor loadings compared to misspecified factor covariances (Hu & Bentler, 1998; see also Moshagen, 2012). In the present algorithm, it is not advisable to incorporate cross-loadings by simply introducing an additional loading parameter, because this would be associated with overly strict boundary conditions of the estimation routine. However, cross-loadings can be modeled by adding additional latent variables that capture cross-loadings to the structural model. Suppose that a manifest variable  $M_1$  is needed to depend on both  $L_1$  and  $L_2$  with loadings  $k_1$  and  $k_2$ . An additional latent variable  $L_{m+1}$  with loading  $k'$  can be introduced, where the correlations  $r(L_1, L_{m+1})$  and  $r(L_2, L_{m+1})$  are set such that  $k'r(L_1, L_{m+1}) = k_1$  and  $k'r(L_2, L_{m+1}) = k_2$  holds. More generally, note that any desired manifest covariance matrix can be represented by a suitable saturated structural equation model. Although the data generation routine builds upon a certain structural model, it is ultimately agnostic to its particular structural assumptions, as the resulting data will be compatible with any equivalent model.

Simulation methods involving the weighted addition of distributions have the consequence that the resulting distributions depend on the predefined loadings (Headrick & Sawilowsky, 1999). Identical linking functions result in different distributions in the manifest variables for different loading structures, which may be undesired in some situations. However, this issue can be circumvented by correlating the errors accordingly. For simplicity, suppose that two standardized loadings  $k_1 > k_2 > 0$  are assumed for two manifest variables  $M_{i_1}$ ,  $M_{i_2}$  and the goal is to replicate the distribution generated for  $k_1$  on both variables in the case of  $k_2$ . Then, the errors for both manifest variables need to be adjusted to

$$(\Sigma_{\xi})_{i_1, i_2} = \frac{k_2^2 - k_1^2}{1 - k_1^2} \quad (31)$$

The generated non-normal variables will have the same distribution.

A modification of the algorithm can be used for simulation studies of methods for dimensionality reduction such as factor analysis or principal component analysis. The purpose of these methods is to reduce manifest variables to fewer latent variables. Therefore, the specific intercorrelations among the manifest variables that are not due to the latent variables seem less important. We suggest correcting only for the covariance change to latent variables and error terms, which would provide a stricter test for dimensionality reduction methods.

In a more general case, it is often desirable to have a non-normal distribution in the manifest variables following a given multivariate density function. The procedure presented herein relies on linking functions to determine the distribution of the manifest variables. An alternative approach would be to parameterize the applied functions and to modify these in an iterative fashion such that the resulting distribution matches a given multivariate target distribution. Consider a set of linking functions depending on one or more parameter. Given Equations 11–12, the estimated correction parameters  $b_j$  and  $c_j$ , and the correlation of the corresponding latent variables, the difference between the target distribution and the distribution for a pair of functions with given parameters can be estimated. By minimizing this difference for the set of functions by means of a suitable optimization algorithm, the parameters of the functions can be iteratively adjusted to increase

the match between actual and target distribution. However, the parameterized functions need to be highly flexible and of sufficient generality in order to capture a wide range of desired target distributions. Moreover, a number of problems might occur during the minimization process that warrant special attention. Nonetheless, this approach provides an interesting avenue for further research.

In summary, the paper presented an algorithm to generate correlated non-normal data using linking functions to transform normal distributions. The algorithm counteracts the deviance in variance, correlation to the latent variables, and covariance among the manifest variables. The linking functions can be estimated to obtain distributions with any single prespecified central moment. The sampling error for the manifest covariance matrix converges to zero for a wide range of different distributions and increasing sample size. The procedure is computationally undemanding and produces a wide range of different distributions valuable for stricter robustness studies concerning non-normality.

### Acknowledgments

This research was supported in parts by a grant from the Baden-Württemberg foundation to the second author

### References

- Bradley, D. R., & Fleisher, C. L. (1994). Generating multivariate data from nonnormal distributions: Mihal and Barrett revisited. *Behavior Research Methods, Instruments, & Computers*, *26*, 156–166. doi:[10.3758/BF03204610](https://doi.org/10.3758/BF03204610).
- Burr, I. W. (1942). Cumulative frequency functions. *The Annals of Mathematical Statistics*, *13*, 215–232. doi:[10.1214/aoms/1177731607](https://doi.org/10.1214/aoms/1177731607).
- Cario, M. C., & Nelson, B. L. (1998). Numerical methods for fitting and simulating autoregressive-to-anything processes. *INFORMS Journal on Computing*, *10*, 72–81.
- Cook, R. D., & Johnson, M. E. (1981). A family of distributions for modelling non-elliptically symmetric multivariate data. *Journal of the Royal Statistical Society. Series B*, *43*, 210–218. doi:[10.2307/2984851](https://doi.org/10.2307/2984851).
- Curran, P. J., West, S. G., & Finch, J. F. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological Methods*, *1*, 16–29.
- Devroye, L. (1986). *Non-uniform random variate generation*. New York: Springer.
- Fang, K.-T., Kotz, S., & Ng, K. W. (1990). *Symmetric multivariate and related distributions*. London: Chapman and Hall.
- Fleishman, A. I. (1978). A method for simulating non-normal distributions. *Psychometrika*, *43*, 521–532. doi:[10.1007/BF02293811](https://doi.org/10.1007/BF02293811).
- Foldnes, N., & Grønneberg, S. (in press). How general is the Vale-Maurelli simulation approach? *Psychometrika*. doi:[10.1007/s11336-014-9414-0](https://doi.org/10.1007/s11336-014-9414-0).
- Headrick, T. C. (2002). Fast fifth-order polynomial transforms for generating univariate and multivariate nonnormal distributions. *Computational Statistics & Data Analysis*, *40*, 685–711. doi:[10.1016/S0167-9473\(02\)00072-5](https://doi.org/10.1016/S0167-9473(02)00072-5).
- Headrick, T. C. (2010). *Statistical simulation: Power method polynomials and other transformations*. Boca Raton, FL: Chapman & Hall/CRC.
- Headrick, T. C., & Kowalchuk, R. K. (2007). The power method transformation: Its probability density function, distribution function, and its further use for fitting data. *Journal of Statistical Computation and Simulation*, *77*, 229–249. doi:[10.1080/10629360600605065](https://doi.org/10.1080/10629360600605065).
- Headrick, T. C., & Mugdadi, A. (2006). On simulating multivariate non-normal distributions from the generalized lambda distribution. *Computational Statistics & Data Analysis*, *50*, 3343–3353. doi:[10.1016/j.csda.2005.06.010](https://doi.org/10.1016/j.csda.2005.06.010).
- Headrick, T. C., & Sawilowsky, S. S. (1999). Simulating correlated multivariate nonnormal distributions: Extending the Fleishman power method. *Psychometrika*, *64*, 25–35. doi:[10.1007/BF02294317](https://doi.org/10.1007/BF02294317).
- Hodis, F. A., Headrick, T. C., & Sheng, Y. (2012). Power method distributions through conventional moments and L-moments. *Applied Mathematical Sciences*, *6*, 2159–2193.
- Hu, L.-T., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods*, *3*, 424–453. doi:[10.1037/1082-989X.3.4.424](https://doi.org/10.1037/1082-989X.3.4.424).
- Joe, H. (1997). *Multivariate models and multivariate dependence concepts*. Boca Raton, FL: Chapman & Hall/CRC.
- Johnson, N. L. (1949). Systems of frequency curves generated by methods of translation. *Biometrika*, *36*, 149–176. doi:[10.2307/2332539](https://doi.org/10.2307/2332539).
- Mair, P., Satorra, A., & Bentler, P. M. (2012). Generating nonnormal multivariate data using copulas: Applications to SEM. *Multivariate Behavioral Research*, *47*, 547–565. doi:[10.1080/00273171.2012.692629](https://doi.org/10.1080/00273171.2012.692629).
- Mattson, S. (1997). How to generate non-normal data for simulation of structural equation models. *Multivariate Behavioral Research*, *32*, 355–373. doi:[10.1207/s15327906mbr3204\\_3](https://doi.org/10.1207/s15327906mbr3204_3).



- Moshagen, M. (2012). The model size effect in SEM: Inflated goodness-of-fit statistics are due to the size of the covariance matrix. *Structural Equation Modeling, 19*, 86–98. doi:[10.1080/10705511.2012.634724](https://doi.org/10.1080/10705511.2012.634724).
- Nagahara, Y. (2004). A method of simulating multivariate nonnormal distributions by the Pearson distribution system and estimation. *Computational Statistics & Data Analysis, 47*, 1–29. doi:[10.1016/j.csda.2003.10.008](https://doi.org/10.1016/j.csda.2003.10.008).
- Ramberg, J. S., & Schmeiser, B. W. (1974). An approximate method for generating asymmetric random variables. *Communications of the ACM, 17*, 78–82. doi:[10.1145/360827.360840](https://doi.org/10.1145/360827.360840).
- Ruscio, J., & Kaczetow, W. (2008). Simulating multivariate nonnormal data using an iterative algorithm. *Multivariate Behavioral Research, 43*, 355–381. doi:[10.1080/00273170802285693](https://doi.org/10.1080/00273170802285693).
- Savalei, V. (2010). Expected versus observed information in SEM with incomplete normal and nonnormal data. *Psychological Methods, 15*, 352–367.
- Tadikamalla, P. R. (1980). On simulating non-normal distributions. *Psychometrika, 45*, 273–279. doi:[10.1007/BF02294081](https://doi.org/10.1007/BF02294081).
- Vale, C. D., & Maurelli, V. A. (1983). Simulating multivariate nonnormal distributions. *Psychometrika, 48*, 465–471. doi:[10.1007/BF02293687](https://doi.org/10.1007/BF02293687).
- Yuan, K.-H., & Bentler, P. M. (1999). On normal theory and associated test statistics in covariance structure analysis under two classes of nonnormal distributions. *Statistica Sinica, 9*, 831–853.

*Manuscript Received: 11 AUG 2013*

*Final Version Received: 31 JAN 2015*

Sampling from arbitrary non-normal distributions with given covariance and central  
moments

Max Auerswald and Morten Moshagen

Ulm University

Author Note

Max Auerswald and Morten Moshagen, Institute of Psychology and Education,  
University of Ulm, Ulm, Germany.

Correspondence concerning this article should be addressed to Max Auerswald,  
Research Methods, Institute of Psychology and Education, University of Ulm,  
Albert-Einstein-Allee 47, 89081 Ulm, Germany. E-Mail: [max.auerswald@uni-ulm.de](mailto:max.auerswald@uni-ulm.de)

## Abstract

The article develops an algorithm to generate multivariate samples with prespecified central moments from a population with a non-normal distribution and given correlation matrix. The algorithm extends the NORTA approach, a method that generates random vectors with arbitrary marginal distributions, by determining an appropriate inverse cumulative distribution function (*CDF*). The inverse *CDF* is estimated as a quantile mixture of prespecified distributions to comply with the desired central moments. A Monte Carlo simulation demonstrates the range of distributions and central moments for which the algorithm is feasible. The algorithm is easy to apply, fast, and implemented in the widely used and open-source R environment.

*Keywords: Non-normal multivariate data; Simulation; Skewness and kurtosis*

Sampling from arbitrary non-normal distributions with given covariance and central moments

Many statistical methods require the assumption of multivariate normality. By applying the statistical test to simulated samples of a population with known covariance, Monte Carlo studies can provide boundaries in which the validity of these methods is not compromised. The samples are generated to violate the normality assumption to a certain degree, while systematically varying the population covariance. Although it is straightforward to generate either normally distributed data with known covariance or non-normally distributed variables with zero or unknown covariance, jointly meeting both requirements is challenging. Distorting a multivariate normal distribution (e.g. by transformation functions) influences the covariance structure. Vice versa, creating covariance by adding non-normally distributed variables results in distributions converging to a normal distribution with increasing number of variables due to the central limit theorem.

A number of strategies have been proposed to simulate samples from a population with non-normal distribution and a prespecified covariance matrix (e.g. Bradley & Fleisher, 1994; Cook & Johnson, 1981; Foldnes & Olsson, 2016; Headrick & Mugdadi, 2006; Koran, Headrick, & Kuo, 2015; Mair, Satorra, & Bentler, 2012; Ruscio & Kaczetow, 2008). Vale and Maurelli (1983) proposed a three-step multivariate power constants approach (MPC) extending the power constants approach (Fleishman, 1978), which is perhaps the most popular approach. First, a system of equations is solved to obtain polynomial transformation functions which result in prespecified first four moments of a distribution. More precisely, for  $Z \sim \mathcal{N}(0, 1)$ , the method solves

$$X = a + bZ + cZ^2 + dZ^3, \tag{1}$$

for parameters  $a, b, c, d$ , to achieve the desired skewness  $\gamma_3$  and kurtosis  $\gamma_4$  in  $X$ .<sup>1</sup> Second, the distorting effect of the transformation on the covariance matrix is determined and

---

<sup>1</sup>Skewness and kurtosis are defined as the standardized third and fourth central moment. Unlike central

counteracted, resulting in an intermediate covariance matrix. Third, normally distributed random variables are sampled according to the intermediate covariance matrix and non-normalized using the polynomial transformation functions. The resulting variables are non-normal according to the prespecified skewness and kurtosis, and comply with the prespecified covariance matrix. The method was further extended to incorporate higher order polynomials (Headrick, 2002; Headrick & Sawilowsky, 1999) and different sampling distributions, e.g. uniform distributions (Hodis, Headrick, & Sheng, 2012).

The ability of MPC to create samples with given skewness and kurtosis is advantageous in the context of Monte Carlo simulations. The possibility to vary continuously both measures allows determining a range for skewness and kurtosis, in which a specific test is robust against violations of distributional assumptions. For any data set and statistical test, skewness and kurtosis can then be assessed with regards to these boundaries. However, MPC also suffers some drawbacks. First, the procedure relies on finding a polynomial transformation function. Certain families of probability distributions, e.g. the  $\chi^2$  or lognormal distributions, cannot be generated relying on polynomials only (Ruscio & Kaczetow, 2008). Second, although different probability distributions can share the same (first) central moments, MPC always generates samples from the one particular distribution (associated with a given skewness and kurtosis). However, the robustness of a statistical test under given first central moments may vary with higher order moments, in turn leading to invalid conclusions for the boundaries of skewness and kurtosis (Astivia & Zumbo, 2014). Third, recent results suggest that the robustness of many statistical tests primarily depends on the underlying multivariate distribution, even if the marginal distributions are similar (Auerswald & Moshagen, 2015; Foldnes & Grønneberg, 2015).

---

moments, skewness and kurtosis take the variance of the random variable into account. However, in the context of sampling correlated, non-normally distributed variables, any manipulation of the standardized moments also affects the (unstandardized) central moments and vice versa. Both manipulations are equivalent because they only differ due to the variance of the random variable, which is easily manipulated by multiplying a constant to said random variable. We therefore use both terms interchangeably.

Foldnes and Grønneberg (2015) investigated the effect of (upper) tail dependence between two random variables  $X_i, X_j$ , a measure of multivariate non-normality. Upper tail dependence is defined as the probability that  $X_i$  exceeds its  $p$ -th quantile, conditional on  $X_j$  exceeding its  $p$ -th quantile. Foldnes and Grønneberg (2015) evaluated a structural equation model with data that were either generated using MPC or a copula based approach, while the latter allowed for higher tail dependence. Maximum likelihood estimation performed less favorable with regards to bias and standard error for the copula based samples, compared to Vale and Maurelli samples, while kurtosis was similar in both cases.

The NORTA (NORmal To Anything) method is an alternative approach that partially addresses the disadvantages of the Vale and Maurelli procedure (Cario & Nelson, 1997). NORTA generates samples of a random vector with given target correlation matrix and marginal distribution. In contrast to MPC, NORTA allows the specification of any marginal distribution, including  $\chi^2$ , lognormal, and also discrete distributions. The specification as a probability distribution also determines the central moments of the resulting random variable. Despite this increased flexibility, the reliance on probability distributions can be disadvantageous when the goal is to find boundaries for measures of non-normality, such as skewness and kurtosis, in which a statistical test is robust. Furthermore, NORTA shares the limitation with MCP of considering the marginal distributions only.

In the present article, an algorithm is introduced that combines the advantages of MPC with those of the NORTA approach and also allows the manipulation of the multivariate distribution independently of univariate skewness and kurtosis. To this end, the NORTA approach is extended to allow for the specification of skewness, kurtosis, and central moments of the desired distribution continuously, while preserving the distributional flexibility of NORTA. The algorithm is computationally undemanding and has been implemented in the NOTAMO (NORmal To Arbitrary MOments) package (see Appendix) for the open-source statistical computing language R (R Core Team, 2016).

### Algorithm description

The purpose of the algorithm is to create multivariate samples from a population that is distributed according to a set of moments (or related distributional measures such as skewness and kurtosis), a set of inverse cumulative distribution functions (*CDFs*), and correlation matrix  $\Sigma_T$ . First, the algorithm determines parameters for a linear combination of inverse *CDFs* to comply with the prespecified set of moments. Second, the NORTA approach is used to estimate an intermediate correlation matrix  $\Sigma_N$  that counters the distortion in correlations introduced in the first step. Finally, the algorithm samples normally distributed random variables according to  $\Sigma_N$  and transforms them to comply with the prespecified moments. The resulting variables are distributed according to the desired correlation matrix  $\Sigma_T$ .

The extension of the NORTA method requires a set of inverse *CDFs* for every random variable. We denote the possible inverse *CDFs* for the  $k$ -th random variable as  $F_{k(1)}^{-1}, \dots, F_{k(m)}^{-1}$ . If only a single inverse *CDF* were to be specified, all moments would be determined by the associated distribution (as is the case with traditional NORTA). Instead, the set of inverse *CDFs* allows to estimate a linear combination of inverse *CDFs* that complies with the prespecified moments. Let  $n$  be the number of random variables and  $\Sigma_T$  the target correlation matrix. The algorithm needs to find parameters  $a_{k(1)}, \dots, a_{k(m)}$  such that

$$F_k^{-1} = \sum_{j=1}^m a_{k(j)} F_{k(j)}^{-1}, \quad (2)$$

where  $F_k^{-1}$  is the inverse *CDF* associated with the prespecified central moments,  $1 \leq k \leq n$ , and

$$\sum_{j=1}^m a_{k(j)} = 1, \quad a_{k(j)} \geq 0. \quad (3)$$

Inverse *CDFs* as in Equation 2 are known as quantile mixtures, in analogy to mixtures of probability density functions (Karvanen, 2006). Note that it is necessary that

$$\min_{k=1..m} \mu_{F_k^{-1}} \leq \mu \leq \max_{k=1..m} \mu_{F_k^{-1}} \quad (4)$$

holds, for the desired moment  $\mu$ . The algorithm cannot find a quantile mixture that complies with moment  $\mu$  if  $\mu$  exceeds the associated moment of every distribution in the set. However, as can be seen in the simulation example, only a small number of inverse *CDFs* is typically sufficient to achieve high distributional flexibility.

The algorithm estimates the parameters  $a_{k(1)}, \dots, a_{k(m)}$  in Equation 2 using a vector of  $p$ -values, similarly to the technique used by Auerswald and Moshagen (2015). The vector of  $p$ -values needs to be equally spaced, e.g. starting at  $10^{-b}$ , increasing in steps of  $10^{-b}$  up to  $1 - 10^{-b}$ , for  $b \in \mathbb{N}$ . For any inverse *CDF*  $F^{-1}$  and uniformly distributed random variable  $U$ ,  $U \sim U[0, 1]$ , the random variable  $F^{-1}(U)$  is distributed according to  $F$ , so  $F^{-1}(U) \sim F$  (e.g. Embrechts & Hofert, 2013). As the vector of  $p$ -values, denoted as  $p_b$ , captures the characteristics of a uniform distribution,  $F_k^{-1}(p_b)$  can be used to estimate (among other aspects) the skewness and kurtosis of  $F_k$ , as will be demonstrated in the simulation example below. The prespecified inverse *CDFs*  $F_{k(j)}^{-1}$  and the resulting parameters  $a_{k(j)}$  then constitute the desired inverse *CDF*  $F_k^{-1}$  for latent variable  $k$ .

The NORTA method is then used to estimate an intermediate correlation matrix  $\Sigma_N$ , that counteracts the distortion introduced by  $F_k^{-1}$ . The NORTA approach generates random samples with given univariate distributions and correlation matrix. Instead of MPC's polynomial transformation functions, NORTA generates non-normal random variables  $X_k$  with  $1 \leq k \leq n$  as

$$X_k = F_k^{-1}(\Phi(Z_k)), \quad (5)$$

where  $\Phi(\cdot)$  denotes the *CDF* of the standard normal distribution,  $F_k^{-1}$  the inverse *CDF* of the distribution associated with  $F_k$ , and standard normally distributed variables  $Z_k$ . Note that the transformation function  $F_k^{-1}(\Phi(\cdot))$  ensures that  $X_k$  is distributed according to  $F_k$ . Similarly to MPC, the crucial part of the algorithm is to determine an intermediate correlation matrix  $\Sigma_N$  that anticipates and counteracts the distortion in correlation, introduced by the non-normality transformation in Equation 5. The correlation of two random variables  $X_{k_1}$  and  $X_{k_2}$  is directly determined by the respective correlation of  $Z_{k_1}$



and  $Z_{k_2}$ , since

$$\text{Corr}(X_{k_1}, X_{k_2}) = \text{Corr}(F_{k_1}^{-1}(\Phi(Z_{k_1})), F_{k_2}^{-1}(\Phi(Z_{k_2}))). \quad (6)$$

Furthermore, the algorithm only needs to consider  $E[X_{k_1}X_{k_2}]$ , the expected value of  $X_{k_1}X_{k_2}$ , because the mean and standard deviation of both random variables is fixed by their respective *CDFs* and

$$\text{Corr}(X_{k_1}, X_{k_2}) = \frac{E[X_{k_1}X_{k_2}] - E[X_{k_1}]E[X_{k_2}]}{(\text{Var}(X_{k_1})\text{Var}(X_{k_2}))^{-\frac{1}{2}}}. \quad (7)$$

The expected value is

$$E[X_{k_1}X_{k_2}] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} F_{X_{k_1}}^{-1}(\Phi(z_i)) F_{X_{k_2}}^{-1}(\Phi(z_j)) \phi_{\rho(i,j)}(z_i, z_j) dz_i dz_j, \quad (8)$$

where  $\phi_{\rho(i,j)}$  is the bivariate standard normal probability density function given correlation  $\rho(i,j)$  (Cario & Nelson, 1997). A closed form expression of Equation 8 is only available for special cases, but numerical approximations provided by Cario and Nelson (1997) converge under mild conditions.

The algorithm can be summarized as follows. First, a quantile mixture is estimated for every variable. The parameters of the quantile mixture are set according to the prespecified central moments. Then, the NORTA approach is used to estimate an intermediate correlation matrix  $\Sigma_N$ . Finally, standard normal variables are sampled according to  $\Sigma_N$  and transformed by the function  $F_k^{-1}(\Phi(\cdot))$ . The resulting variables comply with both the prespecified correlation matrix and the central moments for each variable. The next section addresses a further extension of the algorithm that also allows to alter the multivariate distribution directly.

### Manipulating the multivariate distribution

The algorithm can also be used to vary the multivariate distribution, independently of target skewness and kurtosis. Without loss of generality, assume that the random variables  $X_k$  generated with the described algorithm have unit variance. Skewness and kurtosis are

defined as the standardized third and fourth central moment respectively, so in this case

$$\gamma_3 = \frac{\mu_3}{\sigma^3} = \mu_3, \quad (9)$$

for the third central moment  $\mu_3$ , and

$$\gamma_4 = \frac{\mu_4}{\sigma^4} = \mu_4, \quad (10)$$

for the fourth central moment  $\mu_4$ . The central moments  $\mu_3, \mu_4$  can be used to calculate the corresponding cumulants  $\kappa_3, \kappa_4$  as

$$\kappa_3 = \mu_3, \quad (11)$$

and

$$\kappa_4 = \mu_4 - 3\mu_2^2 = \mu_4 - 3. \quad (12)$$

Cumulants have the property of additivity, so

$$\kappa_n(X_1 + X_2) = \kappa_n(X_1) + \kappa_n(X_2) \quad (13)$$

holds for any cumulant and independent random variables  $X_1, X_2$ . We define

$$X_k = L_k + E_k \quad (14)$$

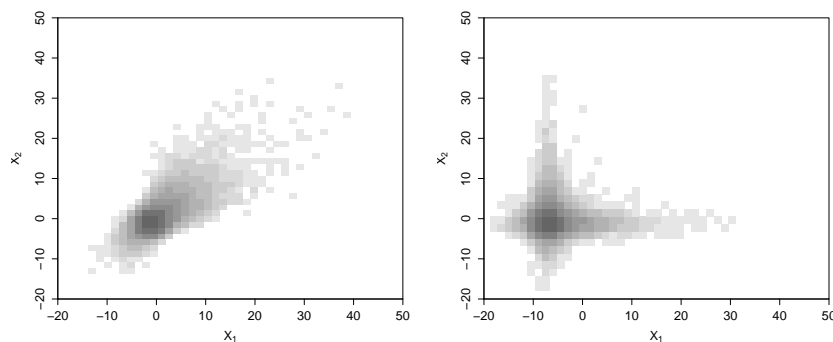
with random variables  $L_k$  and independent random variables  $E_k$ . We also require that  $L_k$  is independent from any variable  $E_k$ . Instead of generating the target random variables  $X_k$  directly, the algorithm can be used to generate either  $L_k$  or  $E_k$ , while the other is normally distributed. For example, if  $E_k$  is normal,  $\gamma_{3E} = 0$  and  $\gamma_{4E} = 3$ . The algorithm can be used to adjust the target correlation, skewness and kurtosis of  $L_k$  according to Equations 9-14, to obtain the desired  $\gamma_3, \gamma_4$  in  $X_k$ . Correspondingly, if  $L_k$  is normal, the algorithm adjusts skewness and kurtosis of  $E_k$ .

The effect on a bivariate distribution with normal  $L_k$  and non-normal  $E_k$  (and vice versa) is illustrated in Figure 1. The target skewnesses for the non-normal variables were 2

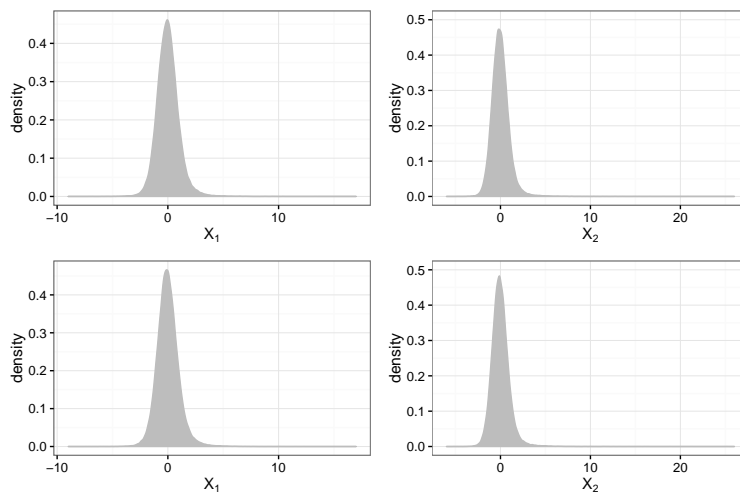
and 5, target kurtoses were 20 and 70, and all variables had unit variance.<sup>2</sup> The target moments and set of inverse *CDFs* together determine the resulting marginal distribution of the generated non-normal variables. Due to the same prespecification, the marginal distribution of the non-normal  $L_1$  is identical to the marginal distribution of the non-normal  $E_1$ . As a consequence,  $X_1$  is in both cases the sum of a normally distributed variable and a non-normally distributed variable specified by the associated quantile mixture.  $X_1$  has the same marginal distribution independent of  $L_1$  or  $E_1$  being the cause for the non-normality, because  $E_1$  and  $L_1$  are independent by definition (see Figure 2). The algorithm is therefore able to manipulate the multivariate distribution while keeping the marginal distributions identical. In this example, the correlation of  $L_1$  and  $L_2$  was  $r = .8$  in both cases, resulting in a correlation of  $r = .4$  for  $X_1$  and  $X_2$ . Note that the assumption of unit variance for  $L_k$  and  $E_k$  would lead to a restricted range of possible correlations among the  $X_k$  as all  $E_k$  are uncorrelated by definition. We therefore only required unit variance in  $X_k$ , allowing for a wider range of possible correlations.

---

<sup>2</sup>The algorithm used the same set of inverse *CDFs* as for the first and second variable in the simulation example.



*Figure 1.* Heatmap of bivariate distributions. The left panel shows the bivariate distribution according to Equation 14 if the dependent  $L_k$  are generated with the proposed algorithm and normal (independent)  $E_k$ . The right panel shows the reversed case. Importantly, the correlation coefficients ( $r = .40$ ) are identical in both cases.



*Figure 2.* Marginal distributions. The left panels show the marginal distributions of  $X_1$ , the right panels show the marginal distributions for  $X_2$ . In the top panels, dependent variables  $L_k$  were generated with the proposed algorithm while  $E_k$  was normally distributed. The bottom panels show the reversed case. Independently of  $L_k$  or  $E_k$  being non-normally distributed, the resulting marginal distributions are identical.

### Simulation example

The proposed algorithm relies on numeric approximation in two parts: (1) the estimation of the appropriate inverse *CDFs*; (2) the estimation of an intermediate correlation matrix  $\Sigma$  that anticipates the effect of the non-normality transformation on the correlation. To evaluate the performance of the algorithm, we conducted a simulation study based on six variables with target correlation matrix

$$\Sigma_T = \begin{pmatrix} 1 & & & & & \\ .49 & 1 & & & & \\ .49 & .49 & 1 & & & \\ .15 & .15 & .15 & 1 & & \\ .15 & .15 & .15 & .49 & 1 & \\ .15 & .15 & .15 & .49 & .49 & 1 \end{pmatrix} \quad (15)$$

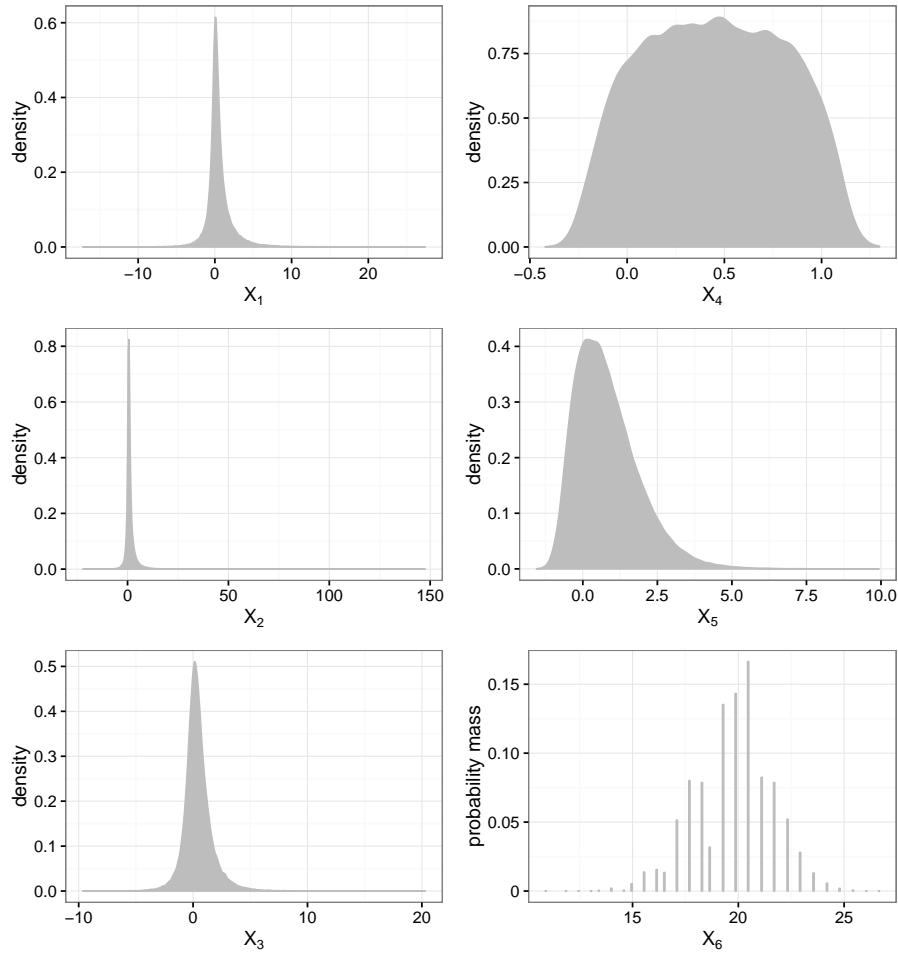
which corresponds to a congeneric two-factor model with loadings of  $\lambda = .7$  and factors correlated to  $r = .3$ .

In defining the distributions for the six variables, we attempted to cover a wide range of distributions, skewnesses, and kurtoses:

- A quantile mixture of a standard normal distribution, a cubic standard normal, and a  $\chi^2$  distribution with  $df = 1$ , with target  $\gamma_3 = 2$  and  $\gamma_4 = 20$ .
- A quantile mixture of a standard normal distribution, a cubic standard normal, and a lognormal distribution with target  $\gamma_3 = 5$  and  $\gamma_4 = 70$
- A quantile mixture of a standard normal distribution, a cubic standard normal, and an exponential distribution with target  $\gamma_3 = 1$  and  $\gamma_4 = 10$
- A quantile mixture of a standard normal distribution and a uniform distribution on the interval  $[0, 1]$  with target  $\gamma_4 = 2$

- A quantile mixture of an exponential distribution with rate  $\lambda = 1$ , a uniform distribution on the interval  $[0, 1]$ , and a standard normal distribution with target  $\gamma_3 = 1$  and  $\gamma_4 = 4.5$
- A quantile mixture of a Poisson distribution with  $\lambda = 1$ , a binomial distribution  $B(n = 30, p = .50)$ , and a binomial distribution  $B(n = 30, p = .99)$  with target  $\gamma_3 = -.2$  and  $\gamma_4 = 3$

The resulting probability density functions (or probability mass function, in case of the discrete variable) corresponding to these quantile mixtures are illustrated in Figure 3. In all distributions, we restricted at least one higher moment to illustrate the extended algorithm (note that restrictions on  $\gamma$  are not required). We used an accuracy value of  $b = 7$ , which appears to be a good trade-off between speed and accuracy. Six sample sizes were used:  $N = (50; 100; 300; 1,000; 10,000; 100,000)$ , with 5,000 replications each. Simulation and estimation was obtained using the R package NOTAMO which in turn incorporates the packages moments (Komsta & Novomestky, 2015), multiroot (Soetaert, 2009), nloptr (Johnson, 2014), and NORTARA (Su, 2014).



*Figure 3.* Distributions generated in the simulation example. For continuous variables  $X_1 - X_5$ , the panel depicts the probability density function of the generated random variables. For the discrete variable  $X_6$ , the panel shows the probability mass function.

Table 1 shows the resulting and target skewness and kurtosis of the generated random variables and sample size 100,000. It is evident that the observed values for skewness and kurtosis closely match the respective target for all random variables. Note that the estimation of the parameters in Equation 2 is independent of sample size. However, using larger sample sizes allows for a more accurate assessment of the resulting skewness and kurtosis in the generated samples.

Table 1

*Mean empirical skewness and kurtosis of the six random variables generated in the simulation example with prespecified targets*

	$X_1$	$X_2$	$X_3$	$X_4$	$X_5$	$X_6$
Empirical skewness	2.00	5.01	1.00	0.00	1.00	-.20
Target skewness	2.00	5.00	1.00	0.00	1.00	-.20
Empirical kurtosis	20.05	72.28	10.02	2.00	4.50	3.00
Target kurtosis	20.00	70.00	10.00	2.00	4.50	3.00

*Note.* Observed and prespecified skewness and kurtosis for the variables in the simulation example. The algorithm estimated a quantile mixture distribution to approximate the prespecified target values. The simulation incorporated 5,000 repetitions of sample size 100,000.

The second numeric approximation used in the proposed algorithm refers to the correction of the correlation matrix and is based on the NORTA approach. We calculated the average root mean square error ( $RMSE$ ) between each sample and target correlation matrix, displayed in Table 2. The sampling error decreases with increasing sample size, indicating convergence. For comparison purposes, Table 2 also illustrates the  $RMSE$  for standard normal random variables with the same target correlation matrix  $\Sigma_T$ . It can be seen that the  $RMSE$  of the proposed algorithm is only slightly higher compared to that of standard normal random variables.



Table 2

*Root mean square error of the difference of simulated and target correlation matrices*

$N$	50	100	300	1,000	10,000	100,000
Algorithm	.120	.086	.051	.029	.011	.006
Standard normal	.116	.081	.047	.026	.008	.003

*Note.* Average root mean square error of the difference of the observed and target correlation matrix for variables generated with the proposed algorithm or standard normal variables. Each cell contains 5,000 repetitions.

### Discussion and conclusion

A common goal in studies investigating the robustness of statistical methods is to generate samples that violate the normality assumption to a specific degree, but maintain a certain covariance structure. Herein, we proposed an algorithm for generating multivariate non-normal random variables based on an extension of the NORTA method that also allows the prespecification of any (combination of) central moments. The key idea is to define the inverse  $CDF$  of the non-normal random variables as a quantile mixture distribution. The algorithm estimates the parameters of the quantile mixture distribution to determine a distribution complying with the prespecified central moments. As the simulation example demonstrates, the algorithm can reproduce a wide range of skewness and kurtosis with high accuracy for both continuous and discrete distributions.

Furthermore, the  $RMSE$  of the difference of empirical and target correlation matrices approximates the  $RMSE$  obtained when generating normally distributed random variables. The algorithm is computationally undemanding, easy to apply, and implemented in the open-source statistical computing language R.

The main advantage of the proposed algorithm to the NORTA method is that central moments can be specified. The ability to vary skewness, kurtosis, and other distributional measures continuously allows for a more systematic investigation of the effect of non-normality and increases the comparability to previous robustness studies. The algorithm can be used in Monte Carlo simulations that attempt to determine a range for skewness and kurtosis (or other standardized central moments) in which a statistical test is robust. However, the algorithm also allows for determining if the common practice of reducing the effect of non-normality to skewness and kurtosis is adequate. Different marginal distributions might have a different effect on the validity of a statistical test even if skewness and kurtosis are similar. The algorithm allows the variation of the marginal distributions while specifying the same target central moments by choosing a different quantile mixture. Furthermore, recent simulation studies emphasized the compromising effect of multivariate non-normality on the robustness of statistical tests (Auerswald & Moshagen, 2015; Foldnes & Grønneberg, 2015). Our algorithm can vary key characteristics of the multivariate distribution independently of marginal distributions and prespecified correlation matrix. The invariance in marginal distributions also implies the same skewness, kurtosis, and any central moment of the marginal distributions. Overall, our algorithm offers researchers conducting Monte Carlo simulations a high flexibility in simulating samples from non-normal distributions and may be useful for assessing the robustness of a wide range of statistical tests and data conditions.

## References

- Astivia, O. L. O., & Zumbo, B. D. (2014). A cautionary note on the use of the Vale and Maurelli method to generate multivariate, nonnormal data for simulation purposes. *Educational and Psychological Measurement, 75*, 541–567.  
doi:10.1177/0013164414548894
- Auerswald, M., & Moshagen, M. (2015). Generating correlated, non-normally distributed data using a non-linear structural model. *Psychometrika, 80*, 920–937.  
doi:10.1007/s11336-015-9468-7
- Bradley, D. R., & Fleisher, C. L. (1994). Generating multivariate data from nonnormal distributions: Mihal and Barrett revisited. *Behavior Research Methods, Instruments, & Computers, 26*, 156–166. doi:10.3758/BF03204610
- Cario, M. C., & Nelson, B. L. (1997). *Modeling and generating random vectors with arbitrary marginal distributions and correlation matrix* (Tech. Rep.). Citeseer.
- Cook, R. D., & Johnson, M. E. (1981). A family of distributions for modelling non-elliptically symmetric multivariate data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 210–218.
- Embrechts, P., & Hofert, M. (2013). A note on generalized inverses. *Mathematical Methods of Operations Research, 77*, 423–432. doi:10.1007/s00186-013-0436-7
- Fleishman, A. I. (1978). A method for simulating non-normal distributions. *Psychometrika, 43*, 521–532. doi:10.1007/BF02293811
- Foldnes, N., & Grønneberg, S. (2015). How general is the Vale–Maurelli simulation approach? *Psychometrika, 80*, 1066–1083. doi:10.1007/s11336-014-9414-0
- Foldnes, N., & Olsson, U. H. (2016). A simple simulation technique for nonnormal data with prespecified skewness, kurtosis, and covariance matrix. *Multivariate Behavioral Research, 1*–13. doi:10.1080/00273171.2015.1133274
- Headrick, T. C. (2002). Fast fifth-order polynomial transforms for generating univariate and multivariate nonnormal distributions. *Computational Statistics & Data Analysis*,

- 40, 685–711. doi:10.1016/S0167-9473(02)00072-5
- Headrick, T. C., & Mugdadi, A. (2006). On simulating multivariate non-normal distributions from the generalized lambda distribution. *Computational Statistics & Data Analysis*, 50, 3343–3353. doi:10.1016/j.csda.2005.06.010
- Headrick, T. C., & Sawilowsky, S. S. (1999). Simulating correlated multivariate nonnormal distributions: Extending the Fleishman power method. *Psychometrika*, 64, 25–35. doi:10.1007/BF02294317
- Hodis, F. A., Headrick, T. C., & Sheng, Y. (2012). Power method distributions through conventional moments and L-moments. *Applied Mathematical Sciences*, 6, 2159–2193.
- Johnson, S. G. (2014). The NLOpt nonlinear-optimization package [Computer software manual]. (R package 1.0.4)
- Karvanen, J. (2006). Estimation of quantile mixtures via L-moments and trimmed L-moments. *Computational Statistics & Data Analysis*, 51, 947–959. doi:10.1016/j.csda.2005.09.014
- Komsta, L., & Novomestky, F. (2015). moments: Moments, cumulants, skewness, kurtosis and related tests [Computer software manual]. Retrieved from <https://CRAN.R-project.org/package=moments> (R package version 0.14)
- Koran, J., Headrick, T. C., & Kuo, T. C. (2015). Simulating univariate and multivariate nonnormal distributions through the method of percentiles. *Multivariate Behavioral Research*, 50, 216–232. doi:10.1080/00273171.2014.963194
- Mair, P., Satorra, A., & Bentler, P. M. (2012). Generating nonnormal multivariate data using copulas: Applications to SEM. *Multivariate Behavioral Research*, 47, 547–565. doi:10.1080/00273171.2012.692629
- R Core Team. (2016). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from <https://www.R-project.org/>
- Ruscio, J., & Kaczetow, W. (2008). Simulating multivariate nonnormal data using an

iterative algorithm. *Multivariate Behavioral Research*, 43, 355–381.

doi:10.1080/00273170802285693

Soetaert, K. (2009). rootsolve: Nonlinear root finding, equilibrium and steady-state analysis of ordinary differential equations [Computer software manual]. (R package 1.6)

Su, P. (2014). NORTARA: Generation of multivariate data with arbitrary marginals [Computer software manual]. Retrieved from <https://CRAN.R-project.org/package=NORTARA> (R package version 1.0.0)

Vale, C. D., & Maurelli, V. A. (1983). Simulating multivariate nonnormal distributions. *Psychometrika*, 48, 465–471. doi:10.1007/BF02293687

## Appendix

The algorithm is available in the NOTAMO R package, which can be downloaded at <https://github.com/NOTAMOr/NOTAMO>. The function `NORTA_function` is used to estimate a quantile mixture for a univariate distribution, given a set of inverse *CDFs* and prespecified central moments. A user can additionally specify alternative values for the accuracy vector, starting values, the maximum number of iterations, and the algorithm used to approximate the central moments. The function `prep_NORTA` converts the results to inverse *CDFs*, which are then used in the NORTA algorithm to generate samples from a multivariate distribution with prespecified correlation matrix (see the NOTAMO reference manual for additional information).

The following code example illustrates the generation of a sample from a bivariate distribution with  $r = .40$  using NOTAMO. Both univariate distributions have prespecified central moments as  $\gamma_3 = 1$  and  $\gamma_4 = 10$  (see distribution 3 in the simulation example). For the first univariate distribution, the quantile mixture consists of a standard normal distribution, an exponential distribution with rate  $\lambda = 1$ , and a cubic standard normal. For the second distribution, an additional exponential distribution with rate  $\lambda = 2$  is considered. The algorithm finds a distribution that is different from the first univariate distribution,

despite having the same skewness and kurtosis.

```
> ### Univariate distributions
>
> # Define set of three inverse CDFs (distribution 1):
> icdf_list <- list(
+   list(qexp),
+   list(qnormcube <- function(p) {return(qnorm(p)^3)}),
+   list(qnorm)
+ )
> # Define target moments:
> moms <- matrix(0,nrow=2,ncol=2)
> moms[1,] <- c(3,1)   #desired skewness is 1
> moms[2,] <- c(4,10) #desired kurtosis is 10
> # Estimate parameters:
> res1 <- NORTA_function(icdf_list,moms)
> # Define set of four inverse CDFs (distribution 2):
> icdf_list2 <- list(
+   list(qexp),
+   list(qnormcube <- function(p) {return(qnorm(p)^3)}),
+   list(qexp,rate=2),
+   list(qnorm)
+ )
> # Estimate parameters:
> res2 <- NORTA_function(icdf_list2,moms)
> # The resulting distribution is different from the first example, despite
> # having the same skewness and kurtosis.
>
```

```
> ### Multivariate distribution
>
> # Define correlation matrix:
> target_cor <- matrix(0,nrow=2,ncol=2)
> target_cor[1,] <- c(1,0.4)
> target_cor[2,] <- c(0.4,1)
> # Define functions for NORTARA:
> f1 <- function(x) {
+   return(prepare_NORTA(res1,x))
+ }
> f2 <- function(x) {
+   return(prepare_NORTA(res2,x))
+ }
> # Generate bivariate distribution with prespecified correlation matrix,
> # skewness, kurtosis, and N=100:
> genNORTARA(100,target_cor,invcdfnames = c('f1','f2'),defaultindex=c(1,2))
```

How to determine the number of factors to retain in exploratory factor analysis? A comparison of extraction methods under realistic conditions.

Max Auerswald and Morten Moshagen

Ulm University

#### Author Note

Max Auerswald and Morten Moshagen, Institute of Psychology and Education, University of Ulm, Ulm, Germany.

Correspondence concerning this article should be addressed to Max Auerswald, Research Methods, Institute of Psychology and Education, University of Ulm, Albert-Einstein-Allee 47, 89081 Ulm, Germany. E-Mail: max.auerswald@uni-ulm.de



## Abstract

Exploratory factor analyses are commonly used to determine the underlying factors of multiple observed variables. Many criteria have been suggested to inform the decision on the number of factors to retain. In this study, we present an extensive Monte Carlo simulation, varying the number of latent factors, the correlation among the factors, the number of items per factor, the magnitude of loadings, the underlying distribution, and the number of observations. We compared traditional parallel analysis (PA) with four recently suggested methods: revised PA, comparison data (CD), the Hull method, and the Empirical Kaiser Criterion (EKC). Whereas traditional PA displayed the highest hit rate (92%) overall, every other method was superior under at least some data conditions. The Hull method and the EKC outperformed traditional PA for unidimensional or orthogonal factor models with a high number of indicators per factor, especially for small sample sizes. In correlated factor designs, CD performed better than PA if the number of indicators was small, whereas revised PA performed better for a higher number of indicators per factor. Given that overall accuracy increases to 98% when traditional PA and either Hull or EKC indicate the same number of factors to retain, we suggest that investigators first apply these methods to determine the number of factors. In the remaining cases where the results of this combination rule are inconclusive, CD or traditional PA achieved the highest overall accuracy. However, disagreement also suggests that factors are in general harder to detect, increasing sample size requirements to  $N = 500$ .

*Keywords: factor analysis, number of factors, Monte Carlo simulation*

How to determine the number of factors to retain in exploratory factor analysis? A comparison of extraction methods under realistic conditions.

Exploratory factor analysis (EFA) is a widely used statistical method to study the underlying latent structure of a large number of observed variables, especially if there is no strong a priori justification for a particular theoretical model. EFA determines the underlying structure in a data-driven approach assuming a common factor model (Thurstone, 1947). In this model, each observed variable is conceptualized as the weighted sum of a set of (potentially correlated) factor variables and a single unique factor.<sup>1</sup> The common factors account for covariances among the observed variables and thus are the factors of theoretical interest. Unique factors, on the other hand, exclusively account for the variances of single observed variables, which is considered to reflect measurement error with regard to the common factors.

One of the key questions in EFA is to decide how many latent factors need to be extracted to account for covariations among the observed variables. Both under- and overestimating the number of factors (referred to as under- and overextraction, respectively) have detrimental effects on the quality of EFA (Comrey, 1978). Underextraction results in substantial error on all factor loadings, irrespective of their weight in a correctly specified model (Wood, Tataryn, & Gorsuch, 1996) and deteriorates the factor scores compared to factor scores in a correctly specified model (Fava & Velicer, 1996). In contrast, overextraction typically results in lower biases in factor scores and loadings (Fava & Velicer, 1992; Wood et al., 1996). However, overextraction can lead to factor splitting, such that manifest variables with loadings on one factor are split on multiple factors after the rotation, which drastically increases biases for loadings (Wood et

---

<sup>1</sup>Principal component analysis (PCA) is also often used as a substitute for EFA. However, in contrast to EFA, PCA is primarily a data reduction technique. If the goal of the analysis is to uncover a latent structure that addresses the covariances among observed variables measured with some random error, which is a more realistic case in psychological research, EFA is usually preferred (e.g. Bentler & Kano, 1990; de Winter & Dodou, 2016). In this article, we therefore focus on EFA.

al., 1996). Overextraction also results in less parsimonious models that include constructs with little to no explanatory value and increase the likelihood of Heywood cases, such as negative variance estimates (de Winter & Dodou, 2012). Several methods are available to determine the number of factors in EFA, such as the widely known Kaiser criterion (Kaiser, 1960), Cattell's scree test (Cattell, 1966), and parallel analysis (PA, Horn, 1965), with the latter generally being considered the state-of-the-art technique (Fabrigar, Wegener, MacCallum, & Strahan, 1999). Recently, a number of new methods were suggested (Braeken & van Assen, in press; Green, Levy, Thompson, Lu, & Lo, 2012; Lorenzo-Seva, Timmerman, & Kiers, 2011; Ruscio & Roche, 2012), each outperforming PA in at least some conditions. However, the performance of these methods has not yet been assessed in comparison with each other. The objective of this study is to fill this gap and compare four modern techniques and PA over a wide range of conditions designed to mimic typical data structures obtained in psychological research. The next section describes the common factor model and introduces the concept of eigenvalues, on which most decision criteria rely. We then present a more detailed review of popular methods and modern techniques for determining the number of factors in EFA.

### The Common Factor model

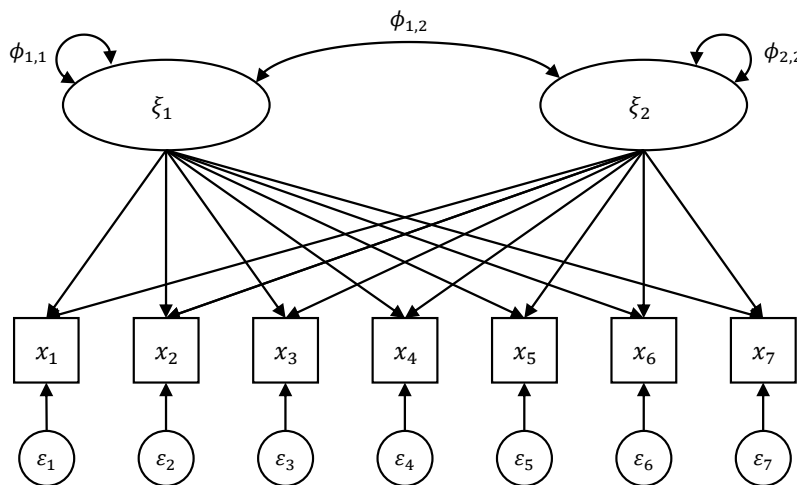
The common factor model (for an overview, see e.g. Jöreskog, 2007) assumes a set of  $m$  latent factors  $\xi_1, \dots, \xi_m$  that explain variations in the  $p$  observed (and mean-centered) random variables  $x_1, \dots, x_p$ . A single observed variable  $x_i$  is assumed to be a linear combination of  $\xi_1, \dots, \xi_m$  and one unique error  $\varepsilon_i$ , similar to a linear regression:

$$x_i = \lambda_{i1}\xi_1 + \lambda_{i2}\xi_2 + \dots + \lambda_{im}\xi_m + \varepsilon_i, \quad 1 \leq i \leq p, \quad (1)$$

where  $\varepsilon_i$  is uncorrelated with all  $\xi_1, \dots, \xi_m$  and all  $\varepsilon_{i'}$  for which  $i \neq i'$ , and  $\lambda_{ij}$  is the loading of the  $i$ -th item on factor  $j$ . The goal is thus to find latent factors, fewer in number than the number of observed variables, that account for the covariances among the observed

variables  $x_1, \dots, x_p$  such that  $x_1, \dots, x_p$  would be uncorrelated conditional on the latent factors  $\xi_1, \dots, \xi_m$ .

Figure 1 shows a common factor model for two latent factors and seven observed variables. In this case, every observed variable  $x_i$  is assumed to depend on two latent factors  $\xi_1, \xi_2$  and the unique error  $\varepsilon_i$  for  $1 \leq i \leq 7$ . The latent factors that are supposed to represent the underlying psychological variables of interest can be correlated. The unique errors measure item-specific variance and are thus assumed to be independent from both the latent factors and other item-specific errors. The loadings are estimated in conjunction with the variances of the unique variables and the (co-)variances of the latent factors. For continuous variables, the estimation is most often based on maximum likelihood or unweighted least squares.



*Figure 1.* A common factor model with two latent factors and seven observed variables. The latent factors can be correlated whereas the unique errors are independent from other unique errors and the latent factors. The arrows from the latent factors to the observed variables indicate the loadings  $\lambda_{1i}, \lambda_{2i}$  for  $1 \leq i \leq 7$  (see Equation 1).

The common factor model can also be denoted in matrix notation. For

$\mathbf{X} = (x_1, \dots, x_p)^T$ ,  $\xi = (\xi_1, \dots, \xi_m)^T$ ,  $\varepsilon = (\varepsilon_1, \dots, \varepsilon_p)^T$ , and a  $p \times m$  matrix of loadings  $\mathbf{\Lambda}$ ,

$$\mathbf{X} = \mathbf{\Lambda} \xi + \varepsilon \quad (2)$$

is an equivalent expression of Equation 1. We can express the covariance matrix of the observed variables  $\mathbf{X}$  as

$$\mathbf{\Sigma} = \mathbf{E}(\mathbf{X}\mathbf{X}^T), \quad (3)$$

where  $\mathbf{E}$  is the expected value, because the observed variables are mean-centered. From Equation 2 follows

$$\mathbf{X}\mathbf{X}^T = (\mathbf{\Lambda} \xi + \varepsilon)(\mathbf{\Lambda} \xi + \varepsilon)^T \quad (4)$$

$$= \mathbf{\Lambda} \xi \xi^T \mathbf{\Lambda}^T + \mathbf{\Lambda} \xi \varepsilon^T + \varepsilon \xi^T \mathbf{\Lambda}^T + \varepsilon \varepsilon^T. \quad (5)$$

We denote the covariance matrix of  $\xi$  as  $\Phi (= \mathbf{E}(\xi\xi^T))$  and the covariance matrix of  $\varepsilon$  as  $\Delta (= \mathbf{E}(\varepsilon\varepsilon^T))$ . Since  $\xi$  and  $\varepsilon$  are independent, the model expresses the covariance matrix as

$$\hat{\mathbf{\Sigma}} = \mathbf{\Lambda} \Phi \mathbf{\Lambda}^T + \Delta. \quad (6)$$

The common factor model thus becomes a statement about the covariance matrix, where the matrices  $\mathbf{\Lambda}$  and  $\Phi$  are only determined up to a rotation (for details, see e.g. Browne, 2001).

The matrix  $\Delta$  in Equation 6 is a diagonal matrix, because the common factor model assumes that all unique errors  $\varepsilon_i, \varepsilon_{i'}, 1 \leq i, i' \leq p$ , are independent for  $i \neq i'$ . The entries  $\delta_i$  of  $\Delta$  are called uniqueness and represent the part of variance of the observed variable  $x_i$  that is independent of the latent factors. The communalities are their counterpart, the part of the variance of  $x_i$  that can be explained by the latent factors.<sup>2</sup> If we consider the correlation

<sup>2</sup>The problem of communalities refers to the difficulty of simultaneously estimating the proportion of variance that can be explained by common factors and the common factor model itself. The common factor model approximates a correlation matrix with communalities on the diagonal, but the communalities are only known after the model is estimated (see e.g. Harman, 1976).

matrix  $\mathbf{R}$  of the observed variables, the common factor model estimates  $\mathbf{\Lambda}$  such that

$$\hat{\mathbf{R}}_{\mathbf{C}} \approx \mathbf{\Lambda}\mathbf{\Lambda}^T, \quad (7)$$

where  $\hat{\mathbf{R}}_{\mathbf{C}}$  is the correlation matrix with communalities on the diagonal. One least squares solution to Equation 7 estimates the loadings in  $\mathbf{\Lambda}$  proportional to the so called eigenvectors of  $\hat{\mathbf{R}}_{\mathbf{C}}$  (Jöreskog, 2007).<sup>3</sup> In general, eigenvectors are vectors  $v$  for which

$$\mathbf{A}v = lv, \quad v \neq 0 \quad (8)$$

holds for an arbitrary square matrix  $\mathbf{A}$  of size  $p \times p$ ,  $v$  a vector of length  $p$ , and  $l$  a scalar, the corresponding eigenvalue. Symmetric, positive semidefinite matrices like covariance matrices or  $\mathbf{R}_{\mathbf{C}}$  always have  $p$  (not necessarily distinct) non-negative eigenvalues. Most importantly, the  $j$ -th largest eigenvalue of  $\mathbf{R}_{\mathbf{C}}$  corresponds to the explained variance of the  $j$ -th factor in a common factor model (see the Appendix for a more technical explanation of this fact).

### Methods to Decide on the Number of Factors to Retain

As an exploratory technique, EFA is typically used whenever there is no strong theoretical reason to expect a particular number of latent factors underlying the observed variables. In this section, we briefly revisit conventional methods and introduce modern techniques that attempt to inform the decision on the number of factors to retain in EFA.

#### Kaiser criterion

One of the most prominent heuristics to decide on the number of factors to retain is the Kaiser criterion (Kaiser, 1960), which extracts all factors with corresponding eigenvalues greater than 1. The rationale behind this rule is that a factor should at least explain as much variance as a single item. However, because sampling error leads to eigenvalues that exceed 1 even in the absence of any factor, the Kaiser criterion severely

---

<sup>3</sup>The solution in fact minimizes  $tr(\hat{\mathbf{R}}_{\mathbf{C}} - \mathbf{\Lambda}\mathbf{\Lambda}^T)^2$

overextracts the number of factors (e.g. Cattell & Vogelmann, 1977; Hakstian, Rogers, & Cattell, 1982; Lance, Butts, & Michels, 2006; Zwick & Velicer, 1986). For example, in a simulation study conducted by Ruscio and Roche (2012), the suggested number of factors for the Kaiser criterion was biased by more than seven factors across all conditions. Despite this substantial tendency to overestimate the number of factors, the Kaiser criterion is commonly used (Henson & Roberts, 2006) and the default in several statistics programs such as SPSS (IBM Corp., 2015).

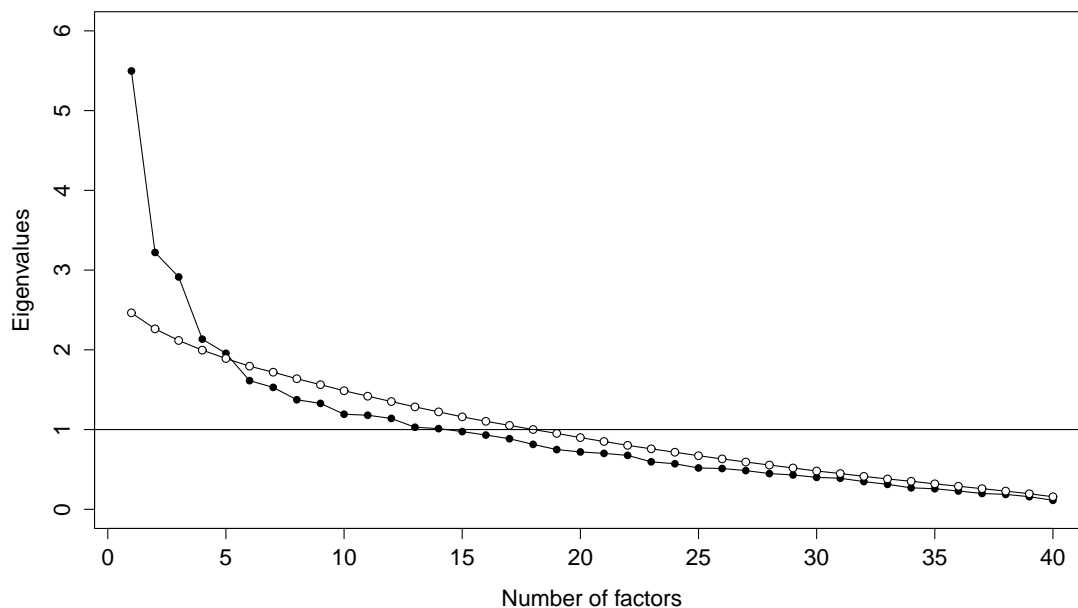
### **Cattell's scree test**

Cattell's (1966) scree test is a graphical method based on the plot of the successive eigenvalues in descending order (the so-called scree plot). The test is performed by searching for an elbow, a point at which the eigenvalues decrease abruptly. The method suggests to extract all factors up to the factor corresponding to the eigenvalue preceding the sharpest decline. Being a graphical approach, the method is obviously subjective and therefore rarely evaluated systematically. Furthermore, scree plots can be ambiguous, either lacking any clear elbow or showing multiple elbows in the same scree plot (Ruscio & Roche, 2012). Raïche, Riopel, and Blais (2006) suggested the optimal coordinate and acceleration factor criteria, which are non-graphical solutions for Cattell's scree test that rely on the change in slope of adjacent eigenvalues. Both methods clearly outperformed the Kaiser criterion, but tended to underestimate the number of factors and were thus still inferior to other approaches, such as PA (Raïche, Walls, Magis, Riopel, & Blais, 2013; Ruscio & Roche, 2012).

### **Traditional and revised parallel analysis**

PA (Horn, 1965) compares the empirical eigenvalues to the mean of eigenvalues obtained from random samples based on uncorrelated variables. The random samples have the same number of observations and variables as the empirical data, so that the eigenvalues of the random samples take sampling error into account. PA extracts all factors

with eigenvalues that exceed the average corresponding eigenvalue of the random samples (see Figure 2 for an example showing the eigenvalues of random independent data and a simulated sample with five underlying factors).



*Figure 2.* Parallel analysis on a simulated sample with  $N = 100$ , 40 manifest variables, and 5 underlying factors. The filled dots represent the sorted eigenvalues of the sample correlation matrix. The empty dots represent the average eigenvalues of correlation matrices from 100 independent random samples. The solid line depicts the threshold for the Kaiser criterion. Parallel analysis correctly identifies the number of factors as five, while the scree test suggests either one or three. The Kaiser criterion suggests 14 factors and thus overestimates the number of factors severely.

The eigenvalues in PA are typically based on the correlation matrix of observed and random samples (e.g. Finch & West, 1997; Steger, 2006), similarly to a principal component analysis (PA-PCA), but can also be based on the correlation matrix with communalities on the diagonal, reflecting a common factor model (Humphreys & Ilgen, 1969). However, Garrido, Abad, and Ponsoda (2013) argued that the common factor model



is inappropriate for PA, because the random samples have uncorrelated variables with communalities of  $h^2 = 0$  in the population, while the common factor model assumes a common cause behind the observed variables. In their simulation study, Garrido et al. (2013) also found higher hit rates for PA-PCA compared to PA with minimum rank factor analysis. Furthermore, the performance of PA is also affected by the method of estimating the communalities. Crawford et al. (2010) found a higher hit rate for PA-PCA unless factors were moderately or highly correlated, compared to PA where communalities are estimated as sample multiple  $R^2$  between the variables and all remaining variables. Overall, PA based on PCA seems to produce better results than PA based on a common factor model. Despite the differences between PCA and EFA (e.g. Fabrigar et al., 1999), the number of common factors directly influences the distribution of eigenvalues of the correlation matrix (Braeken & van Assen, in press). Therefore, PA-PCA can also be used as a criterion for the number of factors, even if an EFA is performed.

PA is supported by strong evidence from simulation studies (Hubbard & Allen, 1987; Humphreys & Montanelli, 1975; Peres-Neto, Jackson, & Somers, 2005; Velicer, Eaton, & Fava, 2000; Zwick & Velicer, 1986) and is generally considered to be the method of choice (e.g. Fabrigar et al., 1999). However, there are two weaknesses associated with PA as suggested by Horn (1965). The first weakness stems from the fact that sampling error can lead to eigenvalues above the average eigenvalue of random samples. For example, if all manifest variables are uncorrelated in the population, the first empirical eigenvalue would exceed the first average eigenvalue from random samples in approximately 50% of all samples, which would in turn lead to a tendency to overestimate the number of factors for PA. One possible solution is to use the 95th percentile of the eigenvalues obtained from random samples as a threshold instead of the mean as in traditional PA (Glorfeld, 1995).

The second weakness of PA involves the choice of the reference eigenvalues for the second and following factors (Turner, 1998). Assume that the empirical data set has one underlying factor that explains a large portion of the item covariances. Any remaining

factor can only explain a fraction of the yet unexplained covariances. However, the items in the random samples that constitute the comparison threshold are uncorrelated, leading to a higher portion of unexplained covariance for the random samples, compared to the empirical sample. This biased comparison due to differing portions of unexplained covariance might leave a second factor undetected. Cho, Li, and Bandalos (2009) showed that both weaknesses tend to partially counteract each other, as PA was more accurate if the average eigenvalues were used as a criterion, compared to the 95th percentile. However, there is no guarantee that these deficiencies have effects to the same extent but in opposing directions. As a remedy, Turner (1998) suggested that the random eigenvalues should be calculated from samples that also have underlying factors equivalent to the factors already extracted in the empirical data. Green et al. (2012) implemented this idea and demonstrated that this revised PA outperforms traditional PA with the 95th percentile as a criterion for highly correlated factors and large loadings (see also Green, Thompson, Levy, & Lo, 2015).

### **Comparison data**

Ruscio and Roche (2012) suggested the comparison data (CD) approach that, similar to revised PA, also takes previous factors into account by generating comparison data of known factorial structure. The CD method finds the number of factors by determining the solution that reproduces the pattern of eigenvalues best. Although both CD and revised PA iteratively compare factor solutions with  $j - 1$  and  $j$  factors, CD differs from revised PA as suggested by Green et al. (2012) in three respects. First, revised PA only compares the eigenvalue of the  $j$ th factor with the  $j$ th eigenvalue of the sampled data, whereas CD always takes all eigenvalues into account. Specifically, CD compares the root mean square error of the difference of all empirical eigenvalues to the eigenvalues of sampled data with underlying factors and tests if the difference becomes significantly smaller when another factor is included. If too many factors are extracted, the eigenvalues for all subsequent

eigenvalues will be lower in the sample data compared to the empirical data, leading to higher misfit in the overall pattern of eigenvalues. The second difference between CD and revised PA concerns the chosen reference value for the sample eigenvalues. CD relies on the average of the sample eigenvalues, which usually lies below the 95th percentile of the eigenvalues as used in revised PA. These two differences of CD and revised PA have different implications for the tendency to under- or overextract factors. The different number of eigenvalues taken into account should lead to less overextractions of CD compared to both traditional and revised PA. However, the lower reference value used in CD should lead to more overextractions compared to revised PA. The third difference between CD and revised PA pertains to the used sampling procedure. Revised PA generates random normally distributed samples with the underlying factor structure. The CD approach, however, reproduces the marginal distributions observed in the empirical data set using an algorithm suggested by Ruscio and Kaczetow (2008). Therefore, CD should be more accurate when data are not normally distributed.

Ruscio and Roche (2012) compared the performance of CD to traditional PA and other methods such as the Kaiser criterion. The loadings in this simulation were set to create challenging conditions for traditional PA, which led to exceptionally low loadings for single factor models ( $\bar{\lambda} = 0.225$ ) and models with uncorrelated factors ( $.275 \leq \bar{\lambda} \leq .425$ ). Overall, CD identified the number of factors more accurately than traditional PA, unless the number of factors was high.

### **Hull method**

The Hull method (Lorenzo-Seva et al., 2011) is an approach based on the Hull heuristic, used in other areas of model selection (e.g. Ceulemans & Kiers, 2006). Similar to non-graphical variants of Cattell's scree plot, the Hull method attempts to find an elbow as justification for the number of common factors. However, instead of using the eigenvalues relative to the number of factors, the Hull method relies on goodness-of-fit indices relative

to the model degrees of freedom of the proposed model. More specifically, the method finds the number of factors in four steps:

1. The method calculates a goodness-of-fit index  $GOF_j$  and model degrees of freedom  $df_j$  of various models with an increasing number of factors  $j$  up to a prespecified maximum  $J$  ( $0 \leq j \leq J$ ).
2. A solution  $s_j$  is considered to be unviable if a less complex model (indicating a lower number of factors) with a higher (better) fit index exists. The  $j$ -th solution is thus unviable if there is a solution  $s_{j'}$  with  $j' < j$  and  $GOF_{j'} > GOF_j$ .
3. The remaining solutions are further identified as unviable if  $GOF_j$  is below the line connecting the adjacent viable solutions in a plot of fit indices and model degrees of freedom. This step is repeated until no remaining solutions can be identified as unviable.
4. The Hull method then suggests the number of factors where

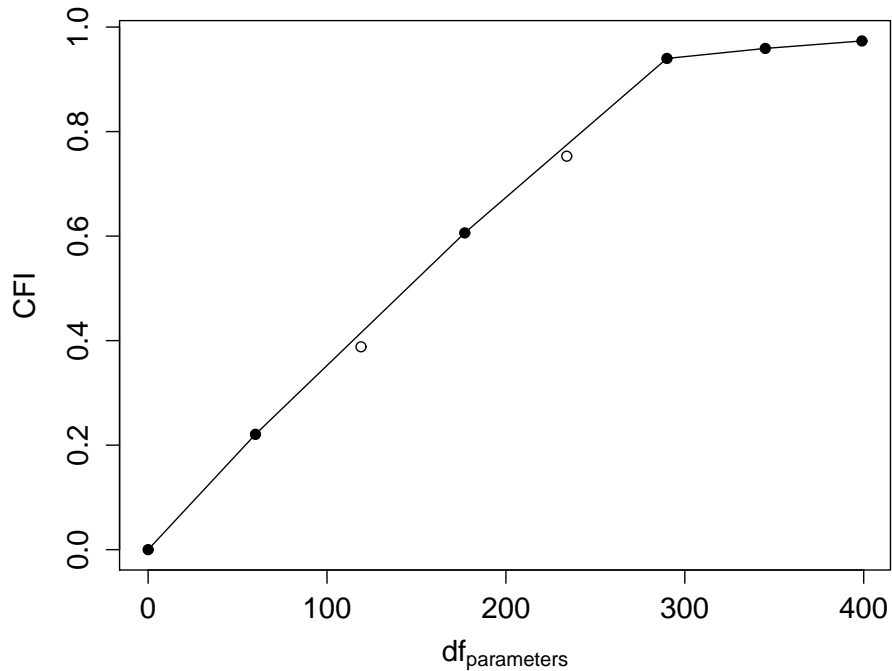
$$\frac{(GOF_j - GOF_{j-1})/(df_j - df_{j-1})}{(GOF_{j+1} - GOF_j)/(df_{j+1} - df_j)} \quad (9)$$

obtains its maximum and  $j$  is a viable solution.<sup>4</sup>

The elbow is identified as the value where, relative to the change in the model  $df$ , model fit increases considerably compared to a lower number of factors ( $j - 1$ ) but increases barely compared to a higher number of factors ( $j + 1$ ). This criterion value is based on every viable fit value relative to both its preceding and subsequent fit values (see Figure 3 for an example). Note that the suggested factor solution therefore cannot be the first or last factor in the range for which the model fit is estimated (unless all other solutions are unviable). This range typically includes a zero factor model as a minimum. In order to avoid overextractions, the suggestion for the maximum is the number of factors extracted based on traditional PA plus one (Lorenzo-Seva et al., 2011).

---

<sup>4</sup>Note that  $j - 1$  and  $j + 1$  are not required to be viable solutions.



*Figure 3.* The Hull method with the comparative fit index (CFI) as a criterion on a simulated sample with five true underlying factors. In this case, the Hull method considers solutions in the range from zero to seven factors. The empty dots are unviable solutions that lie below the line connecting adjacent viable solutions. The filled dots represent viable solutions. The Hull method correctly identifies five factors.

Lorenzo-Seva et al. (2011) compared the Hull method with various goodness-of-fit indices to other selection criteria. The design of the simulation study incorporated both major and minor factors, where major factors constituted the factors of interest. Minor factors were associated with (random) loadings that accounted for 15% of the variance on average and thus represent a comparatively small, systematic error that factor extraction criteria should disregard. While no method consistently outperformed the other approaches across all conditions, the Hull method based on the comparative fit index (CFI, Bentler, 1990) improved upon other methods, including traditional PA, in data conditions where

the number of observed variables or the sample size was large. However, the method has not yet been compared to other variants of PA (Green et al., 2012) or the CD approach (Ruscio & Roche, 2012). Compared to other approaches, the Hull method seems especially suited if the goal is to extract comparatively strong, unambiguous factors, because it successfully ignores small, systematic errors. We therefore expect that the Hull method is particularly useful in the case of single factor models or models with uncorrelated factors, but may fall short, by design, when factors are highly correlated or when some factors account for a small proportion of the variance only.

### Empirical Kaiser criterion

The Empirical Kaiser Criterion (EKC, Braeken & van Assen, in press) is an approach that incorporates random sample variations of the eigenvalues in Kaiser's criterion. On a population level, the criterion is equivalent to Kaiser's criterion and extracts all factors with associated eigenvalues of the correlation matrix greater than one. However, on a sample level, the criterion takes the distribution of eigenvalues for normally distributed data into account. Under the null model, the distribution of eigenvalues asymptotically follows a Marčenko-Pastur distribution (Marčenko & Pastur, 1967). The resulting upper bound of this distribution (the highest value with non-zero density) is the reference value for the first eigenvalue  $l$ , so

$$l_{1,ref} = \left(1 + \sqrt{\frac{p}{N}}\right)^2, \quad (10)$$

for  $N$  observations and  $p$  items. Subsequent eigenvalues are corrected by the explained variance, expressed as the eigenvalues of previous factors. The  $j$ -th reference eigenvalue is

$$l_{j,ref} = \max\left(\frac{p - \sum_{i=0}^{j-1} l_i}{p - j + 1} \left[1 + \sqrt{\frac{p}{N}}\right]^2, 1\right), \quad (11)$$

such that higher previous eigenvalues lower the reference eigenvalue since the proportion of unexplained variance will be lower. In accordance with the original Kaiser criterion, the reference eigenvalue cannot become smaller than one.

Braeken and van Assen (in press) derived theoretical conditions for scale reliability, number of observations, number of factors, and factor correlation, under which the EKC is expected to correctly identify the number of factors. For example, for orthogonal factors, EKC is expected to work if

$$\frac{p_j}{p_j - (p_j - 1)\alpha_j} > \left(1 + \sqrt{\frac{p_j m}{N}}\right)^2, \quad (12)$$

for all  $1 \leq j \leq m$  and  $m$  (overall) underlying factors,  $\alpha_j$  Cronbach's alpha in the population,  $p_j$  the number of items of the respective factor  $j$ , and  $N$  observations. Thus, EKC is especially suited for shorter scales with high reliability. For correlated factors, the conditions that guarantee a high performance for EKC are more complex, but are also more likely to be fulfilled if  $\alpha$  and  $N$  are high, scales are shorter, and factor correlations are low. Corroborating these assumptions, Braeken and van Assen (in press) found that the EKC exhibited a very high hit rate if these conditions were met (.97), but a low hit rate if they were not (.17). In particular, Braeken and van Assen (in press) found that the EKC outperforms traditional PA when factors are correlated and are only measured by few items with very high loadings, yielded comparable results to improved PA and CD in a simulation study with a high number of factors and few observed variables, but that no method outperformed all other methods under all conditions. However, EKC has not yet been compared to improved PA or CD in a more general simulation study that also included the Hull method. In addition, the theoretical conditions guaranteeing a high performance of the EKC require information that is not available to researchers prior to conducting an EFA, so that researchers cannot know in advance whether the EKC can be expected to perform well in their particular analysis scenario.

### **The Present Study**

The goal of the present study is to evaluate the performance of modern techniques for determining the number of factors to retain in EFA. We incorporated a wide range of data conditions that are challenging but realistic in psychological research (Fabrigar et al.,

1999). This allows for assessing the overall performance of factor extraction criteria under conditions relevant in practice. Given that previous simulation studies found that no single method was superior to all other methods under all conditions (Braeken & van Assen, in press; Green et al., 2012; Lorenzo-Seva et al., 2011; Ruscio & Roche, 2012), we (1) focus on identifying conditions under which a particular method performs well and (2) attempt to suggest a combination rule based on information that is available to researchers, in turn allowing for utilizing the strengths of different methods.

## Method

### Extraction Criteria

We considered five methods for determining the number of factors to retain. These include traditional PA as the most often recommended approach and four more recently proposed methods that have been demonstrated to improve upon traditional PA under at least some conditions. We do not consider approaches that have been consistently shown to perform worse than traditional PA (such as the Kaiser criterion or the non-graphical scree plot criteria by Raïche et al., 2013).

**Traditional parallel analysis (PA).** Traditional PA (Horn, 1965) extracts all factors with eigenvalues above the eigenvalues calculated from 100 random samples. Following the suggestions of previous simulation studies, we used the average eigenvalues (Cho et al., 2009) calculated by eigenvalue decomposition of the respective correlation matrix (Garrido et al., 2013) as a criterion.<sup>5</sup> The random samples are generated by (non-parametrically) resampling the input data.

**Revised PA.** Revised PA (Green et al., 2012) sequentially compares the  $j$ th eigenvalue to the 95th percentile of eigenvalues calculated from random samples with  $j - 1$

---

<sup>5</sup>We also calculated the results for traditional PA using the 95th percentile of random eigenvalues and eigenvalues of a common factor model as criteria. The results indicated lower hit rates for both the 95th percentile and PA based on a common factor model. Due to the similarities in methods, we only report the results for PA-PCA based on the average eigenvalue.



underlying factors. As recommended by Green et al. (2012), we used the eigenvalues obtained from an EFA and set the number of random samples to 100.

**Hull method.** The Hull method (Lorenzo-Seva et al., 2011) was implemented using the CFI (Bentler, 1990) to assess the fit of each factor solution. The CFI-based Hull method was superior to every other implementation of the Hull method in the initial simulation study by Lorenzo-Seva et al. (2011).

**Comparison Data (CD).** CD (Ruscio & Roche, 2012) was implemented using an alpha level of .30 and 500 resamples, in line with the recommendations of Ruscio and Roche (2012).

**Empirical Kaiser Criterion (EKC).** The EKC (Braeken & van Assen, in press) was implemented using the eigenvalues of the input correlation matrix.

### Experimental conditions

In realizing the conditions for the simulation study, we attempted to cover a wide range of data conditions plausibly occurring in empirical factor analysis studies. Accordingly, we orthogonally manipulated six independent variables, viz. the number of observations, the number of latent factors, the latent factor correlation, the number of items per factor, the average loading magnitude, and the underlying factor and error distribution.

**Number of observations.** The number of observations was set to 100, 200, 500, or 1,000, thereby covering the sample sizes used in most empirical studies (DiStefano & Hess, 2005; Fabrigar et al., 1999; Jackson, Gillaspay, & Purc-Stephenson, 2009; Worthington & Whittaker, 2006). The condition involving  $N = 1,000$  was included to allow for drawing conclusions about the large sample performance of the approaches under scrutiny.

**Number of latent factors.** Manifest variables were generated with 1, 3, or 5 underlying factors, representing the dimensionality of scales most common in psychometric measurement (DiStefano & Hess, 2005; Jackson et al., 2009).

**Factor intercorrelation.** The intercorrelation among latent factors was set to 0, .25, or .50. Note that we did not include a condition with very high latent correlations,

because extraction decisions in this case primarily depend on theoretical reasoning rather than statistical analysis.

**Items per latent factor.** We examined 4, 8, or 12 items per latent factor. While the majority of scales in psychological assessment comprise 4 to 8 items (DiStefano & Hess, 2005; Fabrigar et al., 1999; Jackson et al., 2009), factor extraction criteria are especially important in the initial development of a measurement instrument. The process of constructing a scale typically involves the elimination of items, so a condition involving 12 items per factor was realized to represent a scale before the elimination process. In conjunction with the manipulated number of latent factors, the total number of items thus ranged from 4 (4 items per latent factor with 1 latent factor) to 60 (12 items per latent factor with 5 latent factors).

**Loading magnitude.** The standardized loadings of the observed variables on the latent factors was set to either (.65, .55, .45, .35) or (.8, .7, .6, .5) for each set of 4 variables (i.e., every loading was assigned three times when a factor was measured by 12 items). The resulting average loadings therefore were .50 or .65, which is typical for psychological research (DiStefano & Hess, 2005). The resulting Cronbach's  $\alpha$  estimates of internal consistency are presented in Table 1. In empirical research, Cronbach's  $\alpha$  is typically between .70 and .89 for published scales (Fabrigar et al., 1999) and thus slightly higher than in this simulation study. Since factor extraction criteria are again often used before item elimination, Cronbach's  $\alpha$  is likely smaller in the development stage when EFA is applied.

Table 1

*Population Cronbach's  $\alpha$  used in the simulation study*

	Number of items per factor		
Average Loading	4	8	12
.5	.57	.73	.80
.65	.75	.85	.90

**Underlying distribution.** Three types of distributions were realized (normal, non-normal based on non-normal errors, non-normal based on non-normal latent factors). Normally distributed data were generated using Cholesky decomposition. Non-normal distributions were generated from a structural model creating non-normality according to the linking functions approach by Auerwald and Moshagen (2015). The linking functions approach generates observed non-normal data by applying non-linear linking functions to the latent part or to the error part (or both). In the present study, we either incorporated non-normal latent factors and normal errors, or vice versa. These types of non-normal distributions were realized in light of evidence indicating that the performance of factor-based models may vary depending on whether non-normality in the observed variables arises from non-normal factors or from non-normal errors (Auerwald & Moshagen, 2015; Foldnes & Grønneberg, 2015; Mair, Satorra, & Bentler, 2012). We used the following linking functions:

- $f_1(x) = x^5 + x^3$
- $f_2(x) = e^{2x}$
- $f_3(x) = \begin{cases} \sqrt{x}, & \text{for } x > 0 \\ -x^2, & \text{for } x \leq 0 \end{cases}$
- $f_4(x) = \begin{cases} -50, & \text{for } x < -3 \\ -1, & \text{for } -3 \leq x < 0 \\ 1, & \text{for } 0 \leq x < 3 \\ 50, & \text{for } x \geq 3 \end{cases}$

When a factor was indicated by 8 or 12 items, each linking function was assigned two or three times (as was done for loading magnitudes, see above). This set of linking functions resulted in non-normal distributions exhibiting an average skewness of

$$\gamma_{3,f_1} = 0 (SD_{\gamma_{3,f_1}} = 0.93), \gamma_{3,f_2} = 0.14 (SD_{\gamma_{3,f_2}} = 0.59), \gamma_{3,f_3} = -0.77 (SD_{\gamma_{3,f_3}} = 0.81),$$

$\gamma_{3,f_4} = 0$  ( $SD_{\gamma_{3,f_4}} = 0.84$ ) in both non-normality conditions. Kurtosis was estimated to be approximately  $\gamma_4 = 12$  for all linking functions and non-normality conditions ( $SD_{\gamma_{4,f_1}} = 11.49$ ,  $SD_{\gamma_{4,f_2}} = 10.97$ ,  $SD_{\gamma_{4,f_3}} = 9.24$ ,  $SD_{\gamma_{4,f_4}} = 4.65$ ). The realized levels of skewness and kurtosis are well within in the boundaries commonly occurring in psychological assessment, without being overly extreme. For example, Cain, Zhang, and Yuan (in press) reported the 95th percentiles across 194 real data samples of empirical skewness and kurtosis to be 2.77 and 12.48, respectively. Other studies report ranges from 1.3 to 40.37 for kurtosis (Micceri, 1989) or  $-2.49$  to 2.33 for skewness and 1.08 to 10.41 for kurtosis, for empirical studies with small sample sizes (Blanca, Arnau, López-Montiel, Bono, & Bendayan, 2013).

### Data generation and Analysis

In total, the design involved 4 (number of observations) x 3 (number of latent factors) x 3 (factor correlation) x 3 (number of indicators per factor) x 2 (loading magnitude) x 3 (underlying distribution) = 648 conditions. For every condition, 500 independent random samples were generated, leading to a total of 324,000 data sets. The data sets were analyzed by all five extraction methods under scrutiny.

Analyses were performed in the statistical computing language R (R Core Team, 2016) using the parallel package to take advantage of multicore processing. All EFA methods used maximum likelihood estimation based on the package psych (Revelle, 2015). For the Hull method, we calculated the CFI using the  $\chi^2$  provided by the psych package. We used R code provided by Ruscio and Roche (2012) for the CD approach and a custom implementation of traditional PA and the EKC.

We recorded the suggested number of factors for each simulated data set and each method as well as their bias to over- or underextract. Bias was defined as the number of suggested factors minus the actual number of factors in the population. Thus, negative values indicate underextraction, positive values indicate overextraction, and zero indicates

no bias.

## Results

We first present the results for the overall performance across conditions as indicated by the percentage of correctly identified factors and over- or underextraction bias. We then evaluate the performance for designs with only one underlying factor, multiple orthogonal factors, and multiple correlated factors. We report estimates of saturated logistic regressions predicting whether the respective method suggested the correct number of factors to quantify the effects. In these logistic regressions, all applicable conditions were effect-coded with the following reference categories: traditional PA,  $N = 500$ , 3 latent variables, orthogonal factors, 8 indicators per factor, average loadings of .5, and normal distribution. Thus, an Odds-Ratio ( $e^\beta$ ,  $OR$ ) of 2 would indicate that the odds of identifying the correct number of factors in this specific condition are twice as high than the grand mean and all else being equal. Furthermore, we computed a linear regression model predicting the extraction biases. We report all main effects and simple interactions with  $|\beta| > .40$  ( $OR < 0.67$  or  $OR > 1.49$ ) in the general logistic regression and with  $|\beta| > .2$  in the linear regression.

Finally, we assessed the performance of combination rules as we assumed that no method would outperform every other method in all conditions. In doing so, we relied on parameters known to investigators (i.e., the sample size, the average correlation among the observed variables, and the number of observed variables) to improve overall performance.

### Overall performance

Table 2 shows the percentage of successfully identified number of factors as a function of condition and extraction method used. As can be seen, all methods had a moderate to high overall success rate. With an overall hit rate of 92%, traditional PA performed best across all conditions. Hull and EKC performed above average ( $OR_{Hull} = 1.95$ ,  $OR_{EKC} = 2.21$ ), whereas revised PA and CD performed below average ( $OR_{CD} = 0.22$ ,

$OR_{PA-R} = 0.29$ ). Averaged across all other conditions, all methods were well-behaved in the sense that an increase in the number of observations increased performance ( $OR_{N=100} = 0.22$ ,  $OR_{N=200} = 0.82$ ,  $OR_{N=1000} = 2.66$ ). This effect was more pronounced for EKC (interaction terms were  $OR_{N=100,EKC} = 0.49$ ,  $OR_{N=1000,EKC} = 1.60$ ) and less pronounced for revised PA (interaction terms were  $OR_{N=100,PA-R} = 2.92$ ,  $OR_{N=1000,PA-R} = 0.49$ ). In contrast to our expectations, all methods were unaffected by the underlying distribution for both non-normal latent ( $OR_{Lat-NN} = 0.74$ ) and non-normal error variables ( $OR_{Err-NN} = 1.45$ ;  $|\beta| < 0.28$  for all interactions).

In general, performance of all methods increased with the number of indicators per factor ( $OR_{\#x=4} = 0.22$ ,  $OR_{\#x=12} = 2.54$ ), especially for improved PA ( $OR_{\#x=4,PA-R} = 0.37$ ), and with the magnitude of loadings ( $OR_{\lambda=.65} = 1.58$ ;  $|\beta| < 0.27$  for all interactions), reflecting that factor recovery improves with factor determination. In contrast, the overall performance decreased considerably when the number of underlying factors increased ( $OR_{\#\xi=1} = 4.39$ ,  $OR_{\#\xi=5} = 0.28$ ). Except for CD ( $OR_{\#\xi=1,CD} = 0.25$ ), all methods showed their highest performance for unidimensional factor models ( $OR_{\#\xi=1,Hull} = 1.72$ ,  $OR_{\#\xi=1,PA-R} = 2.86$ ). CD was the only method to improve with 5 factors, compared to the grand mean ( $OR_{\#\xi=5,CD} = 1.66$ ). Expectably, the difficulties to correctly identify multiple factor models increased with the factor correlation. Performance of all methods was worse when the factor correlation was high ( $OR_{\rho=.5} = 0.36$ ), but improved when factor correlations were small ( $OR_{\rho=.25} = 1.43$ ), compared to the grand mean. Concerning differences between extraction methods, CD was most reliable when factor correlations increased ( $OR_{\rho=.50,CD} = 2.30$ ), whereas Hull and EKC performed worse under this condition ( $OR_{\rho=.50,Hull} = 0.53$ ,  $OR_{\rho=.50,EKC} = 0.50$ ), which is in line with our expectations.

Finally, Table 2 also shows the performance of all methods in conditions for which EKC is predicted to perform well, based on theoretical expectations (Braeken & van Assen, in press). EKC indeed performed considerably better when these conditions were met, but other methods improved as well, albeit to a lesser extent. Notably, traditional PA still

outperformed EKC even under these conditions.

Table 3 displays the average bias for each method, calculated as the number of extracted factors minus the number of correct factors. Except for CD, which was virtually unbiased on average ( $\overline{bias}_{CD} = 0.00$ ), all methods showed an underextraction bias ( $\overline{bias}_{PA-T} = -0.09$ ,  $\overline{bias}_{PA-R} = -0.32$ ,  $\overline{bias}_{Hull} = -0.42$ ,  $\overline{bias}_{EKC} = -0.31$ ). This underextraction bias was larger for all methods when the sample size was small ( $\beta_{N=100} = -0.26$ ,  $\beta_{N=200} = -0.04$ ,  $\beta_{N=1000} = 0.18$ ; all interactions with  $|\beta| < 0.17$ ). The underlying distribution only had a small effect, slightly increasing the number of extracted factors overall ( $\beta_{Lat-NN} = 0.01$ ,  $\beta_{Err-NN} = 0.07$ , all interactions with  $|\beta| < 0.05$ ). Given that the distributional properties affected neither the accuracy nor the bias to a substantial degree, we excluded this factor in the following regressions.

As in the case for overall accuracy, lower loadings increased underextraction biases for all methods ( $\beta_{\lambda=.65} = 0.10$ , all interactions with  $|\beta| < 0.17$ ), as did a smaller number of indicators per factor ( $\beta_{\#x=4} = -0.29$ ,  $\beta_{\#x=12} = 0.18$ ), especially for revised PA ( $\beta_{\#x=4,PA-R} = -0.43$ ,  $\beta_{\#x=12,PA-R} = 0.22$ ). Only CD was again unbiased when the number of indicators per factor was small ( $\beta_{\#x=4,CD} = 0.27$ ). Extraction biases were strongly affected by the true number of underlying factors, with unidimensional factor models being associated with a slight overextraction bias ( $\beta_{\#\xi=1} = 0.27$ ) and models with a large number of factors leading to underextraction ( $\beta_{\#\xi=5} = -0.32$ ), especially when analyzed with the Hull method ( $\beta_{\#\xi=5,Hull} = -0.25$ ). Underextraction biases further increased with the correlation between the factors ( $\beta_{\rho=.50} = -0.31$ ), particularly for the Hull method ( $\beta_{\rho=.50,Hull} = -0.25$ ), whereas small factor correlations again only had a small effect ( $\beta_{\rho=.25} = 0.09$ ).

Table 2

*Percentage of correctly identified number of factors*

Condition	Level	PA-T	PA-R	Hull	CD	EKC
Number of latent variables	1	99	92	100	81	100
	3	93	71	85	85	84
	5	85	68	73	68	70
Number of observations	100	81	71	71	61	67
	200	92	76	83	77	82
	500	97	80	93	86	93
	1,000	99	82	97	87	97
Items per latent variable	4	85	50	76	64	71
	8	95	91	89	82	89
	12	97	91	93	88	94
Underlying distribution	Normal	92	73	84	82	82
	Lat-NN	91	77	85	74	84
	Err-NN	94	82	89	78	88
Intercorrelation	0	97	78	95	79	93
	.25	95	70	88	80	84
	.5	75	60	55	71	54
Average Loading	.50	90	70	82	75	81
	.65	95	84	90	80	89
EKC guarantee	yes	98	84	94	84	94
	no	52	26	28	36	17
Overall		92	77	86	78	85

*Note.* PA-T = traditional parallel analysis, PA-R = revised parallel analysis, Hull = Hull method, CD = comparison data, EKC = Empirical Kaiser Criterion. For the underlying distribution, Lat-NN = non-normal latent variables and normal errors, Err-NN = non-normal error variables and normal latent variables.



Table 3

*Average bias (and standard deviation) of the number of identified factors*

Condition	Level	PA-T	PA-R	Hull	CD	EKC
Number of factors	1	0.01 (0.09)	-0.07 (0.27)	0.00 (0.04)	0.24 (0.52)	0.00 (0.04)
	3	-0.05 (0.32)	-0.32 (0.84)	-0.27 (0.67)	0.00 (0.46)	-0.23 (0.56)
	5	-0.23 (0.76)	-0.58 (1.44)	-1.00 (1.70)	-0.25 (1.07)	-0.71 (1.26)
Number of obs.	100	-0.22 (0.77)	-0.42 (1.04)	-0.82 (1.44)	-0.32 (1.14)	-0.70 (1.17)
	200	-0.11 (0.52)	-0.35 (1.00)	-0.52 (1.25)	0.02 (0.71)	-0.36 (0.90)
	500	-0.03 (0.25)	-0.28 (0.97)	-0.24 (0.91)	0.13 (0.43)	-0.13 (0.58)
	1,000	0.00 (0.10)	-0.24 (0.96)	-0.10 (0.62)	0.16 (0.46)	-0.05 (0.34)
Items per factor	4	-0.21 (0.72)	-1.04 (1.38)	-0.68 (1.35)	-0.02 (1.03)	-0.65 (1.17)
	8	-0.05 (0.35)	0.00 (0.43)	-0.34 (1.06)	0.00 (0.66)	-0.20 (0.66)
	12	-0.02 (0.21)	0.08 (0.34)	-0.24 (0.91)	0.01 (0.50)	-0.09 (0.42)
Distribution	Normal	-0.10 (0.51)	-0.49 (1.13)	-0.50 (1.22)	-0.12 (0.79)	-0.37 (0.92)
	Lat-NN	-0.10 (0.53)	-0.27 (0.96)	-0.43 (1.15)	0.03 (0.80)	-0.32 (0.85)
	Err-NN	-0.07 (0.42)	-0.22 (0.87)	-0.33 (1.03)	0.08 (0.67)	-0.25 (0.76)
Factor correlation	0	0.02 (0.20)	-0.16 (0.81)	-0.09 (0.45)	0.12 (0.55)	-0.10 (0.45)
	.25	-0.03 (0.29)	-0.44 (1.18)	-0.34 (1.01)	-0.03 (0.66)	-0.28 (0.75)
	.5	-0.42 (0.89)	-0.75 (1.41)	-1.47 (1.76)	-0.46 (1.08)	-1.02 (1.34)
Average Loading	.50	-0.11 (0.56)	-0.50 (1.15)	-0.54 (1.26)	-0.09 (0.85)	-0.40 (0.95)
	.65	-0.07 (0.40)	-0.15 (0.78)	-0.30 (0.98)	0.08 (0.65)	-0.23 (0.73)
Overall		-0.09 (0.49)	-0.32 (1.00)	-0.42 (1.14)	0.00 (0.76)	-0.31 (0.85)

*Note.* Bias is calculated as the difference between extracted factors and underlying factors. Positive values indicate overextraction, negative values indicate underextraction, and 0 indicates no bias. PA-T = traditional parallel analysis, PA-R = revised parallel analysis, Hull = Hull method, CD = comparison data, EKC = Empirical Kaiser Criterion. For the underlying distribution, Lat-NN = non-normal latent variables and normal errors, Err-NN = non-normal error variables and normal latent variables.

**Unidimensional factor models.** Figure 4 shows the average accuracies for unidimensional factor models. The performance of the Hull method and the EKC was very high across all conditions (all  $acc_{Hull} > 98\%$ ,  $OR_{Hull} = 5.73$ , all  $acc_{EKC} > 96\%$ ,  $OR_{EKC} = 5.24$ ). Traditional PA also accurately retrieved the number of factors and was only slightly inferior in conditions with few indicators, low loadings, and  $N = 100$  (where  $\overline{acc}_{\#x \leq 8, \bar{\lambda} = .50, N = 100, PA-T} = 95\%$ , all other  $acc_{PA-T} > 97\%$ ). The accuracy of revised PA strongly depended on the number of indicators per factor ( $OR_{\#x=4, PA-R} = 0.13$ ). The performance was very high for all conditions with at least eight indicators (all  $acc_{PA-R, \#x \geq 8} > 99\%$ ), but only moderate in conditions with shorter scales ( $\overline{acc}_{\#x=4, PA-R} = 77\%$ ), where revised PA frequently underestimated the number of factors ( $\overline{bias}_{\#x=4, PA-R} = -0.21$ ). The accuracy of CD was comparatively low ( $\overline{acc}_{CD} = 81\%$ ,  $OR_{CD} = 0.03$ ) due to frequent overextractions ( $\overline{bias}_{CD} = 0.24$ ), especially when the number of indicators was small ( $\overline{acc}_{\#x=4, CD} = 58\%$ ,  $OR_{\#x=12, CD} = 1.66$ ,  $\overline{bias}_{\#x=4, CD} = 0.51$ ). In addition, in contrast to all other methods, the performance of CD decreased with increasing sample size ( $OR_{N=100, CD} = 2.99$ ,  $\overline{acc}_{N=100, CD} = 89\%$ ,  $\overline{bias}_{N=100, CD} = 0.13$ ; compared to  $OR_{N=1,000, CD} = 0.55$ ,  $\overline{acc}_{N=1,000, CD} = 71\%$ ,  $\overline{bias}_{N=1,000, CD} = 0.37$ ).

**Multiple orthogonal factors.** The average accuracies for orthogonal factor models are displayed in Figure 5. Generally, the performance of all methods increased with sample size and the number of indicators per factor. In factor models with at least eight indicators per factor, the Hull method and the EKC exhibited the best performance of all methods when  $N \geq 200$  ( $acc_{N \geq 200, \#x \geq 8, Hull} > 99\%$ ,  $acc_{N \geq 200, \#x \geq 8, EKC} > 99\%$ , for all conditions) and still performed on par with other approaches for smaller samples ( $\overline{acc}_{N=100, \#x \geq 8, Hull} = 96\%$ ,  $\overline{acc}_{N=100, \#x \geq 8, EKC} = 94\%$ ). In conditions with four indicators, Hull and EKC displayed lower hit rates ( $OR_{\#x=4, Hull} = 0.34$ ,  $\overline{acc}_{\#x=4, Hull} = 87\%$ ,  $OR_{\#x=4, EKC} = 0.80$ ,  $\overline{acc}_{\#x=4, EKC} = 83\%$ ) and underestimated the number of factors ( $\overline{bias}_{\#x=4, Hull} = -0.23$ ,  $\overline{bias}_{\#x=4, EKC} = -0.28$ ). Traditional PA was slightly inferior to the Hull method and the EKC in the conditions involving at least eight indicators

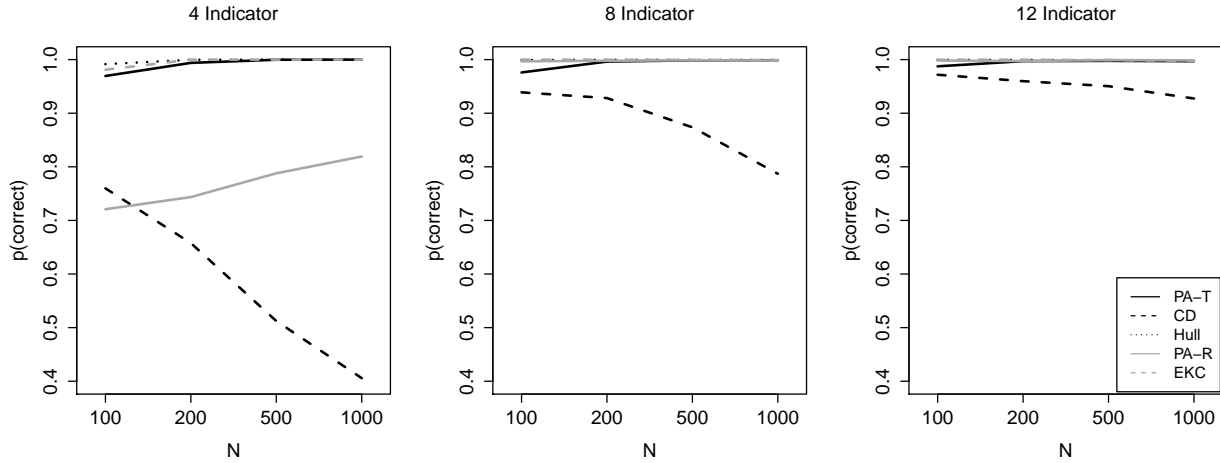


Figure 4. Accuracy of factor extraction criteria for unidimensional factor models depending on the number of indicators per factor and sample size. PA-T - traditional Parallel analysis, CD - Comparison Data, Hull - Hull method, PA-R - revised Parallel analysis, EKC - Empirical Kaiser Criterion.

( $\overline{acc}_{\#x \geq 8, PA-T} = 98\%$ ), but outperformed all other methods in conditions with four indicators ( $\overline{acc}_{\#x=4, PA-T} = 94\%$ ). However, the average accuracy of traditional PA was only moderate when sample sizes were small ( $\overline{acc}_{N=100, \#x=4, PA-T} = 81\%$ ), in part due to a slight tendency to overestimate the number of factors ( $\overline{bias}_{N=100, \#x=4, PA-T} = 0.08$ ). Revised PA and CD performed only moderately in all conditions with orthogonal factors ( $\overline{acc}_{PA-R} = 78\%$ ,  $\overline{acc}_{CD} = 79\%$ ), where revised PA underestimated the number of factors ( $\overline{bias}_{PA-R} = -0.16$ ), whereas CD extracted too many factors on average ( $\overline{bias}_{CD} = 0.12$ ).

**Multiple correlated factors.** Figure 6 summarizes the results for conditions with correlated factors. As was to be expected, all methods exhibited a weaker performance compared to orthogonal factor models, especially when the number of observations or the number of indicators per factor was small ( $OR_{N=100} = 0.10$ ,  $OR_{\#x=4} = 0.13$ ). When factor correlations were low, traditional PA retrieved the number of factors with very high accuracy when the sample size was large ( $acc_{\rho=.25, N \geq 500, PA-T} > 99\%$ , for all conditions). Although the accuracy of traditional PA varied depending on the number of indicators per

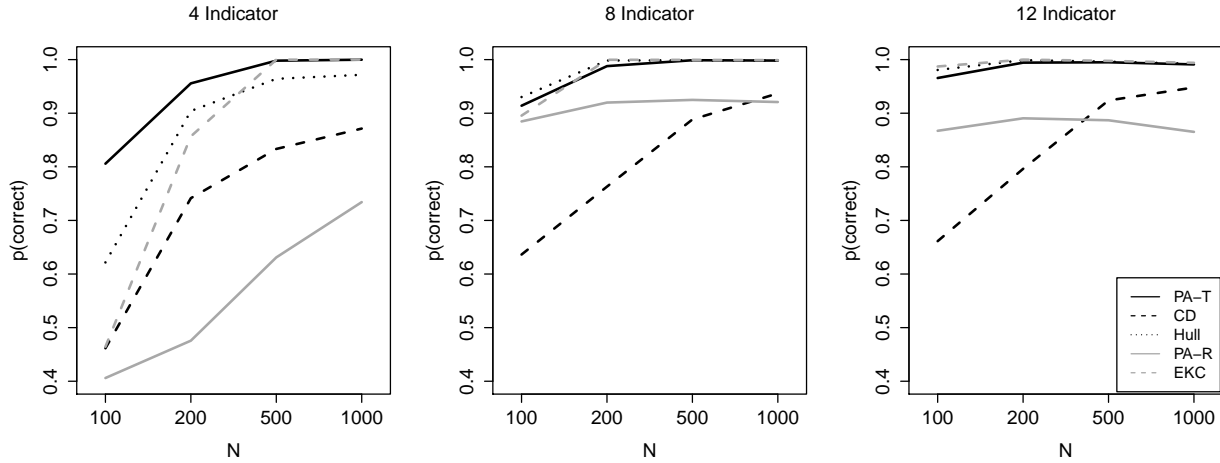


Figure 5. Accuracy of factor extraction criteria for orthogonal factor models depending on number of indicators per factor and sample size. PA-T - traditional Parallel analysis, CD - Comparison Data, Hull - Hull method, PA-R - revised Parallel analysis, EKC - Empirical Kaiser Criterion.

factor for smaller sample sizes ( $\overline{acc}_{\rho=.25, N \leq 200, \#x=4, PA-T} = 77\%$ ,

$\overline{acc}_{\rho=.25, N \leq 200, \#x=12, PA-T} = 98\%$ ), no other method outperformed traditional PA under these conditions. In line with our expectations, most of these errors were underextractions

( $\overline{bias}_{\rho=.25, N \leq 200, \#x=4, PA-T} = -0.18$ ).

Performance was lower overall in conditions with highly correlated factors, especially when only four indicators per factor were used ( $OR_{\rho=.50} = 0.37$ ,  $OR_{\rho=.50, \#x=4} = 0.65$ ). In

addition, all methods underestimated the number of factors ( $\overline{bias}_{\rho=.50, PA-T} = -0.42$ ,

$\overline{bias}_{\rho=.50, CD} = -0.46$ ,  $\overline{bias}_{\rho=.50, Hull} = -1.47$ ,  $\overline{bias}_{\rho=.50, PA-R} = -0.75$ ,  $\overline{bias}_{\rho=.50, EKC} = -1.02$ ),

again reflecting lower factor determinacy and thus greater difficulties to correctly identify the number of factors. With only four indicators per factor – in contrast to all other

conditions considered thus far – CD exhibited the best performance of all methods under scrutiny ( $OR_{\#x=4, CD} = 2.92$ ). For large sample sizes, CD displayed moderate to high

accuracies ( $\overline{acc}_{\#x=4, N \geq 500, \rho=.50, CD} = 86\%$ ), even when loadings were low

( $\overline{acc}_{\#x=4, N \geq 500, \rho=.50, \bar{\lambda}=.50, CD} = 71\%$ ), and virtually no underextraction bias

( $\overline{bias}_{\#x=4, N \geq 500, \rho=.50, CD} = -0.01$ ).

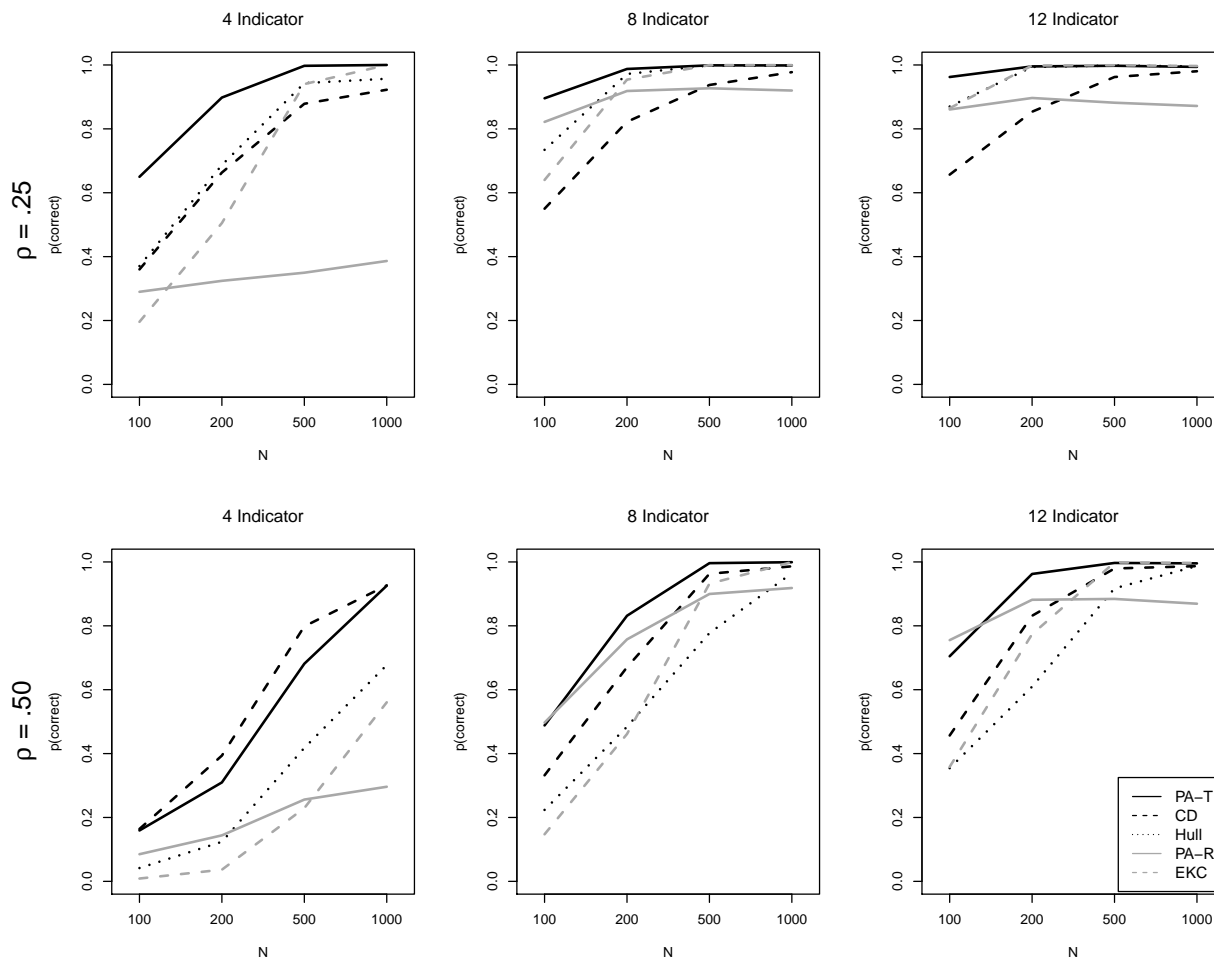


Figure 6. Accuracy of factor extraction criteria for correlated factor models depending on number of indicators per factor, factor correlation, and sample size. The top panels display the accuracy for low factor correlations ( $\rho = .25$ ), the bottom panels for high factor correlations ( $\rho = .50$ ). PA-T - traditional Parallel analysis, CD - Comparison Data, Hull - Hull method, PA-R - revised Parallel analysis, EKC - Empirical Kaiser Criterion.

Performance generally improved with the number of indicators per factor ( $OR_{\#x=12} = 3.46$ ), particularly when applying traditional or revised PA ( $OR_{\#x=4, PA-R} = 0.42$ ). With large sample sizes, traditional PA consistently retrieved the correct number of factors ( $acc_{\rho=.50, \#x \geq 8, N \geq 500, PA-T} > 99\%$  for all conditions). With smaller

sample sizes, traditional and revised PA performed best, but still obtained only moderate hit rates ( $\overline{acc}_{\rho=.50, \#x \geq 8, N \leq 200, PA-T} = 75\%$ ,  $\overline{acc}_{\rho=.50, \#x \geq 8, N \leq 200, PA-R} = 72\%$ ), due to the expected underextraction bias ( $\overline{bias}_{\rho=.50, \#x \geq 8, N \leq 200, PA-T} = -0.35$ ,  $\overline{bias}_{\rho=.50, \#x \geq 8, N \leq 200, PA-R} = -0.20$ ).

The EKC showed a higher performance in those conditions where traditional PA performed well, but exhibited less robust results with small sample sizes or short scales

( $OR_{\#x=4, EKC} = 0.51$ ,  $OR_{N=100, EKC} = 0.29$ ). As predicted, the hit rate of the Hull method decreased overall if factor correlation was high ( $OR_{\rho=.50, Hull} = 0.54$ ). Compared to the other methods in this study, both EKC and Hull frequently underestimated the number of factors and did not perform well under these conditions ( $\overline{acc}_{Hull} = 71\%$ ,  $\overline{bias}_{Hull} = -0.90$ ,  $\overline{acc}_{EKC} = 69\%$ ,  $\overline{bias}_{EKC} = -0.65$ ), unless the sample size was very large ( $\overline{acc}_{N \geq 500, Hull} = 89\%$ ,  $\overline{bias}_{N \geq 500, Hull} = -0.39$ ,  $\overline{acc}_{N \geq 500, EKC} = 89\%$ ,  $\overline{bias}_{N \geq 500, EKC} = -0.21$ ).

### Combination rules

The results presented thus far indicate that traditional PA displayed the highest average accuracy across conditions. However, every other method considered outperformed PA in at least some conditions: EKC and Hull provided very high hit rates for unidimensional or orthogonal factor models even when the sample size was small. Revised PA and CD were more suitable when factors were highly correlated. As such, the question arises whether extraction methods can be beneficially used in conjunction with each other to inform on the number of factors to retain. However, a complication is that investigators obviously have no access to information regarding the true number of factors, the correlation between the factors, or the average loading magnitude before applying EFA and deciding how many factors to extract. In this section, we thus attempt to determine combination rules only considering information that are available to researchers prior to conducting EFA, namely the number of observations, the number of observed variables, the average correlation among the observed variables in the sample, and, of course, the results of all factor extraction criteria.

In principle, the results of various extraction criteria can be combined according to very different schemes. In what follows, we consider a combination rule based on the idea that evidence to extract a particular number of factors is strongest when two criteria agree with respect to the suggested number of factors to retain.<sup>6</sup>

Table 4 shows the conditional hit rates of all pairs of extraction criteria, given that both methods suggest to extract the same number of factors. For instance, in the  $N = 100$  condition, traditional PA and CD agreed regarding the number of factors to retain in 66% of cases (coverage rate). The hit rate of this particular combination, given that they agreed on the suggested number of factors, was 88%. As can be seen from Table 4, all combination rules exhibited a very high accuracy, often close to 100% when  $N \geq 500$ , thereby illustrating the utility of combining the information provided by various criteria to increase overall accuracy. Across conditions, combinations of traditional PA with either the Hull method or the EKC were associated with an overall accuracy of 98%, which substantially improves over traditional PA alone (overall accuracy of 92%). At the same time, combining traditional PA with either the Hull method or the EKC covered 87% and 86%, respectively, of all simulated samples. Only the combination of Hull and EKC provided a slightly larger coverage (88%), however, this was accompanied by a lower accuracy (94%). Taken together, combining traditional PA and either the Hull method or the EKC provided excellent hit rates given that they agree on the number of factors and cover a wide range of conditions.

---

<sup>6</sup>We also compared all triplets of factor extraction criteria where the resulting number of retrieved factors was equal to the median of the suggested number of each triplet. The resulting overall accuracies never exceeded 91%, which is less accurate than traditional PA alone (92%).

Table 4

*Percentage of correctly identified factors for pairs of methods given that both methods agree on the number of factors*  
*(Percentage of cases for which pairs of methods agree on the number of factors)*

Condition	Level	PA-T		Hull		PA-R		CD		Hull		PA-R		EKC	
		CD	Hull	PA-R	EKC	Hull	PA-R	EKC	Hull	PA-R	EKC	PA-R	EKC	PA-R	EKC
N	100	88 (66)	94 (73)	90 (75)	93 (70)	87 (63)	84 (63)	79 (68)	87 (71)	83 (79)	83 (72)				
	200	95 (79)	98 (84)	97 (78)	97 (84)	98 (71)	96 (65)	96 (72)	92 (77)	93 (85)	91 (78)				
	500	99 (86)	100 (93)	100 (80)	99 (94)	100 (81)	99 (71)	99 (81)	96 (82)	98 (92)	98 (81)				
	1,000	100 (87)	100 (97)	100 (82)	100 (97)	100 (84)	99 (72)	100 (84)	97 (84)	100 (96)	100 (82)				
# Items	≤ 15	97 (81)	98 (93)	98 (79)	98 (92)	98 (77)	97 (66)	98 (77)	94 (82)	95 (94)	94 (82)				
	16 – 35	91 (75)	96 (77)	93 (66)	95 (72)	93 (69)	87 (63)	89 (68)	81 (73)	85 (79)	81 (70)				
	> 35	97 (80)	99 (85)	96 (87)	98 (87)	98 (75)	96 (74)	93 (81)	99 (77)	98 (85)	99 (78)				
Sample $\overline{ r }$	≤ .20	94 (78)	97 (81)	95 (70)	97 (80)	95 (72)	92 (61)	90 (75)	87 (71)	91 (83)	88 (75)				
	> .20	98 (81)	99 (95)	98 (91)	99 (95)	100 (79)	98 (76)	99 (78)	99 (89)	98 (95)	99 (89)				
Overall		96 (79)	98 (87)	97 (79)	98 (86)	97 (75)	95 (68)	94 (76)	93 (78)	94 (88)	93 (78)				

*Note.* PA-T = traditional parallel analysis, CD = comparison data, Hull = Hull method, PA-R = revised parallel analysis, EKC = Empirical Kaiser Criterion. # Items = number of items,  $\overline{|r|}$  = average absolute sample correlation.



While concurrence between traditional PA and Hull or EKC reliably indicated that the suggested number of factors is correct, we also examined the conditions in which these methods disagreed to evaluate whether there is an optimal strategy in situations where the proposed combination rule provides conflicting results. In conditions where PA and Hull disagreed (Table 5), the hit rates of all methods decreased considerably and underextractions occurred frequently. The Hull method displayed very low hit rates, especially for large sample sizes ( $\overline{acc}_{N \geq 500, Hull} = 1\%$ ), and consistently underestimated the number of factors ( $\overline{bias}_{Hull} = -2.96$ ). The hit rate of EKC was also low ( $\overline{acc}_{EKC} = 20\%$ ) and only slightly improved with larger sample sizes ( $\overline{acc}_{N \geq 500, EKC} = 37\%$ ). Traditional PA retrieved the correct number of factors in 54% of all cases and obtained acceptable accuracy if the sample size was large ( $\overline{acc}_{N \geq 500, PA-T} = 72\%$ ). Revised PA was superior to other methods when the number of items was large ( $\overline{acc}_{\#items > 25, PA-R} = 65\%$ ), but only showed low hit rates overall ( $\overline{acc}_{PA-R} = 35\%$ ). The performance of CD strongly depended on sample size with comparatively high hit rates when the sample size was large ( $\overline{acc}_{N \geq 500, CD} = 81\%$ ), but low hit rates for smaller samples ( $\overline{acc}_{N \leq 200, CD} = 34\%$ ). The overall pattern of results given that traditional PA and the EKC disagreed on the number of factors was highly similar to the results presented in Table 5 ( $\overline{acc}_{PA-T} = 60\%$ ,  $\overline{acc}_{CD} = 44\%$ ,  $\overline{acc}_{PA-R} = 33\%$ ), with two exceptions. The Hull method obtained higher hit rates ( $\overline{acc}_{Hull} = 28\%$ ), whereas the EKC rarely identified the correct number of factors when it deviated from traditional PA ( $\overline{acc}_{EKC} = 5\%$ ).

In 65% (69%) of all considered cases where traditional PA and the Hull method (traditional PA and the EKC) disagreed on the number of factors, at least one of CD or traditional PA suggested the correct number of factors. No other pair of methods obtained a higher overall hit rate (all  $acc < 63\%$  for traditional PA and Hull, all  $acc < 65\%$  for traditional PA and EKC). For sample sizes of at least 500, either traditional PA or CD identified the number of factors correctly in 88% of both conditions, also superior to every other pair of methods (all  $acc_{N \geq 500} < 86\%$ ). Thus, in cases where the combination of

traditional PA and Hull or EKC provides inconclusive results, considering traditional PA or CD yielded the highest hit rate. Clearly, however, determining the number of factors to retain is difficult under these conditions, in particular when the sample is small.

Table 5

*Percentage of correctly identified number of factors (and average bias) given that traditional PA and Hull provide different solutions*

Condition	Level	PA-T	CD	Hull	PA-R	EKC
Number of observations	100	45 (-0.50)	26 (-1.30)	8 (-2.75)	36 (-1.08)	13 (-1.96)
	200	57 (-0.52)	47 (-0.64)	4 (-3.10)	39 (-1.32)	20 (-1.76)
	$\geq 500$	72 (-0.35)	82 (-0.03)	1 (-3.32)	28 (-2.03)	37 (-1.40)
Number of Items	$\leq 25$	49 (-0.54)	45 (-0.75)	6 (-2.65)	18 (-2.07)	15 (-1.98)
	$> 25$	62 (-0.38)	40 (-1.01)	4 (-3.49)	65 (-0.14)	27 (-1.47)
Sample $ \bar{r} $	$\leq .20$	56 (-0.43)	41 (-0.95)	5 (-2.94)	33 (-1.48)	21 (-1.77)
	$> .20$	42 (-0.74)	55 (-0.27)	4 (-3.08)	52 (-0.57)	14 (-1.88)
Overall		54 (-0.48)	43 (-0.85)	5 (-2.96)	35 (-1.34)	20 (-1.79)

*Note.* PA-T = traditional parallel analysis, CD = comparison data, Hull = Hull method, PA-R = revised parallel analysis, EKC = Empirical Kaiser Criterion. Sample  $|\bar{r}|$  = average absolute sample correlation.

## Discussion

In psychological research, it is often of key interest to determine the number of latent factors underlying multiple observed variables. To this end, EFA is often employed. An important issue in this context pertains to the number of latent factors required to adequately describe the covariance structure among the observed data. A large number of criteria that attempt to inform the decision of how many factors to extract have been suggested in the last decades. Early (but still prominent) criteria, such as Kaiser's criterion or the scree test, have been shown to yield severely biased solutions and, consequently, have

been superseded by other approaches, in particular PA. While the latter approach is often considered the method of choice, a number of new techniques informing factor extraction have been put forward more recently. Each of these methods has been shown to improve upon PA under at least some conditions; however, a thorough comparative evaluation among alternative criteria under a wide range of conditions was still lacking.

Correspondingly, the present study subjected these approaches to a critical test by realizing data conditions that are often encountered in psychological research, systematically varying the number of factors, the factor correlations, the number of indicators, the magnitude of loadings, and the underlying distributions.

Across all conditions, traditional PA (based on the sample correlation matrix and mean eigenvalues) provided the highest hit rate, followed by the Hull method and the EKC. Since traditional PA was superior over all other approaches considered, traditional PA should be chosen to inform factor extraction, if the decision on the number of factors to retain should be based on a single criterion. However, every other method considered outperformed PA in at least one condition. For a sufficient number of indicators per factor, the Hull method and the EKC performed well in unidimensional or orthogonal factor designs. For small sample sizes, revised PA also improved upon traditional PA when the number of indicators per factor was large and factors were correlated. Unlike all other approaches, CD worked comparatively well in conditions with short, highly correlated scales.

Given that each approach has merits under at least one condition, we investigated whether overall performance can be maximized by jointly considering the outcomes of different extraction criteria. Indeed, overall performance increased considerably when multiple factor extraction criteria were used simultaneously. When traditional PA and either the Hull method or the EKC agree (which occurred in 87% and 86% of all simulated data sets, respectively), the number of factors is almost always correctly identified (hit rate of 98%). In the remaining data sets where both methods disagreed, confident judgements

could only be made when sample sizes were large ( $N \geq 500$ ). In these cases, traditional PA or CD correctly identified the number of factors with a probability of 88%. While all approaches exhibited a rather poor performance under these conditions with smaller sample sizes, traditional PA still displayed the highest accuracy. Clearly, the cases in which the combination of traditional PA and Hull or EKC provide conflicting results represent conditions where factor recovery is generally more difficult. This mirrors the fact that sample size requirements for EFA mostly depend on the signal-to-noise ratio in the data (Fabrigar et al., 1999). Whereas conditions with single factors, high loadings, and 12 indicators were easily identified even if  $N = 100$ , correlated factors with low loadings and 4 indicators each were much harder to detect.

The present study also showed that all of the extraction criteria under scrutiny were highly robust under commonly observed values of skewness and kurtosis in the manifest variables, thereby replicating and extending previous results (Dinno, 2009; Garrido et al., 2013; Glorfeld, 1995; Peres-Neto et al., 2005). Note that previous studies investigating non-normality only evaluated traditional PA and only varied the marginal distributions, but neither considered other extraction criteria nor manipulated the multivariate distribution itself. Evidence from studies performed in a confirmatory factor model framework indicates that similar marginal distributions may arise from highly different multivariate distributions, with a differential effect of the latter on the model parameters and goodness-of-fit (e.g. Auerwald & Moshagen, 2015; Foldnes & Grønneberg, 2015; Mair et al., 2012). Consequently, the present study more comprehensively evaluated the performance of a wider array of extraction criteria under non-normal data by considering the multivariate distribution itself. Nevertheless, all criteria were virtually unaffected by non-normality. Whereas non-normal latent variables led to a small overall decrease in accuracy, the average accuracy for non-normal errors was even slightly higher compared to normal distributions, possibly due to the extraction of an additional factor that in part counteracted the observed underextraction bias. Interestingly, although both the Hull

method and the EKC explicitly assume a normal distribution, either by using the CFI which is based on the  $\chi^2$  value of the corresponding structural equation model (Lorenzo-Seva et al., 2011) or as a prerequisite of the Marčenko-Pastur distribution (Braeken & van Assen, in press), their performance was not negatively affected by the non-normality conditions implemented in this study. Consequently, the results of the present study indicate that the investigated extraction criteria can be applied safely under a wide range of distributional properties of the observed data.

### **Issues in Implementing PA**

When PA is employed to inform the extraction of the number of factors, two choices need to be made. The first choice pertains to how to summarize the random reference eigenvalues to which the empirical eigenvalues are compared. Previous studies reported mixed results regarding the tendency to over- or underextract of traditional PA based on the average of random eigenvalues (Buja & Eyuboglu, 1992; Cho et al., 2009; Garrido et al., 2013; Glorfeld, 1995; Peres-Neto et al., 2005; Ruscio & Roche, 2012; Weng & Cheng, 2005). Especially the results of Glorfeld (1995) speak against the use of the average eigenvalue criterion, instead suggesting the 95th percentile to avoid the reported overextraction bias. In contrast, Garrido et al. (2013) as well as the results of our study indicated an underextraction bias. As Peres-Neto et al. (2005) demonstrated, the tendency of PA to overextract mainly occurs in the presence of (at least some) uncorrelated variables. In the rather unrealistic case of a population model with zero factors and uncorrelated observed variables, an average-based PA would overextract at least one factor with a probability of 50%. In these cases, Bartlett's test can be used to determine whether the first eigenvalue is significantly different from the remaining eigenvalues (Bartlett, 1954), but this solution is only available if all observed variables are uncorrelated in the population. While investigators obviously cannot know if there are few systematically uncorrelated items in the data set, the resulting bias of overextractions would likely be less severe, unless an

orthogonal rotation is used (Wood et al., 1996). Overextractions would typically lead to one or more additional factor(s) with overall weak loadings and one high loading for the otherwise uncorrelated item. Such factors and items would likely be excluded, reducing the number of factors to the correct number. Underextractions however lead to substantially stronger biases and are harder to detect (Wood et al., 1996). We therefore recommend the average random eigenvalue instead of using the 95th percentile rule.

The second choice investigators have to make when using PA pertains to the matrix from which the empirical and sampled eigenvalues are derived. The eigenvalues can be obtained either from the correlation matrix, corresponding to a PCA, or from a matrix in which the diagonal of the correlation matrix is replaced with the item communalities estimated by a common factor model. Since the primary purpose of empirical studies often is to uncover a set of latent variables that explain covariations among observed variables, the common factor model is usually recommended over PCA (e.g. Fabrigar et al., 1999; McArdle, 1990; Widaman, 1993). Traditional PA, on the other hand, typically uses the eigenvalues of the correlation matrix as a criterion, which could be considered inconsistent, because the suggested number of components to retain is then used to inform factor extraction in an EFA (Ford, MacCallum, & Tait, 1986; Humphreys & Montanelli, 1975). However, any specification of a common factor model likewise determines the eigenvalues of both matrices under consideration. Indeed, Braeken and van Assen (in press) derived the distribution of eigenvalues of the correlation matrix for normally distributed observed variables from a common factor model. In contrast, the eigenvalues of a common factor model additionally depend on the method that estimates the communalities. Given that both variants of PA seem theoretically appropriate, the hit rates from Monte Carlo simulations should inform the decision which reference eigenvalue should be chosen. As such, the results of our study are in line with Garrido et al. (2013) in suggesting that PA based on a common factor model performs on average worse than PA-PCA.

The inferiority of a common factor PA may also explain why revised PA exhibited

lower accuracies than traditional PA in our study. Following the recommendations of Green et al. (2012), we implemented revised PA based on the common factor model, but implemented traditional PA based on PCA (in line with previous simulation results). This difference likely resulted in the relatively low overall performance of revised PA despite its theoretical advantages. A second downside of revised PA (compared to traditional PA) is that the theoretical maximum hit rate is bound by 95%. Suppose that the correct number of factors was already extracted. Revised PA then proceeds by comparing the next factor to the random sample based eigenvalues (while accounting for previously explained variance). Given that the previous number of factors was already correct, the next eigenvalues of both empirical and random samples should, to the same extent, depend on random error. The empirical eigenvalue will then lie above the 95th percentile of sampled eigenvalues in 5% of cases, thereby leading to the extraction of an additional factor. However, as this behavior should lead to overextractions, this issue arguably played a minor role in our simulation, where revised PA on average underestimated the number of factors. Third, unlike in the study by Green et al. (2012), our design did not incorporate conditions with very high factor intercorrelations (like  $\rho = .80$ ). Under these conditions, revised PA displayed a clear advantage over traditional PA in Green et al. (2012). Thus, when investigators expect such high factor correlations, revised PA based on principal axis factoring and the 95th percentile of sampled eigenvalues could still be a viable alternative. However, we would argue that the expectation of strongly correlated factors requires hypotheses concerning the number and nature of the factors, so that confirmatory factor models are more suited in this context.

### **Limitations**

The results of Monte Carlo studies should only be interpreted within the bounds of the realized conditions. One limitation of our study is that we only considered continuous response variables, because we were also interested in the effect of non-normality in the

observed variables. Changes in the distribution of the latent variables translate to changes in the distribution of the observed variables in a non-trivial way. More specifically, the same values of skewness and kurtosis in the observed ordinal variables can result from different skewness and kurtosis in the underlying continuous variables depending on the thresholds chosen to obtain the ordinal variables. Prespecifying skewness and kurtosis for the underlying continuous variables might therefore not provide valid guidelines for practice, since investigators can only compute skewness and kurtosis of the observed variables. For ordinal variables, Garrido et al. (2013) compared normal and skewed variables and found that traditional PA is also robust against skewness in ordinal variables, if PA is based on the polychoric correlation matrix. Nevertheless, future studies should also examine the performance of other factor extraction criteria for ordinal or dichotomous observed variables.

A second limitation pertains to the selection of the examined extraction criteria. While we included a number of modern techniques that have not yet been thoroughly investigated, we did not consider methods that have been shown to be inferior to traditional PA in previous simulation studies (Peres-Neto et al., 2005; Raïche et al., 2013; Ruscio & Roche, 2012; Zwick & Velicer, 1986). These include indices that incorporate the fit of different structural equation models or test the model fit directly (e.g. Ruscio & Roche, 2012), the minimum average partial method (Velicer, 1976), or several non-graphical solutions for Cattell's Scree test (e.g. Raïche et al., 2013). Overall, there are more than 40 criteria to assess the dimensionality of observed variables (Peres-Neto et al., 2005; Raïche et al., 2013; Ruscio & Roche, 2012) and our selection was based on their relevance for factor analysis in psychology and performance in previous simulation studies. However, it might be possible that a criterion not considered here may improve overall hit rates when used in conjunction with another criterion. Future studies might consider this issue. Finally, it should be noted that we applied a rather rigid criterion to determine the accuracy of the extraction methods. The data were generated using a predefined number of



factors and a method was considered to provide valid results if it successfully recovered this number. This approach assumes that each factually existent latent factor should be recovered, regardless of whether it represents a large or a small proportion of the common variance. Likewise, we also excluded conditions with very high factor correlations in which the explanatory value of additional factors is rather low. We pursued this particular approach based on the rationale that the decision whether an additional correlated factor is to be considered as small, but meaningful or as minor and insignificant mainly depends on theoretical considerations. For example, one condition in the simulation by Green et al. (2012) realized a common factor model with two factors, loadings of  $\lambda = .40$  each, the same number of indicators per factor, and a factor correlation of  $\rho = .80$ . In this condition, the (unrotated) second factor only explains 1.6% of the common variance. By comparison, the minor factors that methods were supposed to ignore in the study conducted by Lorenzo-Seva et al. (2011) on average accounted for 15% of the common variance. Clearly, we cannot expect one statistical method to appropriately differentiate between these conditions, because the decision of which result is to be considered the correct one would also depend on the interpretation of the extracted factor solution. Thus, we included neither highly correlated nor minor factors. Nevertheless, our approach may have led to an overly critical assessment of the accuracy of extraction methods that specifically aim to extract major factors only, even in the presence of minor factors (e.g., the Hull method; Lorenzo-Seva et al., 2011). At the same time, the results obtained herein should not be readily transferred to situations in which the goal is to uncover factors with very high intercorrelations (e.g. Green et al., 2012).

### Conclusion

We investigated the performance of various criteria to decide on the number of factors to retain in EFA. Our results indicate that the highest accuracy can be obtained when considering the outcomes of several criteria simultaneously. In particular, within the

bounds of this simulation study, we recommend a decision heuristic viable over a wide range of data conditions. First, investigators should compare the results of traditional PA and the Hull method or the EKC. If both methods suggest the same number of factors, this most likely reflects the correct number of underlying factors. If both methods disagree, both traditional PA and CD are viable extraction criteria when the sample is large, whereas traditional PA should be chosen when the sample is small. However, the latter conditions are generally associated with greater difficulties to identify the number of factors for all approaches we investigated. Thus, under these conditions, confident decisions require larger sample sizes. In the suggested decision rule, disagreement between traditional PA and the Hull method or the EKC can thus serve as an indicator that the latent structure is more difficult to uncover.

Finally, we want to stress that decisions on the number of factors should also involve theoretical considerations. While the suggested strategy is a helpful tool in assessing the number of factors and the confidence investigators should have in this number, it should not be interpreted as a strict and rigid rule. The interpretability of the resulting loading patterns, theoretical considerations concerning the relevance of an item for a scale, and the resulting scale reliabilities are all equally important and should all be taken into account when deciding how many factors to retain in EFA.

## References

- Auerswald, M., & Moshagen, M. (2015). Generating correlated, non-normally distributed data using a non-linear structural model. *Psychometrika*, *80*, 920–937.  
doi:10.1007/s11336-015-9468-7
- Bartlett, M. S. (1954). A note on the multiplying factors for various  $\chi^2$  approximations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 296–298.  
doi:10.4236/ib.2015.73013
- Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*, 238–246. doi:10.1037/0033-2909.107.2.238
- Bentler, P. M., & Kano, Y. (1990). On the equivalence of factors and components. *Multivariate Behavioral Research*, *25*, 67–74. doi:10.1207/s15327906mbr2501\_8
- Blanca, M. J., Arnau, J., López-Montiel, D., Bono, R., & Bendayan, R. (2013). Skewness and kurtosis in real data samples. *Methodology*, *9*, 78–84.  
doi:10.1027/1614-2241/a000057
- Braeken, J., & van Assen, M. A. (in press). An empirical Kaiser criterion. *Psychological Methods*. doi:10.1037/met0000074
- Browne, M. W. (2001). An overview of analytic rotation in exploratory factor analysis. *Multivariate Behavioral Research*, *36*, 111–150. doi:10.1207/S15327906MBR3601\_05
- Buja, A., & Eyuboglu, N. (1992). Remarks on parallel analysis. *Multivariate Behavioral Research*, *27*, 509–540. doi:10.1207/s15327906mbr2704\_2
- Cain, M. K., Zhang, Z., & Yuan, K.-H. (in press). Univariate and multivariate skewness and kurtosis for measuring nonnormality: Prevalence, influence and estimation. *Behavior Research Methods*. doi:10.3758/s13428-016-0814-1
- Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, *1*, 245–276. doi:10.1207/s15327906mbr0102\_10
- Cattell, R. B., & Vogelmann, S. (1977). A comprehensive trial of the scree and KG criteria for determining the number of factors. *Multivariate Behavioral Research*, *12*,

- 289–325. doi:10.1207/s15327906mbr1203\_2
- Ceulemans, E., & Kiers, H. A. (2006). Selecting among three-mode principal component models of different types and complexities: A numerical convex hull based method. *British Journal of Mathematical and Statistical Psychology*, *59*, 133–150.  
doi:10.1348/000711005X64817
- Cho, S.-J., Li, F., & Bandalos, D. (2009). Accuracy of the parallel analysis procedure with polychoric correlations. *Educational and Psychological Measurement*, *69*, 748–759.  
doi:10.1177/0013164409332229
- Comrey, A. L. (1978). Common methodological problems in factor analytic studies. *Journal of Consulting and Clinical Psychology*, *46*, 648–659. doi:10.1037/0022-006X.46.4.648
- Crawford, A. V., Green, S. B., Levy, R., Lo, W.-J., Scott, L., Svetina, D., & Thompson, M. S. (2010). Evaluation of parallel analysis methods for determining the number of factors. *Educational and Psychological Measurement*, *70*, 885–901.  
doi:10.1177/0013164410379332
- Dinno, A. (2009). Exploring the sensitivity of Horn's parallel analysis to the distributional form of random data. *Multivariate Behavioral Research*, *44*, 362–388.  
doi:10.1080/00273170902938969
- DiStefano, C., & Hess, B. (2005). Using confirmatory factor analysis for construct validation: An empirical review. *Journal of Psychoeducational Assessment*, *23*, 225–241. doi:10.1177/073428290502300303
- Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, *4*, 272–299. doi:10.1037/1082-989X.4.3.272
- Fava, J. L., & Velicer, W. F. (1992). The effects of overextraction on factor and component analysis. *Multivariate Behavioral Research*, *27*, 387–415.  
doi:10.1207/s15327906mbr2703\_5
- Fava, J. L., & Velicer, W. F. (1996). The effects of underextraction in factor and

- component analyses. *Educational and Psychological Measurement*, *56*, 907–929.  
doi:10.1177/0013164496056006001
- Finch, J. F., & West, S. G. (1997). The investigation of personality structure: Statistical models. *Journal of Research in Personality*, *31*, 439–485. doi:10.1006/jrpe.1997.2194
- Foldnes, N., & Grønneberg, S. (2015). How general is the Vale–Maurelli simulation approach? *Psychometrika*, *80*, 1066–1083. doi:10.1007/s11336-014-9414-0
- Ford, J. K., MacCallum, R. C., & Tait, M. (1986). The application of exploratory factor analysis in applied psychology: A critical review and analysis. *Personnel psychology*, *39*, 291–314. doi:10.1111/j.1744-6570.1986.tb00583.x
- Garrido, L. E., Abad, F. J., & Ponsoda, V. (2013). A new look at Horn's parallel analysis with ordinal variables. *Psychological Methods*, *18*, 454–474. doi:10.1037/a0030005
- Glorfeld, L. W. (1995). An improvement on Horn's parallel analysis methodology for selecting the correct number of factors to retain. *Educational and Psychological Measurement*, *55*, 377–393. doi:10.1177/0013164495055003002
- Green, S. B., Levy, R., Thompson, M. S., Lu, M., & Lo, W.-J. (2012). A proposed solution to the problem with using completely random data to assess the number of factors with parallel analysis. *Educational and Psychological Measurement*, *72*, 357–374. doi:10.1177/0013164411422252
- Green, S. B., Thompson, M. S., Levy, R., & Lo, W.-J. (2015). Type I and type II error rates and overall accuracy of the revised parallel analysis method for determining the number of factors. *Educational and Psychological Measurement*, *75*, 428–457. doi:10.1177/0013164414546566
- Hakstian, A. R., Rogers, W. T., & Cattell, R. B. (1982). The behavior of number-of-factors rules with simulated data. *Multivariate Behavioral Research*, *17*, 193–219. doi:10.1207/s15327906mbr1702\_3
- Harman, H. H. (1976). *Modern factor analysis* (3rd ed.). Chicago: University of Chicago

Press.

- Henson, R. K., & Roberts, J. K. (2006). Use of exploratory factor analysis in published research: Common errors and some comment on improved practice. *Educational and Psychological Measurement, 66*, 393–416. doi:10.1177/0013164405282485
- Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30*, 179–185. doi:10.1007/BF02289447
- Hubbard, R., & Allen, S. J. (1987). An empirical comparison of alternative methods for principal component extraction. *Journal of Business Research, 15*, 173–190. doi:10.1016/0148-2963(84)90047-X
- Humphreys, L. G., & Ilgen, D. R. (1969). Note on a criterion for the number of common factors. *Educational and Psychological Measurement, 29*, 571–578. doi:10.1177/001316446902900303
- Humphreys, L. G., & Montanelli, R. G. (1975). An investigation of the parallel analysis criterion for determining the number of common factors. *Multivariate Behavioral Research, 10*, 193–205. doi:10.1207/s15327906mbr1002\_5
- IBM Corp. (2015). IBM SPSS Statistics for Windows, Version 23.0. Armonk, NY: IBM Corp.
- Jackson, D. L., Gillaspay, J. A., & Purc-Stephenson, R. (2009). Reporting practices in confirmatory factor analysis: An overview and some recommendations. *Psychological Methods, 14*, 6. doi:10.1037/a0014694
- Jöreskog, K. G. (2007). Factor analysis and its extensions. In R. Cudeck & R. C. MacCallum (Eds.), *Factor analysis at 100: Historical developments and future directions* (p. 47-77). Mahwah, New Jersey: Lawrence Erlbaum.
- Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement, 20*, 141–151. doi:10.1177/001316446002000116
- Lance, C. E., Butts, M. M., & Michels, L. C. (2006). The sources of four commonly

- reported cutoff criteria: What did they really say? *Organizational Research Methods*, *9*, 202–220. doi:10.1177/1094428105284919
- Lorenzo-Seva, U., Timmerman, M. E., & Kiers, H. A. (2011). The Hull method for selecting the number of common factors. *Multivariate Behavioral Research*, *46*, 340–364. doi:10.1080/00273171.2011.564527
- Mair, P., Satorra, A., & Bentler, P. M. (2012). Generating nonnormal multivariate data using copulas: Applications to SEM. *Multivariate Behavioral Research*, *47*, 547–565. doi:10.1080/00273171.2012.692629
- Marčenko, V. A., & Pastur, L. A. (1967). Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, *1*, 457–483. doi:10.1070/SM1967v001n04ABEH001994
- McArdle, J. J. (1990). Principles versus principals of structural factor analyses. *Multivariate Behavioral Research*, *25*, 81–87. doi:10.1207/s15327906mbr2501\_10
- Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, *105*, 156.
- Peres-Neto, P. R., Jackson, D. A., & Somers, K. M. (2005). How many principal components? Stopping rules for determining the number of non-trivial axes revisited. *Computational Statistics & Data Analysis*, *49*, 974–997. doi:10.1016/j.csda.2004.06.015
- R Core Team. (2016). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from <https://www.R-project.org/>
- Raïche, G., Riopel, M., & Blais, J.-G. (2006). *Nongraphical solutions for the Cattell's scree test*. (Paper presented at the annual meeting of the Psychometric Society, Montreal, Quebec, Canada)
- Raïche, G., Walls, T. A., Magis, D., Riopel, M., & Blais, J.-G. (2013). Non-graphical solutions for Cattell's scree test. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *9*, 23–29. doi:10.1027/1614-2241/a000051

- Revelle, W. (2015). psych: Procedures for psychological, psychometric, and personality research [Computer software manual]. Evanston, Illinois. Retrieved from <http://CRAN.R-project.org/package=psych> (R package version 1.5.8)
- Ruscio, J., & Kaczetow, W. (2008). Simulating multivariate nonnormal data using an iterative algorithm. *Multivariate Behavioral Research, 43*, 355–381.  
doi:10.1080/00273170802285693
- Ruscio, J., & Roche, B. (2012). Determining the number of factors to retain in an exploratory factor analysis using comparison data of known factorial structure. *Psychological Assessment, 24*, 282–292. doi:10.1037/a0025697
- Steger, M. F. (2006). An illustration of issues in factor extraction and identification of dimensionality in psychological assessment data. *Journal of Personality Assessment, 86*, 263–272. doi:10.1207/s15327752jpa8603\_03
- Thurstone, L. L. (1947). *Multiple factor analysis*. Chicago: University of Chicago Press.
- Turner, N. E. (1998). The effect of common variance and structure pattern on random data eigenvalues: Implications for the accuracy of parallel analysis. *Educational and Psychological Measurement, 58*, 541–568. doi:10.1177/0013164498058004001
- Velicer, W. F. (1976). Determining the number of components from the matrix of partial correlations. *Psychometrika, 41*, 321–327. doi:10.1007/BF02293557
- Velicer, W. F., Eaton, C. A., & Fava, J. L. (2000). Construct explication through factor or component analysis: A review and evaluation of alternative procedures for determining the number of factors or components. In *Problems and solutions in human assessment: Honoring Douglas N. Jackson at seventy* (pp. 41–71). Boston, MA: Kluwer Academic.
- Weng, L.-J., & Cheng, C.-P. (2005). Parallel analysis with unidimensional binary data. *Educational and Psychological Measurement, 65*, 697–716.  
doi:10.1177/0013164404273941
- Widaman, K. F. (1993). Common factor analysis versus principal component analysis:



- Differential bias in representing model parameters? *Multivariate Behavioral Research*, *28*, 263–311. doi:10.1207/s15327906mbr2803\_1
- de Winter, J. C., & Dodou, D. (2012). Factor recovery by principal axis factoring and maximum likelihood factor analysis as a function of factor pattern and sample size. *Journal of Applied Statistics*, *39*, 695–710. doi:10.1080/02664763.2011.610445
- de Winter, J. C., & Dodou, D. (2016). Common factor analysis versus principal component analysis: A comparison of loadings by means of simulations. *Communications in Statistics - Simulation and Computation*, *45*, 299-321. doi:10.1080/03610918.2013.862274
- Wood, J. M., Tataryn, D. J., & Gorsuch, R. L. (1996). Effects of under- and overextraction on principal axis factor analysis with varimax rotation. *Psychological Methods*, *1*, 354–365. doi:10.1037/1082-989X.1.4.354
- Worthington, R. L., & Whittaker, T. A. (2006). Scale development research: A content analysis and recommendations for best practices. *The Counseling Psychologist*, *34*, 806–838. doi:10.1177/0011000006288127
- Zwick, W. R., & Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin*, *99*, 432-442. doi:10.1037/0033-2909.99.3.432

### Appendix

The goal of this section is to explain the correspondence between explained variance of the common factor model and the eigenvalues of the matrix of correlations  $\mathbf{R}_C$  with communalities on the diagonal, assuming that the (hypothetical) data fit the common factor model perfectly. Note that the explained variance in a PCA can be similarly derived if we used the correlation matrix  $\mathbf{R}$  instead of  $\mathbf{R}_C$ . Suppose we have standardized observed variables  $\mathbf{X}_C = (x_{C1}, \dots, x_{Cm})^T$  from which we partialled out the uniqueness, such that

$$\mathbf{R}_C = E(\mathbf{X}_C \mathbf{X}_C^T) \quad (13)$$

is the covariance matrix of  $\mathbf{X}_C$ . We denote the observations as  $x_{Ck}$ ,  $1 \leq k \leq N$  for  $N$  observations and try to find factors that linearly explain variations in  $\mathbf{X}_C$ . This is equivalent to finding lines on which we project each observation  $x_{Ck}$  such that the variance of the length of projections is maximal (and the variance of the distances to the line is minimal). A line is a set of points that satisfy

$$x = \alpha v, \quad (14)$$

where  $v$  is a vector of length  $p$  and  $\alpha \in \mathbb{R}$ . The length of the projection of  $x_{Ck}$  on this line is

$$\frac{\langle x_{Ck}, v \rangle}{\|v\|}. \quad (15)$$

Note that the length of  $v$  does not change the line in Equation 14, so that we can set  $\|v\| = 1$  without loss of generality. The length of projections then is  $\langle x_{Ck}, v \rangle$ . In order to maximize the variance of  $\langle x_{Ck}, v \rangle$ , we first obtain the average of the projections. The vector  $v$  is part of an orthonormal basis of our space, which we denote as

$$\{v, v'_2, \dots, v'_p\}. \quad (16)$$

We can rewrite every observation as

$$x_{Ck} = \alpha_{1k}v + \alpha_{2k}v'_2 + \dots + \alpha_{pk}v'_p, \quad (17)$$

so that

$$x_{Ck} = \alpha_{1k}v + \sum_{i=2}^p \alpha_{ik}v'_i \quad (18)$$

$$\Rightarrow v^T x_{Ck} = v^T \alpha_{1k}v + v^T \sum_{i=2}^p \alpha_{ik}v'_i \quad (19)$$

$$\Rightarrow v^T x_{Ck} = \alpha_{1k}v^T v + \sum_{i=2}^p \alpha_{ik}v^T v'_i \quad (20)$$

$$\Rightarrow v^T x_{Ck} = \alpha_{1k}. \quad (21)$$

In the last step, we used that  $v \perp v'_i$ ,  $2 \leq i \leq p$  and  $v^T v = \|v\| = 1$ . The mean of projections therefore is

$$\sum_{k=1}^N \alpha_{1k}v = \sum_{k=1}^N v^T x_{Ck}v \quad (22)$$

$$= v^T \left( \sum_{k=1}^N x_{Ck} \right) v \quad (23)$$

$$= 0 \quad (24)$$

because  $x_{Ck}$  is standardized. We can therefore obtain the variance of the length of projections of  $x_{Ck}$  as

$$\frac{1}{N-1} \sum_{k=1}^N \langle x_{Ck}, v \rangle^2 = \frac{1}{N-1} \sum_{k=1}^N (x_{Ck} \cdot v)^2 \quad (25)$$

$$= \frac{1}{N-1} \sum_{k=1}^N v^T x_{Ck}^T x_{Ck} v \quad (26)$$

$$= \frac{1}{N-1} v^T \left( \sum_{k=1}^N x_{Ck}^T x_{Ck} \right) v \quad (27)$$

$$= v^T \mathbf{R}_C v. \quad (28)$$

The variance of the length of projections is  $v^T \cdot \mathbf{R}_C \cdot v$ , we try to obtain the maximum.

We denote the eigenvectors of  $\mathbf{R}_C$  as  $e_1, \dots, e_p$  and the corresponding eigenvalues as  $l_1, \dots, l_p$  such that  $l_1 \geq l_2 \geq \dots \geq l_p$ . If we choose  $v = e_1$ , the variance is

$$e_1^T \mathbf{R}_C e_1 = e_1^T (l_1 e_1) = l_1. \quad (29)$$

The first eigenvalue corresponds to the explained variance if we choose the eigenvector  $e_1$  as a projection line. Suppose we choose any other vector as a projection line. The eigenvectors  $e_1, \dots, e_p$  form an orthonormal basis of our space. We can therefore rewrite  $v$  as

$$v = \langle e_1, v \rangle e_1 + \langle e_2, v \rangle e_2 + \dots + \langle e_p, v \rangle e_p = \sum_{i=1}^p \langle e_i, v \rangle e_i. \quad (30)$$

The variance of the length of projections for  $v$  then is

$$\left( \sum_{i=1}^p \langle e_i, v \rangle e_i \right)^T \mathbf{R}_C \left( \sum_{i=1}^p \langle e_i, v \rangle e_i \right) = \left( \sum_{i=1}^p \langle e_i, v \rangle e_i \right)^T \left( \sum_{i=1}^p \langle e_i, v \rangle \mathbf{R}_C e_i \right) \quad (31)$$

$$= \left( \sum_{i=1}^p \langle e_i, v \rangle e_i \right)^T \left( \sum_{i=1}^p \langle e_i, v \rangle l_i e_i \right) \quad (32)$$

$$= \sum_{i=1}^p \langle e_i, v \rangle^2 l_i \|e_i\|^2 \quad (33)$$

In the last step, we used that  $e_i \perp e_{i'}$  for  $1 \leq i, i' \leq p$  and  $i \neq i'$ . Note that the eigenvectors are standardized, so that  $\|e_i\| = 1$ . Further note that  $\langle e_i, v \rangle^2 \geq 0$  and

$$\sum_{i=1}^p \langle e_i, v \rangle^2 = 1 \quad (34)$$

because  $\|v\| = 1$ . Therefore, the variance of the length of projections for  $v$  is a weighted sum of eigenvalues where the weights are all non-negative and sum to one, such that

$$v^T \mathbf{R}_C v = \sum_{i=1}^p \langle e_i, v \rangle^2 l_i \leq l_1. \quad (35)$$

Hence,  $v = e_1$  obtains a maximum of explained variance. If we choose a second factor, we choose a line orthogonal to  $e_1$  and, by analogy, arrive at the conclusion that  $v = e_2$  with corresponding explained variance  $l_2$ . For  $m$  extracted factors, the explained variance is

$$\sum_{j=1}^m l_j. \quad (36)$$