Journal of Modern Applied Statistical Methods

Volume 18 | Issue 1 Article 2

2-13-2020

Analytical Closed-Form Solution for General Factor with Many Variables

Stan Lipovetsky
GfK North America, stan.lipovetsky@gmail.com

Vladimir Manewitsch
NIM e.V., Nuremberg, Germany, vladimir.manewitsch@nim.org

Recommended Citation

Lipovetsky, S., & Manewitsch, V. (2019). Analytical closed-form solution for general factor with many variables. Journal of Modern Applied Statistical Methods, 18(1), eP2616. doi: 10.22237/jmasm/1556668980

Analytical Closed-Form Solution for General Factor with Many Variables

Erratum

The original version of this article had character encoding errors for several summation and products in formulas. This has been corrected.

Journal of Modern Applied Statistical Methods
May 2019, Vol. 18, No. 1, eP2616
doi: 10.22237/jmasm/1556668980
בס"ד

ISSN 1538 − 9472

Analytical Closed-Form Solution for General Factor with Many Variables

Stan Lipovetsky
GfK North America
Minneapolis, MN

Vladimir Manewitsch
NIM e.V.
Nuremberg, Germany

The factor analytic triad method of one-factor solution gives the explicit analytical form for a common latent factor built by three variables. The current work considers analytical presentation of a general latent factor constructed in a closed-form solution for multivariate case. The results can be supportive to theoretical description and practical application of latent variable modeling, especially for big data because the analytical closed-form solution is not prone to data dimensionality.

Keywords: Factor loadings, latent variable, Spearman method of triads, multivariate closed-form solution

---

Introduction

The statistical method of Factor Analysis (FA) is widely and heavily used in many disciplines. Main capability of this method is about dimensionality reduction of multivariate problems which allows for focusing on essential features of the underlying research problem, interpretation and visualization of high dimensional data, and discovering latent constructs or dimensions behind multiple variables. Particularly in marketing research, FA has been extensively applied for describing consumer attitudes and their personal characteristics, identifying service dimensions, product positioning and modelling of microeconomic hypotheses.

Spearman (1904, 1927) originated FA principles for a common latent factor based on the triad relation $r_{12} = r_{13}r_{23}$ between pair correlations of three variables and derived from it the tetrads relations $r_{12}r_{34} = r_{13}r_{24} = r_{14}r_{23}$ for four variables correlations. From statistical point of view, Factor Analysis, and related to it the principal component analysis (PCA), and singular value decomposition (SVD) are well known tools of multivariate statistics (Harman, 1976; Dillon and Goldstein,

---

doi: 10.22237/jmasm/1556668980 | Accepted: August 3, 2017; Published: February 13, 2020.
Correspondence: Stan Lipovetsky, stan.lipovetsky@gmail.com

1984; Elden, 2007; Izenman, 2008). Modern FA, PCA and SVD working with many variables and factors have been developed in numerous works (Bartholomew and Knott, 1999; Skrondal and Rabe-Hesketh, 2004; Drton et al., 2007; Härdle and Hlavka, 2007; Härdle and Simar, 2012; Gorsuch, 2014; Brown, 2015) and implemented in software packages.

Although FA can be performed in maximum likelihood and eigenproblem approaches (Lawley and Maxwell, 1971; Lipovetsky and Conklin 2005; Lipovetsky, 2009, 2015, 2017) the analytical formulae show explicitly how a general factor solution is defined by data incorporated into it. For instance, the so-called triads method is applied in various practical research, particularly, for measurement of validity using biomarkers (Ocké and Kaaks, 1997; Kabagambe et al., 2001; McNaughton et al., 2005; de Carvalho Yokota et al., 2010, Djekic-Ivankovic et al., 2016), where the validity coefficient for a questionnaire can be defined via pair correlations among three characteristics as

$$q = \sqrt{r_{q1} r_{q2} / r_{12}} \tag{1}$$

which corresponds to a general FA solution with three variables, as is shown below.

Consider analytical derivation of a general latent factor and its presentation in a closed-form solution for multivariate case of any number of variables. It demonstrates how the general FA loadings can be constructed in algebraic formulae directly from the correlation matrix. Having such closed-form solution facilitates building a general latent variable especially for big data because the analytical solution does not require the iterative calculations and so does not depend on a data dimensionality. In contrast to numerical methods, the presented solution enables direct and interpretable linking of factor solution to the underlying correlations making apparent the structure and the formation of the latent construct. This property of the closed-form solution offers new prospects for FA focusing on interpretation of factors that is mostly the case in applications for marketing. In the following sections, the paper presents solutions for the general factor constructed by several variables, considers the least squares solution for the general factor by multiple variables, describes additional specific features of the obtained solutions, and demonstrates numerical examples.

General Factor Solution for Several Variables

Factor analysis aims to present all the measured variables via a smaller number of some latent unobserved variables called factors. Suppose the variables are standardized so that their means equal zero and variances equal one (FA is scale-invariant so this transformation only simplifies the derivation). For an a-priori fixed set of variables x_j (where j = 1, 2, …, n – number of variables) the model presenting the variables via a common, or general, factor f and the specific factors u_j in each point of observations is:

$$x_j = q_j f + u_j \tag{2}$$

where the unknown parameters q_j are the so-called loadings of the variables on the factor. Simplifying normalizing conditions (expectations equal zero, variance of the factor f equals one) and assumptions of independence of general and specific factors (covariance equals zero) are usually imposed:

$$\begin{aligned} &E(f) = E(u_j) = 0,\ \ \mathrm{var}(f) = 1,\ \ \mathrm{var}(u_j) = c_j, \\ &\mathrm{cov}(f, u_j) = \mathrm{cov}(u_j, u_k) = 0,\ j \neq k \end{aligned} \tag{3}$$

with c_j denoting variances of specific factors. Covariance matrix of the standardized variables coincides with the correlation matrix $R = X'X$, where X is the data matrix of N by n order (of N observations by n variables), and prime denotes transposition. Then correlation matrix for the model (2)-(3) can be presented as follows:

$$R = qq' + diag(c) \tag{4}$$

where q is the vector of unknown loadings q_j, the outer product of this vector is qq', and $diag(c)$ is the diagonal matrix of the specific factors' variances.

All the parameters of loadings q_j and specific variances c_j are unknown and have to be estimated by the known symmetric matrix of correlation (4) built from the data. The system (4) can be represented in explicit form:

$$\begin{pmatrix} 1 & r_{12} & r_{13} & \cdots & r_{1n} \\ r_{21} & 1 & r_{23} & \cdots & r_{2n} \\ r_{31} & r_{32} & 1 & \cdots & r_{3n} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ r_{n1} & r_{n2} & r_{n3} & \cdots & 1 \end{pmatrix} = \begin{pmatrix} q_1^2 + c_1 & q_1 q_2 & q_1 q_3 & \cdots & q_1 q_n \\ q_2 q_1 & q_2^2 + c_2 & q_2 q_3 & \cdots & q_2 q_n \\ q_3 q_1 & q_3 q_2 & q_3^2 + c_3 & \cdots & q_3 q_n \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ q_n q_1 & q_n q_2 & q_n q_3 & \cdots & q_n^2 + c_n \end{pmatrix} \tag{5}$$

Consider the general FA for several examples of number of variables. For $n = 3$, the off-diagonal relations (5) yield the following system of equations:

$$q_1 q_2 = r_{12},\ q_1 q_3 = r_{13},\ q_2 q_3 = r_{23} \tag{6}$$

Multiplying the first two equations (6) containing q_1 gives $q_1^2 q_2 q_3 = r_{12} r_{13}$ and substituting into it the third equation (6) yields solution $q_1^2 = r_{12} r_{13} / r_{23}$. Taking square root of it, and repeating similar derivation for other parameters q we obtain the closed-form solution for all loadings:

$$q_1 = \sqrt{\frac{r_{12} r_{13}}{r_{23}}},\ q_2 = \sqrt{\frac{r_{12} r_{23}}{r_{13}}},\ q_3 = \sqrt{\frac{r_{13} r_{23}}{r_{12}}} \tag{7}$$

The triads model (1) coincides with one of the loadings (7). With the solutions (7), taking equalities for the diagonal elements in (5) produces estimates for the specific variances:

$$c_j = 1 - q_j^2 \tag{8}$$

where the squared parameters q_j^2 are called communalities. The results (7)-(8) are known (for instance, see Härdle and Hlavka, 2007, pp. 187-188; Härdle and Simar, 2012, pp. 315-316), let us extend them to higher dimensions.

For number of variables $n = 4$, similarly to (6) we take the off-diagonal relations (5) and get the system of six equations:

$$q_1 q_2 = r_{12},\ q_1 q_3 = r_{13},\ q_1 q_4 = r_{14},\ q_2 q_3 = r_{23},\ q_2 q_4 = r_{24},\ q_3 q_4 = r_{34} \tag{9}$$

Product of the first three relations (9) containing q_1 yields the equation $q_1^3 q_2 q_3 q_4 = r_{12} r_{13} r_{14}$ and the product of the rest three relations (9) without q_1 yields another equation $(q_2q_3q_4)^2 = r_{23}r_{24}r_{34}$. Substituting the latter one into the former one produces the result $q_1^6 r_{23}r_{24}r_{34} = (r_{12}r_{13}r_{14})^2$. Then the solution for q_1 and similarly derived other loadings are:

$$q_1 = \sqrt[6]{\frac{(r_{12}r_{13}r_{14})^2}{r_{23}r_{24}r_{34}}},\ q_2 = \sqrt[6]{\frac{(r_{12}r_{23}r_{24})^2}{r_{13}r_{14}r_{34}}},\ q_3 = \sqrt[6]{\frac{(r_{13}r_{23}r_{34})^2}{r_{12}r_{14}r_{24}}},\ q_4 = \sqrt[6]{\frac{(r_{14}r_{24}r_{34})^2}{r_{12}r_{13}r_{23}}} \tag{10}$$

By analogue with (7) we can call (10) the closed-form solution for the tetrads model of general factor by four variables. The corresponding estimates for the specific variances are given by the same relation (8). By comparison of (7) with (10) it is not yet clear how to compose loadings for any number of variables.

Consider the case of $n = 5$, where we take the off-diagonal relations (5) and get the system of ten equations:

$$\begin{gathered} q_1q_2 = r_{12},\ q_1q_3 = r_{13},\ q_1q_4 = r_{14},\ q_1q_5 = r_{15}, \\ q_2q_3 = r_{23},\ q_2q_4 = r_{24},\ q_2q_5 = r_{25},\ q_3q_4 = r_{34},\ q_3q_5 = r_{35},\ q_4q_5 = r_{45} \end{gathered} \tag{11}$$

The product of the first four relations (11) containing q_1 yields the equation $q_1^4 q_2q_3q_4q_5 = r_{12}r_{13}r_{14}r_{15}$ and the product of the rest six relations (9) without q_1 yields another equation $(q_2q_3q_4q_5)^3 = r_{23}r_{24}r_{25}r_{34}r_{35}r_{45}$. Substituting the latter one into the former produces $q_1^{12} r_{23}r_{24}r_{25}r_{34}r_{35}r_{45} = (r_{12}r_{13}r_{14}r_{15})^3$. Solving with respect to q_1 and similar derivations for other loadings yields:

$$q_1 = \sqrt[12]{\frac{(r_{12}r_{13}r_{14}r_{15})^3}{r_{23}r_{24}r_{25}r_{34}r_{35}r_{45}}},\ q_2 = \sqrt[12]{\frac{(r_{12}r_{23}r_{24}r_{25})^3}{r_{13}r_{14}r_{15}r_{34}r_{35}r_{45}}},\ \cdots,\ q_5 = \sqrt[12]{\frac{(r_{15}r_{25}r_{35}r_{45})^3}{r_{12}r_{13}r_{14}r_{23}r_{24}r_{34}}} \tag{12}$$

By analogue with (7) and (10) we can call the closed-form solution (12) the pentads model of general factor by five variables. The estimates for the specific variances are given by the relation (8) as well. Comparing (7), (10), and (12) it is possible to estimate how the loadings incorporate the pair correlations in these closed-form solutions for more variables.

General Factor Solution for Multiple Variables

Consider how to extend the obtained results to deriving the closed-form solution for common factor by any number of variables. Suppose all correlations positive, otherwise we flip scales of variables which can make the correlation matrix to be compatible with the factor model (2), so to have positive elements only.

For $n = 3$, the system of three equations (6) with three unknown variables q_j yields the unique solution (7). But for $n = 4$, the system of $n(n - 1)/2 = 6$ equations (9) with four unknown variables q_j is over-identified, so it needs some optimizing criterion for incorporating all the relations into a solution (10). Similarly with $n = 5$, the system of $n(n - 1)/2 = 10$ equations (11) with five unknown variables q_j needs a criterion for collapsing all the relations into a solution (12).

Such a needed criterion for solving over-identified systems with number of equations more than number of variables can be found in the Least Squares (LS) objective well-known in the regression modeling. For a positive matrix of correlations, consider a model with multiplicative error term in approximation off-diagonal elements of correlation matrix (4):

$$r_{jk} = q_j q_k \delta_{jk} \tag{13}$$

where δ_{jk} is a relative error in the jk-th correlation presented via the product of the loadings q_j and q_k. Taking logarithm of the relations (13) yields a linearized model:

$$\ln r_{jk} = \ln q_j + \ln q_k + \ln \delta_{jk} \tag{14}$$

Denoting logarithm of correlations as $y_{jk} = \ln r_{jk}$, absolute errors $\varepsilon_{jk} = \ln \delta_{jk}$, and the unknown coefficients

$$b_j = \ln q_j \tag{15}$$

rewrite (14) as a linear regression:

$$y_{jk} = b_j + b_k + \varepsilon_{jk} \tag{16}$$

The LS objective for the model (16) is:

$$LS = \sum_{j>k}^{n} \left(\varepsilon_{jk}\right)^2 = \sum_{j>k}^{n} \left(y_{jk} - b_j - b_k\right)^2 \to \min \tag{17}$$

For more explicit presentation, a design matrix of the predictors for the regression model (16)-(17) can be seen in the following example of $n = 5$ variables when (14) in a plain form is:

$$\begin{pmatrix} \ln r_{12} \\ \ln r_{13} \\ \ln r_{14} \\ \ln r_{15} \\ \ln r_{23} \\ \ln r_{24} \\ \ln r_{25} \\ \ln r_{34} \\ \ln r_{35} \\ \ln r_{45} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 \end{pmatrix} \cdot \begin{pmatrix} \ln q_1 \\ \ln q_2 \\ \ln q_3 \\ \ln q_4 \\ \ln q_5 \end{pmatrix} + \begin{pmatrix} \ln \delta_{12} \\ \ln \delta_{13} \\ \ln \delta_{14} \\ \ln \delta_{15} \\ \ln \delta_{23} \\ \ln \delta_{24} \\ \ln \delta_{25} \\ \ln \delta_{34} \\ \ln \delta_{35} \\ \ln \delta_{45} \end{pmatrix} \tag{18}$$

Denoting the design matrix at the right-hand side (18) as Z, the system (18) can be presented in the matrix form as

$$y = Zb + \varepsilon \tag{19}$$

where the vectors y, b, and epsilon correspond to the notations used in description (14)-(17). The system (19) can be solved in the LS approach (17) as a regression model of y by z variables given by columns in the matrix Z. Minimization by unknown parameters in (17) yields the so-called normal equations

$$Z'Zb = Z'y \tag{20}$$

where prime denotes the transposed matrix Z'. The solution of (20) is given via the matrix inversion as follows:

$$b = \left(Z'Z\right)^{-1} Z'y \tag{21}$$

For an explicit example, consider this solution for the case $n = 5$ with the equations (18), when the system (20) can be presented as:

$$\begin{pmatrix} 4 & 1 & 1 & 1 & 1 \\ 1 & 4 & 1 & 1 & 1 \\ 1 & 1 & 4 & 1 & 1 \\ 1 & 1 & 1 & 4 & 1 \\ 1 & 1 & 1 & 1 & 4 \end{pmatrix} \cdot \begin{pmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \\ b_5 \end{pmatrix} = \begin{pmatrix} \sum_{j\neq 1}^{n} \ln r_{1j} \\ \sum_{j\neq 2}^{n} \ln r_{2j} \\ \sum_{j\neq 3}^{n} \ln r_{3j} \\ \sum_{j\neq 4}^{n} \ln r_{4j} \\ \sum_{j\neq 5}^{n} \ln r_{5j} \end{pmatrix} \tag{22}$$

Diagonal elements in the matrix $Z'Z$ (22) are the sums of squares in each column of the design matrix in (18), so for n variables they equal $n – 1$ (as number of pairs of each one variable with $n – 1$ others), and the off-diagonal elements in (22) equal one as the scalar products of columns in (18). The sums of logarithms in the vector $Z'y$ (22) can be presented as logarithms of products of all the elements in each row of the correlation matrix at the left-hand side (5) (for simplicity we can incorporate the diagonal elements $r_{jj} = 1$). So the general expression for the normal system (20) with arbitrary n is:

$$\begin{pmatrix} (n-1) & 1 & 1 & \cdots & 1 \\ 1 & (n-1) & 1 & \cdots & 1 \\ 1 & 1 & (n-1) & \cdots & 1 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & 1 & 1 & \cdots & (n-1) \end{pmatrix} \cdot \begin{pmatrix} b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_5 \end{pmatrix} = \begin{pmatrix} \ln \prod_{j=1}^{n} r_{1j} \\ \ln \prod_{j=1}^{n} r_{2j} \\ \cdots \\ \ln \prod_{j=1}^{n} r_{nj} \end{pmatrix} \tag{23}$$

To obtain solution b (21) from this system, we need to invert the n-th order matrix at the left-hand side (23). This matrix can be represented as

$$\begin{pmatrix} (n-1) & 1 & 1 & \cdots & 1 \\ 1 & (n-1) & 1 & \cdots & 1 \\ 1 & 1 & (n-1) & \cdots & 1 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & 1 & 1 & \cdots & (n-1) \end{pmatrix} = (n-2)I + ee' \tag{24}$$

where I and e are the identity matrix and vector of the n-th order, respectively, so $(n-2)I = A$ is the scalar matrix (denoted A), and ee' is the vectors outer product which defines the matrix of n-th order with each element equals one.

Applying the Sherman-Morrison formula for matrix inversion yields:

$$(A+ee')^{-1} = A^{-1} - \frac{A^{-1}ee'A^{-1}}{1+e'A^{-1}e} = \frac{1}{n-2}I - \frac{ee'}{2(n-1)(n-2)} \tag{25}$$

Using (25), the solution of (23) can be presented as follows:

$$\begin{aligned} \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix} &= \left(\frac{1}{n-2}I - \frac{ee'}{2(n-1)(n-2)} \right) \cdot \begin{pmatrix} \ln \prod_{j=1}^{n} r_{1j} \\ \ln \prod_{j=1}^{n} r_{2j} \\ \cdots \\ \ln \prod_{j=1}^{n} r_{nj} \end{pmatrix} \\ &= \begin{pmatrix} \ln\left(\prod_{j=1}^{n} r_{1j}\right)^{\frac{1}{n-2}} - \ln\left(\prod_{i,j=1}^{n} r_{ij}\right)^{\frac{1}{2(n-1)(n-2)}} \\ \ln\left(\prod_{j=1}^{n} r_{2j}\right)^{\frac{1}{n-2}} - \ln\left(\prod_{i,j=1}^{n} r_{ij}\right)^{\frac{1}{2(n-1)(n-2)}} \\ \cdots\cdots\cdots\cdots\cdots\cdots \\ \ln\left(\prod_{j=1}^{n} r_{nj}\right)^{\frac{1}{n-2}} - \ln\left(\prod_{i,j=1}^{n} r_{ij}\right)^{\frac{1}{2(n-1)(n-2)}} \end{pmatrix} \end{aligned} \tag{26}$$

Exponent of (26) due to the relation (15) yields the loadings of the general factor for any number n of variables. Each j-th loading value is proportional to the root of power $n-2$ from the product of all elements in the j-th row of correlation matrix, and reciprocally proportional to the root of power $2(n-1)(n-2)$ from the product of all elements of the correlation matrix:

$$\begin{pmatrix} q_1 \\ q_2 \\ \vdots \\ q_n \end{pmatrix} = \begin{pmatrix} \ln \prod_{j=1}^{n} r_{1j} \\ \ln \prod_{j=1}^{n} r_{2j} \\ \cdots \\ \ln \prod_{j=1}^{n} r_{nj} \end{pmatrix} \frac{1}{\left(\prod_{i,j=1}^{n} r_{ij}\right)^{\frac{1}{2(n-1)(n-2)}}} \tag{27}$$

If the intent is to use only the elements in a half of symmetrical correlation matrix then it is possible to skip the term 2 and represent (27) within the same power of the root as follows:

$$\begin{pmatrix} q_1 \\ q_2 \\ \vdots \\ q_n \end{pmatrix} = \left[\begin{pmatrix} \ln \prod_{j=1}^{n} r_{1j} \\ \ln \prod_{j=1}^{n} r_{2j} \\ \cdots \\ \ln \prod_{j=1}^{n} r_{nj} \end{pmatrix} \frac{1}{\prod_{i \geq j}^{n} r_{ij}} \right]^{\frac{1}{(n-1)(n-2)}} \tag{28}$$

This general expression reduces to those used in the examples with $n = 3$, 4, and 5 above. Indeed, for $n = 3$, the formula (28) becomes

$$\begin{pmatrix} q_1 \\ q_2 \\ q_3 \end{pmatrix} = \left[\begin{pmatrix} (r_{12}r_{13})^2 \\ (r_{12}r_{23})^2 \\ (r_{13}r_{23})^2 \end{pmatrix} \frac{1}{r_{12}r_{13}r_{23}} \right]^{\frac{1}{2}} = \begin{pmatrix} (r_{12}r_{13})/r_{23} \\ (r_{12}r_{23})/r_{13} \\ (r_{13}r_{23})/r_{12} \end{pmatrix}^{\frac{1}{2}} \tag{29}$$

which coincides with the results in (7). For $n = 4$, the formula (28) reduces to

$$\begin{pmatrix} q_1 \\ q_2 \\ q_3 \\ q_4 \end{pmatrix} = \left[\begin{pmatrix} (r_{12}r_{13}r_{14})^3 \\ (r_{12}r_{23}r_{24})^3 \\ (r_{13}r_{23}r_{34})^3 \\ (r_{14}r_{24}r_{34})^3 \end{pmatrix} \frac{1}{r_{12}r_{13}r_{14}r_{23}r_{24}r_{34}} \right]^{\frac{1}{6}} = \begin{pmatrix} (r_{12}r_{13}r_{14})^2 / (r_{23}r_{24}r_{34}) \\ (r_{12}r_{23}r_{24})^2 / (r_{13}r_{14}r_{34}) \\ (r_{13}r_{23}r_{34})^2 / (r_{12}r_{14}r_{24}) \\ (r_{14}r_{24}r_{34})^2 / (r_{12}r_{13}r_{23}) \end{pmatrix}^{\frac{1}{6}} \tag{30}$$

it is the same result as (10). And for $n = 5$, the formula (28) is

$$\begin{aligned} \begin{pmatrix} q_1 \\ q_2 \\ \vdots \\ q_5 \end{pmatrix} &= \left[\begin{pmatrix} (r_{12}r_{13}r_{14}r_{15})^4 \\ (r_{12}r_{23}r_{24}r_{25})^4 \\ \dots \\ (r_{15}r_{25}r_{35}r_{45})^4 \end{pmatrix} \frac{1}{r_{12}r_{13}r_{14}r_{15}r_{23}r_{24}r_{25}r_{34}r_{35}r_{45}} \right]^{\frac{1}{12}} \\ &= \begin{pmatrix} (r_{12}r_{13}r_{14}r_{15})^3 / (r_{23}r_{24}r_{25}r_{34}r_{35}r_{45}) \\ (r_{12}r_{23}r_{24}r_{25})^3 / (r_{13}r_{14}r_{15}r_{34}r_{35}r_{45}) \\ \dots \\ (r_{15}r_{25}r_{35}r_{45})^3 / (r_{12}r_{13}r_{14}r_{23}r_{24}r_{34}) \end{pmatrix}^{\frac{1}{12}} \end{aligned} \tag{31}$$

which reproduces the result (12).

Thus, the formula (28) presents the closed-form solution for the general factor by any number of variables. Together with the partial cases (29)-(31), it shows that each j-th loading depends directly on the correlations of x_j with the other variables, and depends reciprocally on the correlations among all other variables without x_j. In the regression modeling approach (13)-(21), besides estimation of the loadings as regression coefficients, it is possible to obtain evaluation of their errors, t-statistics, and overall quality of the fit given in the residual variance and coefficient of multiple determination. The regression modeling should be used if the interest is in estimation of the quality of the factor loadings, otherwise the same results on the loadings themselves can be obtained by the analytical formulae (7)-(12) or (27)-(31).

Specific Features of the Obtained Solutions for a General Factor

Consider some additional specifics of the general factor solutions. From relations (27)-(28) it is easy to find that the product of loadings $q_1 q_2 \ldots q_n$ equals the product of all different r_{ij} (in the upper correlation matrix triangle, $j > i$) in power $1/(n = 1)$, or in other words, the geometric mean of loadings equals square root of the geometric means of r_{ij} where $j > i$. It means that any one loading can always be obtained from the others. The structure of loadings is defined by the proportions of the numerators in (28) (the product in denominator (28) is common for all loadings) which can be written as follows:

$$q_1 : q_2 : \ldots. : q_n = \sqrt[n-2]{\prod_{j=1}^{n} r_{1j}} : \sqrt[n-2]{\prod_{j=1}^{n} r_{2j}} : \ldots. : \sqrt[n-2]{\prod_{j=1}^{n} r_{nj}}$$
$$= G_1^{n/(n-2)} : G_2^{n/(n-2)} : \ldots. : G_n^{n/(n-2)} \qquad (32)$$

where G_j denote geometric means of the elements in each j-th row of correlation matrix. So for n variables the loadings in the general factor are proportional to these geometric means powered to $n/(n = 2)$. Similar relations can be obtained with covariance matrix as well.

Another property of the general solution in (28) is that the j-th loading can be obtained as the geometric mean of all possible triad solutions (29) involving x_j, e.g. for $j = 1$ and $n = 4$:

$$q_1 = \left[\left(r_{12} r_{13} r_{14} \right)^2 / \left(r_{23} r_{24} r_{34} \right) \right]^{\frac{1}{6}} = \left[\left(r_{12} r_{13} / r_{23} \right)^{\frac{1}{2}} \left(r_{12} r_{14} / r_{24} \right)^{\frac{1}{2}} \left(r_{13} r_{14} / r_{34} \right)^{\frac{1}{2}} \right]^{\frac{1}{3}} \qquad (33)$$

Similarly, for $n = 5$

$$q_1 = \left[\left(r_{12} r_{13} r_{14} r_{15} \right)^3 / \left(r_{23} r_{24} r_{25} r_{34} r_{35} r_{45} \right) \right]^{\frac{1}{12}}$$
$$= \begin{bmatrix} \left(r_{12} r_{13} / r_{23} \right)^{\frac{1}{2}} \left(r_{12} r_{14} / r_{24} \right)^{\frac{1}{2}} \left(r_{13} r_{14} / r_{34} \right)^{\frac{1}{2}} \\ \left(r_{12} r_{15} / r_{25} \right)^{\frac{1}{2}} \left(r_{13} r_{15} / r_{35} \right)^{\frac{1}{2}} \left(r_{14} r_{15} / r_{45} \right)^{\frac{1}{2}} \end{bmatrix}^{\frac{1}{6}} \qquad (34)$$

This result can be extended for the case of general $n > 5$ by simple algebra. Furthermore, this result can be generalized in the sense, that the solution for j-th loading from the space of n variables can be represented as geometric mean of all possible solutions involving j-th variable in a lower dimensional space $m < n$. For example, for $n = 5$ and $m = 4$:

$$q_1 = \left[(r_{12}r_{13}r_{14}r_{15})^3 / (r_{23}r_{24}r_{25}r_{34}r_{35}r_{45}) \right]^{\frac{1}{12}}$$
$$= \begin{bmatrix} \left[\left(r_{12}r_{13}r_{14} \right)^2 / (r_{23}r_{24}r_{34}) \right]^{\frac{1}{6}} \left[\left(r_{12}r_{13}r_{15} \right)^2 / (r_{23}r_{25}r_{35}) \right]^{\frac{1}{6}} \\ \left[\left(r_{12}r_{15}r_{14} \right)^2 / (r_{25}r_{24}r_{45}) \right]^{\frac{1}{6}} \left[\left(r_{15}r_{13}r_{14} \right)^2 / (r_{35}r_{45}r_{34}) \right]^{\frac{1}{6}} \end{bmatrix}^{\frac{1}{4}} \tag{35}$$

Concerning the assumption of positive correlations, consider several violations of it. The case when some correlations equal zero is unlikely in practical applications, and close to zero correlations are not important and can be taken by absolute value. Even if a correlation equals exactly zero, $r_{ij} = 0$, exclusion of one of the involved variables from the analysis can be considered, because the FA model (3)-(6) with $r_{ij} = q_i q_j$, implies that one of the loadings is zero. Thus, at least one of the variables x_i or x_j is not driven by the common factor f, but by the corresponding specific factor u only.

For the case of (significant) negative correlations, we can consider flipping of the scale of a corresponding variable by multiplying it by –1. Because flipping a variable changes the sign of all correlations involving that variable, the question is if the correlation matrix be turned to be positive by flipping a subset of variables, which can be answered by considering all possible triads of correlations given in the matrix. Let us call "triad" a set of three correlations (r_{ij}, r_{jk}, r_{ki}) for three variables x_i, x_j, and x_k in the correlation matrix R. Call R admissible if it can be turned into a positive matrix, with only positive entries, by flipping some x-variables. Any triad describes a 3×3 matrix which is admissible if its triad has one of two patterns: (+, +, +) or (+, –, –), which we can call admissible triads. The other triads such as (+, +, –) and (–, –, –) are called inadmissible and can exist because correlations in a triad do not have to be transitive (Langford et al., 2001; Lipovetsky, 2002; Lipovetsky and Conklin, 2004). Then it can be shown by induction that a correlation matrix R of an order $n > 3$ is admissible if and only if it contains only admissible triads.

Hence, checking that all triads are admissible ensures that the whole correlation matrix is suitable for closed-form solution in (28), and if yes, which variables have to be flipped. However, the reverse argument is not true for the proposed solution in form of eq. (28), i.e., the correlation matrix does not have to be admissible but the solution (28) could be applied to it anyway. For instance, look at a 5×5 correlation matrix having the following pattern of signs of its elements:

$$R = \begin{pmatrix} 1 & - & + & + & + \\ \ldots & 1 & + & - & + \\ \ldots & \ldots & 1 & + & + \\ \ldots & \ldots & \ldots & 1 & + \\ \ldots & \ldots & \ldots & \ldots & 1 \end{pmatrix} \tag{36}$$

There are several inadmissible triads, e.g. (r_{12}, r_{13}, r_{23}), but the solution (28) still can be calculated. Thus, on one hand, the admissibility of a correlation matrix is sufficient, but not necessary for applicability of the closed-form solution in eq. (28). However, inadmissibility of a correlation matrix may prevent using of (28), whereas other FA techniques (e.g., PC, PA, etc.) can be applied anyway. Nevertheless, it is not a shortcoming of the proposed solution but it is rather indication that an inadmissible correlation matrix is not compatible with the factor model (2) or assumptions in (3). Indeed, suppose there is an inadmissible triad $(r_{12}, r_{13}, r_{23}) = (+, +, –)$ within a correlation matrix. By Eq. (4) which follows from (2) and (3) we get the triad representation via the loadings:

$$\left(q_1 q_2,\ q_1 q_3,\ q_2 q_3\right) = \left(+,+,-\right) \tag{37}$$

Thus, it can be assumed without loss of generality $q_1 > 0$, then it can be concluded from the first two elements in the triad that $q_2 > 0$ and $q_3 > 0$, which leads to the inequality $r_{23} = q_2 q_3 > 0$, but that contradicts to the condition (37). An inadmissible triad (+, +, –) violates the relations (2) or (3). If some solution can be calculated for an inadmissible correlation matrix, this solution is not compatible with the standard factor model.

It could be argued such solutions are not meaningful in most applications. This can be illustrated by an attempt of geometrical interpretation of the situation in (37). The correlation between two variables corresponds to the cosine of the angle between corresponding vectors. An inadmissible triad like in (37) implies that there is always an obtuse angle (i.e., negative cosine) between some pair of vectors in the triad, regardless which direction we choose for them. In most applications, an obtuse angle between the vectors contradicts intuition behind the factor solution, especially its interpretation as common latent construct underlying the manifest variables.

Furthermore, if the closed-form solution is represented as a geometric mean of triad solutions (as in (33) and (34)), it becomes applicable if and only if the correlation matrix is admissible. In this form, the analytical solution is calculable if and only if the correlation matrix is compatible with factor model (2) and (3), whereas numerical factor analysis methods would produce some, likely senseless, results for correlation matrices not compatible with the underlying factor model.

Numerical examples

Consider a classical example given in Harman (1967, p. 244) on nine cognitive variables by 696 respondents from a work by Holzinger, and available in `psych` package of R software procedures for psychological, psychometric, and personality research (Revelle, 2017). Table 1 shows the matrix of pair correlations for this Harman data.

Table 1. Correlation matrix by Harman data.

	***x*1**	***x*2**	***x*3**	***x*4**	***x*5**	***x*6**	***x*7**	***x*8**	***x*9**
***x*1**	1.00	0.75	0.78	0.44	0.45	0.51	0.21	0.30	0.31
***x*2**	0.75	1.00	0.72	0.52	0.53	0.58	0.23	0.32	0.30
***x*3**	0.78	0.72	1.00	0.47	0.48	0.54	0.28	0.37	0.37
***x*4**	0.44	0.52	0.47	1.00	0.82	0.82	0.33	0.33	0.31
***x*5**	0.45	0.53	0.48	0.82	1.00	0.74	0.37	0.36	0.36
***x*6**	0.51	0.58	0.54	0.82	0.74	1.00	0.35	0.38	0.38
***x*7**	0.21	0.23	0.28	0.33	0.37	0.35	1.00	0.45	0.52
***x*8**	0.30	0.32	0.37	0.33	0.36	0.38	0.45	1.00	0.67
***x*9**	0.31	0.30	0.37	0.31	0.36	0.38	0.52	0.67	1.00

Presented in Table 2 is the regression solution b (21) with t-statistics for parameters of the model and its characteristic of quality of fit: the coefficient of multiple determination R^2. The last two columns show the FA-1 loadings $q = exp(b)$ and their errors which can be calculated as $s(q) = qb/t$. We see that the t-statistics is very good, mostly above 2, so the parameters are significantly different from zero, and the quality of total fit is high, $R^2 = 0.898$.

Table 2. Regression model and loadings, by Harman data.

	b	*t*-statistics	loadings *q*	*s*(*q*)
***x*1**	-0.429	3.501	0.651	-0.080
***x*2**	-0.358	2.917	0.699	-0.086
***x*3**	-0.312	2.545	0.732	-0.090
***x*4**	-0.322	2.628	0.724	-0.089
***x*5**	-0.278	2.266	0.757	-0.093
***x*6**	-0.223	1.817	0.800	-0.098
***x*7**	-0.736	5.998	0.479	-0.059
***x*8**	-0.554	4.514	0.575	-0.071
***x*9**	-0.547	4.456	0.579	-0.071
R^2	0.898			

In Table 3, FA-1 solutions obtained by several methods are compared, based on principal components (PC, the main component), maximum likelihood (ML), principal axes (PA), minimum residuals (MR), and closed-form (CF) analytic solution. Squared norm of each of these vectors define the loadings' variance (eigenvalue) presented in the bottom row. By these variances we see that the CF is close to the ML solution.

Table 3. FA-1 solutions obtained by several methods, by Harman data.

	PC	*ML*	*PA*	*MR*	*CF*
***x*1**	0.746	0.634	0.706	0.683	0.651
***x*2**	0.781	0.696	0.750	0.734	0.699
***x*3**	0.783	0.666	0.750	0.713	0.732
***x*4**	0.798	0.868	0.774	0.832	0.724
***x*5**	0.803	0.844	0.780	0.818	0.757
***x*6**	0.835	0.879	0.824	0.856	0.800
***x*7**	0.529	0.424	0.461	0.435	0.479
***x*8**	0.604	0.465	0.535	0.491	0.575
***x*9**	0.607	0.461	0.537	0.488	0.579
Variance	4.771	4.178	4.297	4.276	4.082

Shown in Table 4 are the pair correlations between these solutions, and the mean value in each column. By correlations, CF is similar to PC and PA, and close to other methods as well. By the mean values, CF yields an intermediate solution between ML and other methods.

Table 4. Correlation matrix of FA-1 solutions, by Harman data.

	PC	ML	PA	MR	CF
PC	1.000	0.934	0.999	0.978	0.977
ML	0.934	1.000	0.945	0.987	0.931
PA	0.999	0.945	1.000	0.984	0.976
MR	0.978	0.987	0.984	1.000	0.959
CF	0.977	0.931	0.976	0.959	1.000
mean	0.972	0.949	0.976	0.977	0.961

In accordance with the model (2)-(3), for a given number of variables n the values for loadings q_j were generated by the uniform distribution on the 0-1 interval, and for a given number of observations N a vector of factor scores f was generated by the normal distribution with zero mean value and standard deviation equals one. Using q_j values, the estimates for the specific factor variances c_j were obtained by the relation (8), and then the specific factors u_j (2) were generated by the normal distributions for the sample size N, zero means, and the variances c_j. With all these constructs, the vectors of variables x_j were built by the formula (2), and these variables x_j were used for making their correlation matrix R as at the left-hand side in (4)-(5). By this correlation matrix, the loadings were estimated in several classical approaches and the analytical solution. For a numerical illustration with $n = 5$ variables and $N = 100$ observations, the correlation matrix is shown in Table 5.

Table 5. Correlation matrix, by simulated data.

	***x*1**	***x*2**	***x*3**	***x*4**	***x*5**
***x*1**	1.000	0.434	0.567	0.563	0.477
***x*2**	0.434	1.000	0.636	0.722	0.658
***x*3**	0.567	0.636	1.000	0.739	0.513
***x*4**	0.563	0.722	0.739	1.000	0.615
***x*5**	0.477	0.658	0.513	0.615	1.000

For the matrix in Table 5, the simulated loadings q and its estimations by the same methods described for Table 3 are presented in Table 6. The loadings' variance is shown in the bottom row, and we can see that all solutions are very close to the original loadings. Particularly, it is so for the closed-form analytical solution CF, which in regression estimation yields the coefficient of multiple determination $R^2 = 0.977$, and the adjusted to degrees of freedom $R^2 = 0.954$, with the corresponding F-statistic equal 42.21 and its p-value 0.0004, so the quality of fit is very high.

Table 6. Original loadings and FA-1 solutions in several methods, by simulated data.

	q	*PC*	*ML*	*PA*	*MR*	*CF*
***x*1**	0.659	0.725	0.630	0.633	0.630	0.632
***x*2**	0.784	0.847	0.807	0.807	0.807	0.794
***x*3**	0.823	0.846	0.809	0.804	0.809	0.806
***x*4**	0.870	0.894	0.896	0.891	0.894	0.891
***x*5**	0.745	0.792	0.709	0.720	0.710	0.724
Variance	3.039	3.384	3.008	3.010	3.005	2.999

Presented in Table 7 is the matrix of correlations between original *q* and all solutions from Table 6, so we see that CF is the closest solution to the original simulated loadings. In the last two rows Table 7 shows also the mean absolute error (MAE) and relative the mean absolute error (RMAE, in %) of each solution compared with the original simulated loadings *q*, by which we can see that the analytic solution yields very good results.

Table 7. Correlation matrix of original loadings and FA-1 solutions, by simulated data.

	q	*PC*	*ML*	*PA*	*MR*	*CF*
q	1.000	0.983	0.980	0.981	0.980	0.990
PC	0.983	1.000	0.994	0.997	0.995	0.995
ML	0.980	0.994	1.000	0.999	1.000	0.996
PA	0.981	0.997	0.999	1.000	0.999	0.998
MR	0.980	0.995	1.000	0.999	1.000	0.996
CF	0.990	0.995	0.996	0.998	0.996	1.000
MAE		4.440	2.558	2.253	2.487	1.905
RMAE, %		5.954	3.371	2.960	3.284	2.515

Other factor solutions can be a bit closer to the original loadings than the closed-form solution. For instance, presenting in Table 8 are descriptive statistics for MAE, relative RMAE, and correlations of several solutions with the original loadings for n = 10 and N = 1000, run in 1000 samples.

The results for minimum, mean, and maximum values are very close among all the solutions, and in general the difference is very small and negligible for any reasonable precision with which the loading values are used in all practical needs.

Table 8. Descriptive statistics for MAE, RMAE, and correlations, by 1000 samples.

		PC	*MLE*	*PA*	*MR*	*CF*
MAE	*min*	0.106	0.040	0.042	0.041	0.043
	mean	0.284	0.116	0.117	0.116	0.120
	max	0.536	0.303	0.305	0.303	0.305
RMAE, %	*min*	1.898	2.129	2.131	2.129	2.131
	mean	3.122	3.523	3.523	3.523	3.525
	max	4.407	5.010	5.011	5.010	5.016
Correlations	*min*	0.921	0.926	0.928	0.926	0.925
	mean	0.987	0.988	0.988	0.988	0.987
	max	0.999	0.999	0.999	0.999	0.999

For comparison of runtime of the solutions, take correlation matrix as a starting point, so the results do not depend on a number of observations but only on a varying number of variables in the simulated data. Figure 1 shows increasing saving in runtime for CF over ML and MR methods when the number of variables increases. The results underlie advantages of the proposed solution for big data having wide format, i.e., many variables, while for a moderate number of variables the methods exhibit comparable running time.

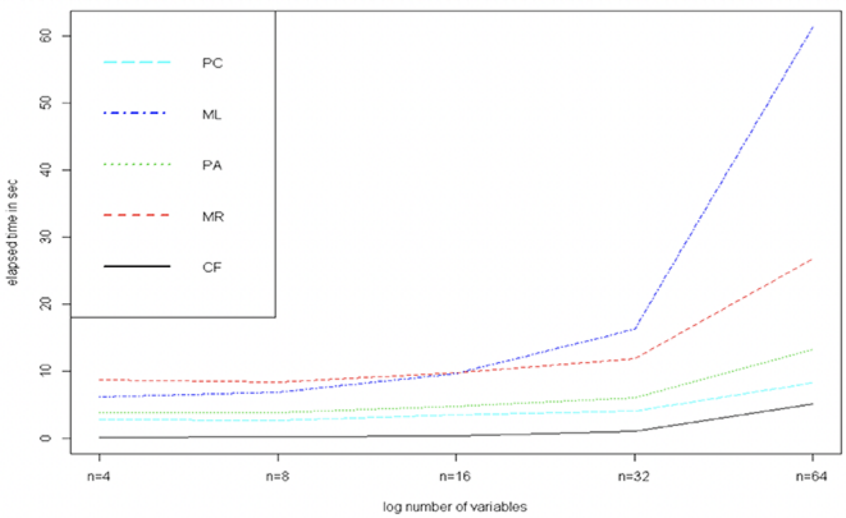

Figure 1. Runtime of calculation with *n*=4, 8, 16, 32 and 64 by 100 samples.

Summary

An analytical derivation of the loadings for a general latent factor is presented. The results are expressed in a closed-form solution for multivariate case and can be seen as extension of Spearman triads to any number of variables. The obtained algebraic formulae show explicitly how the general factor loadings are constructed from the correlation matrix, so they are useful for analysis and for practical applications.

This approach can be useful for confirmatory factor analysis, especially because in its regression solution various characteristics of the model quality and fit are produced, for instance such a convenient measure as coefficient of multiple determination. When the factors structure rather corresponds to several factors, the closed-form solution can be applied to each group of variables separately to check their quality of fit in the confirmatory factor analysis. Extension of the analytical approach to constructing several factors in the exploratory factor analysis by the subsequently reduced matrix could be tried in future research.

The closed-form solution facilitates building a general latent variable, does not require any specialized software, and is very convenient for working with big data because analytical formulae do not depend on the dimensionality in the loading calculations. Furthermore, it facilitates the interpretation of latent factors and enables deeper insights into a studied problem, which is crucial for many FA applications in social, psychologic, economics, marketing research, and other areas of human interests and activity.

References

Bartholomew, D. J. & Knott, M. (1999). *Latent Variable Models and Factor Analysis*. London: Arnold.

Brown, T. A. (2015). *Confirmatory Factor Analysis for Applied Research*, (2nd Ed.). NY: Guilford Publication Inc.

de Carvalho Yokota, R. T., Miyazaki, E. S., & Ito, M. K. (2010). Applying the triads method in the validation of dietary intake using biomarkers. *Reports in Public Health* (Cadernos de Saúde Pública, Rio de Janeiro, Brazil), 26(11), 2027-2037. doi: 10.1590/s0102-311x2010001100004

Dillon, W. R., & Goldstein, M. (1984). *Multivariate Analysis: Methods and Applications.* NY: Wiley.

Djekic-Ivankovic, M., Weiler, H. A., Nikolic, M., Kadvan, A., Gurinovic, M., Mandic, L. M., & Glibetic, M. (2016). Validity of an FFQ assessing the vitamin D intake of young Serbian women living in a region without food fortification: the method of triads model. *Public Health Nutrition, 19*(3), 437-445. doi: 10.1017/s136898001500138x

Drton, M., Sturmfels, B., & Sullivant, S. (2007). Algebraic factor analysis: tetrads, pentads and beyond. *Probability Theory Related Fields, 138*(3/4), 463–493. doi: 10.1007/s00440-006-0033-2

Elden, L. (2007). *Matrix Methods in Data Mining and Pattern Recognition*. Philadelphia, PA: SIAM. doi: 10.1137/1.9780898718867

Gorsuch, R. L. (2014). *Factor Analysis*. Abington, UK: Routledge. 10.4324/9781315735740

Härdle, W., & Hlavka, Z. (2007) *Multivariate Statistics: Exercises and Solutions*. NY: Springer.

Härdle, W., & Simar, L. (2012). *Applied Multivariate Statistical Analysis*. NY: Springer. doi: 10.1007/978-3-642-17229-8

Harman, H. H. (1976). *Modern Factor Analysis*. Chicago, IL: University of Chicago Press.

Izenman, A. J. (2008). *Modern Multivariate Statistical Techniques*. NY: Springer. doi: 10.1007/978-0-387-78189-1

Kabagambe, E. K., Baylin, A., Allan, D. A., Siles, X., Spiegelman, D., & Campos, H. (2001). Application of the Method of Triads to Evaluate the Performance of Food Frequency Questionnaires and Biomarkers as Indicators of Long-term Dietary Intake. *American Journal of Epidemiology*, *154*(12), 1126-1135. doi: 10.1093/aje/154.12.1126

Langford, E., Schwertman, N., & Owens, M. (2001). Is The Property Of Being Positively Correlated Transitive? *The American Statistician*, *55*(4), 322-325. doi: 10.1198/000313001753272286

Lawley, D. N., & Maxwell, A. E. (1971). *Factor Analysis as a Statistical Method*. NY: Elsevier.

Lipovetsky, S. (2002). Comment on the paper by Langford et al. “Is the Property of Being Positively Correlated Transitive?” *The American Statistician, 56*, 341-342.

Lipovetsky, S. (2009). PCA and SVD with nonnegative loadings. *Pattern Recognition, 42*(1), 68-76. doi: 10.1016/j.patcog.2008.06.025

Lipovetsky, S. (2015). MANOVA, LDA, and FA Criteria in Clusters Parameter Estimation. *Cogent Mathematics, 2*(1). doi: 10.1080/23311835.2015.1071013

Lipovetsky, S. (2017). Factor Analysis by Limited Scales: Which Factors to Analyze? *Journal of Modern Applied Statistical Methods, 16*(1), 233-245. doi: 10.22237/jmasm/1493597520

Lipovetsky, S., & Conklin, M. (2004) Enhance-Synergism and Suppression Effects in Multiple Regression. *Int. J. of Mathematical Education in Science and Technology, 35*(3), 391-402. doi: 10.1080/00207390310001658492

Lipovetsky, S., & Conklin, M. (2005) Singular Value Decomposition in Additive, Multiplicative, and Logistic Forms. *Pattern Recognition, 38*(7), 1099-1110. doi: 10.1016/j.patcog.2005.01.010

McNaughton, S. A., Marks, G. C., Gaffney, P., Williams, G., & Green, A. (2005). Validation of a food-frequency questionnaire assessment of carotenoid and vitamin E intake using weighed food records and plasma biomarkers: the method of triads model. *European Journal of Clinical Nutrition, 59*(2), 211-218. doi: 10.1038/sj.ejcn.1602060

Ocké, M. C., & Kaaks, R. J. (1997). Biochemical markers as additional measurements in dietary validity studies: application of the method of triads with examples from the European Prospective Investigation into Cancer and Nutrition. *American Journal of Clinical Nutrition, 65*(4), 1240S-1245S. doi: 10.1093/ajcn/65.4.1240s

Revelle, W. (2017). Procedures for Personality and Psychological Research, Northwestern University, Evanston, Illinois, USA. Retrieved from https://CRAN.R-project.org/package=psych Version = 1.7.5.

Skrondal, A. & Rabe-Hesketh, S. (2004). *Generalized Latent Variable Modeling*. Boca Raton, FL: Chapman & Hall/CRC. doi: 10.1201/9780203489437

Spearman, C. (1904). General Intelligence Objectively Determined and Measured. *American Journal of Psychology*, *15*(2), 201-293. doi: 10.1037/11491-006

Spearman, C. (1927). *The Abilities of Man.* NY: Macmillan.