CONNECTIONS BETWEEN SINGLE-FACTOR ANALYSIS AND GRAPHICAL GAUSSIAN MODELS

MARIA DE FÁTIMA SALGUIERO, PETER W. F. SMITH,
JOHN W. MCDONALD

ABSTRACT

The classical single-factor model is parametrized as a graphical Gaussian model. The relationship between the classical parametrization of the single-factor model and this alternative parametrization is derived. This relationship provides extra insights into the single-factor model, which facilitates power calculations. The overall power of the first step of a backward elimination model selection procedure to detect an association structure between manifest variables compatible with a single-factor model is investigated. The results are illustrated using a one-factor congeneric measurement model.

Southampton Statistical Sciences Research Institute
Methodology Working Paper M06/13

Connections between single-factor analysis and graphical Gaussian models

M. Fátima Salgueiro [*,1]

Departamento de Métodos Quantitativos, ISCTE Business School,
Av. Forças Armadas 1649-026 Lisboa, Portugal

Peter W.F. Smith and John W. McDonald

Southampton Statistical Sciences Research Institute, University of Southampton,
Southampton SO17 1BJ, United Kingdom

Abstract

The classical single-factor model is parametrized as a graphical Gaussian model. The relationship between the classical parametrization of the single-factor model and this alternative parametrization is derived. This relationship provides extra insights into the single-factor model, which facilitates power calculations. The overall power of the first step of a backward elimination model selection procedure to detect an association structure between manifest variables compatible with a single-factor model is investigated. The results are illustrated using a one-factor congeneric measurement model.

Key words: Factor analysis, Graphical Gaussian model, Parallel measures, Partial correlation, Power

1 Introduction

Factor analysis is a classical approach to modeling multivariate data where all variables are treated on an equal footing. Graphical Gaussian models can be

[*] Corresponding author. Phone: +351-963077654. Fax:+351-217903941
Email addresses: fatima.salgueiro@iscte.pt (M. Fátima Salgueiro), pws@soton.ac.uk (Peter W.F. Smith), bigmac@soton.ac.uk (John W. McDonald).

[1] This research was funded by the Fundação para a Ciência e a Tecnologia and by the Southampton Statistical Sciences Research Institute.

considered an alternative approach for investigating the association structure of multivariate normal random variables. Instead of covariances or correlations between manifest variables, a graphical modeler is interested in partial correlations.

Several authors have considered the use and importance of incorporating latent variables in the graphical models framework, mainly concerning the identification of factor analysis models with correlated residuals. Giudici and Stanghellini [3] define the *graphical factor analysis model* as a factor model with correlated residuals and give a sufficient condition for the identification of a factor model with an arbitrary number of factors, to some extent generalizing Stanghellini [12]. Grzebyk et al. [4] build on the work of Vicard [14] and propose conditions for the identification of multi-factor models with correlated residuals. Stanghellini and Wermuth [13] consider path analysis models with uncorrelated residuals, with one hidden (latent) variable, and address the question of the identification of such models.

The current paper investigates the relationship between the classical single-factor model with no correlated residuals and graphical Gaussian models. Models are represented by undirected conditional independence graphs, and associations between each manifest variable and the latent variable are measured by partial correlation coefficients. By relating the two parameterizations, the paper presents results that provide extra insights into the single-factor model, which facilitates power calculations for single-factor models. Sections 2 and 3 review single-factor models and graphical Gaussian models, respectively. Section 4.1 demonstrates how to parameterize the single-factor model using partial correlations. Section 4.2 further investigates the relationship between the two parametrizations of the single-factor model. Implications for the single-factor model with parallel measures are considered in Section 4.3. Section 5 presents formulas to estimate the power of selecting a graphical Gaussian model consistent with a single-factor model, and illustrates their use for a single-factor model with parallel measures. Section 6 contains a discussion.

2 The classical single-factor model

2.1 The classical parametrization

The classical single-factor model can be written as $\boldsymbol{X_M} = \boldsymbol{\lambda} L + \boldsymbol{\delta}$, where $\boldsymbol{X_M}$ is the vector of the p manifest variables $(X_1, X_2, \ldots, X_p)$, L is the factor or latent variable, $\boldsymbol{\lambda}$ is a $p \times 1$ vector of factor loadings, and $\boldsymbol{\delta}$ is a vector of p variables representing random measurement error and indicator specificity. Variables are considered to be measured as deviations from their means, that is $E[\boldsymbol{X_M}] = \mathbf{0}$ and $E[L] = 0$. The model assumes that $E[L\boldsymbol{\delta}] = \mathbf{0}$, $E[\boldsymbol{\delta}] = \mathbf{0}$, $\mathrm{var}[\boldsymbol{\delta}]$ is diagonal and $\boldsymbol{X_M}$ and $\boldsymbol{\delta}$ are multivariate normal. To avoid the basic problem of identification the latent variable L is scaled to have unit variance. The variance matrix for $\boldsymbol{X_M}$, with elements denoted by σ_{ij}, is $\boldsymbol{\Sigma_M} = \boldsymbol{\lambda}\boldsymbol{\lambda}^T + \boldsymbol{\Theta}$, where $\boldsymbol{\Theta}$ is the $p \times p$ diagonal variance matrix of $\boldsymbol{\delta}$.

Anderson and Rubin [1] gave a necessary and sufficient condition for the identification of a single-factor model: at least three factor loadings have to be non-zero. Thus, the elements of $\boldsymbol{\Theta}$ and $\boldsymbol{\lambda}$ can be expressed uniquely (up to the simultaneous sign change) as a function of the elements in $\boldsymbol{\Sigma_M}$. Also, Anderson and Rubin [1, Theorem 4.2] stated that '*a necessary and sufficient condition that $\boldsymbol{\Sigma_M}$ is a variance matrix of a factor analysis model with one factor is that $p(p-1)/2 - p$ independent tetrad conditions are satisfied and*

$$0 \leq \frac{\sigma_{ki}\sigma_{ij}}{\sigma_{kj}} \leq \sigma_{ii}$$

for one pair $(j \neq k)$ for each i.' Note that throughout this paper different letters denote distinct indices, which range from 1 to p. The $p(p-1)/2 - p$ tetrad conditions to be satisfied are of the type $\sigma_{ki}\sigma_{lj} - \sigma_{li}\sigma_{kj} = 0$, for all i, j, k and l. When $p = 3$ no additional conditions have to be satisfied, whereas when $p = 4$ the two tetrad conditions are given by $\sigma_{12}\sigma_{34} - \sigma_{14}\sigma_{23} = 0$ and $\sigma_{13}\sigma_{24} - \sigma_{14}\sigma_{23} = 0$.

Alternatively, if the population correlation matrix, denoted by $\boldsymbol{P}$ with elements ρ_{ij}, is used, the following results hold, provided the tetrad conditions $\rho_{ki}\rho_{lj} - \rho_{li}\rho_{kj} = 0$, for all i, j, k and l hold:

$$\begin{aligned}
&\rho_{ij} = \lambda_i\lambda_j \text{ and } \lambda_i^2 = \frac{\rho_{ij}\rho_{ik}}{\rho_{jk}};\\
&\theta_{ii} \geq 0 \Rightarrow \lambda_i^2 \leq 1 \text{ and } \frac{\rho_{ij}\rho_{ik}}{\rho_{jk}} \leq 1;\\
&\lambda_i^2 \geq 0 \Rightarrow \frac{\rho_{ij}\rho_{ik}}{\rho_{jk}} \geq 0.
\end{aligned} \tag{1}$$

2.2 Parallel measures

Tests of parallelism and parallel measures are particularly important in psychology. Parallel measures have equal true score variances and equal error variances. If the single-factor model is thought of as a one-factor congeneric measurement model, the observed measures $X_1, X_2, \ldots, X_p$ are parallel measures if $\lambda_1 = \lambda_2 = \cdots = \lambda_p$ and $\theta_{11} = \theta_{22} = \cdots = \theta_{pp}$. The implications of parallelism in terms of partial correlations between observed measures will be considered in Section 4.3.

3 Graphical Gaussian models

Graphical Gaussian models are parametric statistical models for multivariate normal random variables. A graphical Gaussian model is specified by setting one or more elements of the inverse covariance matrix to zero. The test of whether an element in the inverse covariance matrix is zero is equivalent to the test of conditional independence between the corresponding variables, given the remaining variables. The independence structure of the variables is displayed using a mathematical graph, the conditional independence graph. Each variable is represented by a vertex (node), associations between variables being represented by edges: either lines or arrows. The interpretation of the association structure among the variables can be directly read from the graph, using the Markov properties. In brief: two vertices are connected if there is an association between them; two vertices are not connected if the corresponding variables are conditionally independent. A complete independence graph represents a model with no conditional independencies between variables. For an introduction to graphical Gaussian models see, for example, Edwards [2], Lauritzen [5] or Whittaker [15].

3.1 Notation and definitions

For the vector of random variables $\boldsymbol{X}$, of dimension q, the corresponding set of vertices is given by $\mathcal{V} = \{1, 2, \ldots, q\}$. An undirected graph is the *conditional independence graph* for $\boldsymbol{X}$ if there is no edge between X_i and X_j when X_i and X_j are conditionally independent given the remaining $q-2$ variables, that is,

$$X_i \perp\!\!\!\perp X_j \mid X_{\mathcal{V}\setminus(i,j)} \Leftrightarrow (i,j) \ \notin \ \mathcal{E},$$

where $\mathcal{E}$ is the edge set. Directed independence graphs allow for the representation of the lack of symmetry in the roles played by the variables. Markov properties relate the conditional independence structure of the random vector to the structure of a graph, and may differ for directed and undirected graphs. Figure 1 displays two Markov equivalent graphs: a directed acyclic graph (DAG), in panel a), and an undirected conditional independence graph, in panel b). Both graphs state that variables X_1, X_2 and X_3 are conditionally independent, given variable X_4.

The edge set of the complete graph is given by $\{(i,j) : i,j \in \mathcal{V},\ i < j\}$ and has $q(q-1)/2$ elements. The scaled (to have ones on the diagonal) inverse variance matrix of the underlying multivariate normal distribution is denoted by $sc(\boldsymbol{\Sigma}^{-1})$. This matrix has diagonal elements equal to unity. The off-diagonal element (i,j) equals the negative of the partial correlation between variables i and j, after conditioning on the remaining variables, $X_{\mathcal{V}\setminus(i,j)}$ ("the rest"),

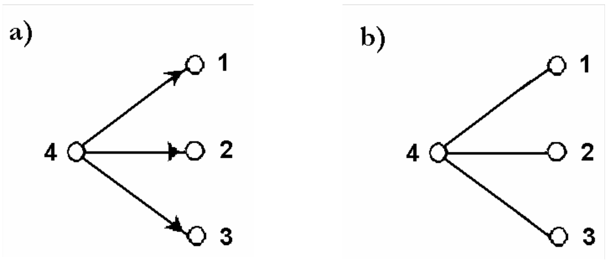

Fig. 1. Two Markov equivalent graphs: a DAG (panel a) and an undirected graph (panel b).

$\rho_{ij.R}$. It is well known that $X_i \perp\!\!\!\perp X_j \,|\, X_{\mathcal{V}\setminus(i,j)} \Leftrightarrow \rho_{ij.R} = 0$; (see, for example, Whittaker [15, p. 143]).

3.2 Edge exclusion tests

When searching for a well-fitting graphical model it is usual first to test for exclusion of each edge, in turn, from the saturated model, i.e., to perform the first step of a backward elimination model selection procedure. Traditionally the likelihood ratio test has been used, although the Wald or the efficient score test can be used. Closed-form expressions for the test statistics for single-edge exclusion from the saturated graphical Gaussian model were derived by Smith and Whittaker [11]. Under the null hypothesis that variables i and j are conditionally independent given the remaining variables in the model, i.e., the edge between i and j is absent from the independence graph, the non-signed versions of the three test statistics for single-edge exclusion have asymptotically a χ^2_1 distribution and the signed square-root versions have asymptotically a $N(0, 1)$ distribution.

Salgueiro et al. [8] studied the distributions of the Wald, the score and the likelihood ratio test statistics under the alternative hypothesis that the saturated model holds and proposed asymptotic normal approximations. Formulas for the means, variances and covariances of the test statistics, in the asymptotic distribution, are presented in detail there. The authors concluded that, at small distances from the null, the approximations for the signed square-root test statistics perform better than those for the non-signed test statistics. They also concluded that there was no difference in performance of the three test statistics. Therefore, for simplicity, in this paper attention is restricted to the signed square-root score test statistic $S_{ij} = n^{1/2}\widehat{\rho}_{ij.R}$, where the hat denotes the maximum likelihood estimate.

Formulas for the expectations, variances and covariances of the signed square-root version of the score test statistic, in the asymptotic distribution, based on a sample of size n, are (Salgueiro et al. [8, Table 2])

$$\begin{aligned} E(S_{ij}) &= n^{1/2}\,\rho_{ij.R} \\ \operatorname{var}(S_{ij}) &= (1-\rho^2_{ij.R})^2 \\ \operatorname{cov}(S_{ij},S_{kl}) &= \frac{1}{4}\,C_\rho\,(\rho_{ij.R}\rho_{kl.R})^{-1}, \text{ where} \end{aligned} \tag{2}$$

$$\begin{aligned} C_\rho = {} & 4\,\rho_{ij.R}\rho_{ik.R}\rho_{jl.R}\rho_{kl.R} + 4\,\rho_{ij.R}\rho_{il.R}\rho_{jk.R}\rho_{kl.R} + 4\,\rho^2_{ij.R}\rho_{ik.R}\rho_{il.R}\rho_{kl.R} \\ & +4\,\rho^2_{ij.R}\rho_{jk.R}\rho_{jl.R}\rho_{kl.R} + 4\,\rho_{ij.R}\rho_{ik.R}\rho_{jk.R}\rho^2_{kl.R} + 4\,\rho_{ij.R}\rho_{il.R}\rho_{jl.R}\rho^2_{kl.R} \\ & +2\,\rho^2_{ij.R}\rho^2_{jl.R}\rho^2_{kl.R} + 2\,\rho^2_{ij.R}\rho^2_{jk.R}\rho^2_{kl.R} + 2\,\rho^2_{ij.R}\rho^2_{il.R}\rho^2_{kl.R} + 2\,\rho^2_{ij.R}\rho^2_{kl.R}\rho^2_{ik.R}. \end{aligned}$$

4 The relationship between the classical and the graphical Gaussian parametrization

4.1 Parameterizing the single-factor graphical Gaussian model

The single-factor model is often represented by a directed acyclic graph. However, due to Markov equivalence of the two graphs presented in Figure 1, it is possible to represent the single-factor graphical Gaussian model using an undirected conditional independence graph. Indeed, if in Figure 1, panel b), variable X_4 is not manifest but latent, the conditional independence graph represents a single-factor model with three manifest random variables, X_1, X_2 and X_3 and a single latent variable L.

Partition $\boldsymbol{X}$ as $\left[\boldsymbol{X_M}^T \middle| L\right]^T$, with positive definite scaled inverse variance matrix $\boldsymbol{T}$ partitioned as

$$\boldsymbol{T} = \left[\begin{array}{c|c} \boldsymbol{I} & \boldsymbol{T_{ML}} \\ \hline \boldsymbol{T_{LM}} & 1 \end{array}\right], \tag{3}$$

where $\boldsymbol{I}$ is a $p\times p$ identity matrix since the single-factor model assumes that the manifest variables are conditionally independent, given the latent variable. The $p\times 1$ vector $\boldsymbol{T_{ML}} = \boldsymbol{T_{LM}}^T$ contains the non-zero elements $-\tau_{iL.R}$, the negative of the partial correlation between manifest variable X_i and latent variable L, where R denotes the remaining $p-1$ variables in $\boldsymbol{X}$, after removing X_i and L.

Salgueiro et al. [10, Result 2] have shown that marginalizing the single-factor graphical Gaussian model over the latent variable L yields a joint distribution for the manifest variables with no conditional independencies, that is, with no zero entries in the scaled inverse variance matrix of the manifest variables. Consequently, the corresponding independence graph is a complete graph, with all edges present. One should note that, in the three manifest variables case, when no tetrad conditions have to be satisfied, the resulting model for the manifest variables is the saturated model. When there are four or more manifest variables, the tetrad conditions have to be satisfied, imposing a structure on the partial correlations among manifest variables, and therefore the model obtained has a complete graph, but is not necessarily the saturated model.

For the scaled inverse variance matrix of the manifest variables to be positive definite all variances $1-\tau^2_{iL.R}$ have to be positive, implying that $0<\tau^2_{iL.R}<1$. Note that whereas in the classical factor model the factor loadings $\lambda^2_i \in [0,1]$ (since positive semidefinite matrices are allowed), in the current paper $\tau^2_{iL.R} \in \ (0,1)$, since only positive definite matrices are considered. The partial correlations between manifest variables can be written as a function of the τs (Salgueiro et al. [10, Result 3])

$$\rho_{ij.R} = \frac{\tau_{iL.R}\,\tau_{jL.R}}{\{(1-\tau^2_{iL.R})(1-\tau^2_{jL.R})\}^{1/2}}, \qquad i \neq j \in \{1,2,\ldots,p\}. \tag{4}$$

4.2 The relationship between the two parametrizations

From Section 2.1, using the classical parametrization of the single-factor model, given the latent variable L, the manifest variables are normally distributed with mean vector $\boldsymbol{\lambda}L$ and variance matrix $\boldsymbol{\Theta}$. If L has variance one, the vector of factor loadings is $\boldsymbol{\lambda}=\boldsymbol{\Sigma_{ML}}$, the $p\times 1$ vector of covariances between the manifest variables and the latent variable. The diagonal matrix with the variances of the measurement error terms is $\boldsymbol{\Theta} = \text{diag}\,\{\boldsymbol{\Sigma_M}-\boldsymbol{\Sigma_{ML}}\boldsymbol{\Sigma_{ML}}^T\}$. From Equation 3, the inverse of $\boldsymbol{T}$ is given by

$$\boldsymbol{T}^{-1} = \left[\begin{array}{c|c} \boldsymbol{I} & \boldsymbol{T_{ML}} \\ \hline \boldsymbol{T_{LM}} & 1 \end{array}\right]^{-1} = \begin{bmatrix} (\boldsymbol{I}-\boldsymbol{T_{ML}}\boldsymbol{T_{LM}})^{-1} & -(\boldsymbol{I}-\boldsymbol{T_{ML}}\boldsymbol{T_{LM}})^{-1}\boldsymbol{T_{ML}} \\ -\boldsymbol{T_{LM}}(\boldsymbol{I}-\boldsymbol{T_{ML}}\boldsymbol{T_{LM}})^{-1} & (1-\boldsymbol{T_{LM}}\boldsymbol{T_{ML}})^{-1} \end{bmatrix}. \tag{5}$$

By noting that, up to scaling, $\boldsymbol{\Sigma_{ML}}$ is the upper right submatrix of $\boldsymbol{T}^{-1}$, it follows from Equation 5 when

$$\boldsymbol{\lambda}^* = -(\boldsymbol{I}-\boldsymbol{T_{ML}}\boldsymbol{T_{LM}})^{-1}\boldsymbol{T_{ML}} \tag{6}$$

is a vector of factor loadings, $\text{var}(L)=(1-\boldsymbol{T_{LM}}\boldsymbol{T_{ML}})^{-1}$, and

$$\boldsymbol{\Theta}^* = \text{diag}\,\{\boldsymbol{\Sigma_M}-\text{var}(L)\boldsymbol{\lambda}\boldsymbol{\lambda}^T\} \tag{7}$$

is the corresponding diagonal matrix with the variances of the error terms.

Consequently, each λ^*_i can be written as a function of the τs as

$$\lambda^*_i = \frac{\tau_{iL.R}}{1-\sum_{k=1}^{p}\tau^2_{kL.R}}. \tag{8}$$

The proof follows. From Rao [7, p. 33]

$$(\boldsymbol{I} - \boldsymbol{T_{ML}}\boldsymbol{T_{LM}})^{-1} = \boldsymbol{I} + \boldsymbol{T_{ML}}\boldsymbol{T_{LM}}/(1 - \boldsymbol{T_{LM}}\boldsymbol{T_{ML}}).$$

Since $\boldsymbol{T_{LM}}\boldsymbol{T_{ML}}$ is a scalar and equals $\sum_{k=1}^{p} \tau_{kL.R}^2$,

$$\boldsymbol{\lambda}^* = -\frac{1}{1 - \sum_{k=1}^{p} \tau_{kL.R}^2}\boldsymbol{T_{ML}} \quad \text{and} \quad \lambda_i^* = \frac{\tau_{iL.R}}{1 - \sum_{k=1}^{p} \tau_{kL.R}^2}.$$

Because the denominator is always positive, λ_i^* has the same sign as the corresponding $\tau_{iL.R}$.

Also, the diagonal elements in $\boldsymbol{\Theta}^*$ can be written as a function of the τs as

$$\theta_{ii}^* = \frac{1 - \sum_v \tau_{vL.R}^2}{1 - \sum_{k=1}^{p} \tau_{kL.R}^2} - \frac{\tau_{iL.R}^2}{(1 - \sum_{k=1}^{p} \tau_{kL.R}^2)^3}, \tag{9}$$

with $v = 1, \ldots, p$ and $v \neq i$.

The "classical standardized solution" for the factor loadings is obtained by dividing each λ_i^* by the square root of the diagonal elements of $\boldsymbol{T}^{-1}$ corresponding to manifest variable X_i and latent variable L. Let $\boldsymbol{\lambda}^{sc}$ denote the vector of factor loadings in the "classical standardized solution". The variance matrix in the "classical standardized solution" is given by $\boldsymbol{\Theta}^{sc} = \text{diag}\ \{\boldsymbol{I} - \boldsymbol{\lambda}^{sc}(\boldsymbol{\lambda}^{sc})^T\}$.

The derived relationships between the classical parametrization of the single-factor model and the parametrization of the single-factor graphical Gaussian model hold theoretically (for the population parameters), for the general p manifest variables case once the tetrad conditions are fulfilled.

4.3 Implications of parallel measures on the structure of partial correlations

In a single-factor model the observed measures $X_1, X_2, \ldots, X_p$ are parallel measures if they have equal factor loadings (λ) and equal error variances (θ). Consequently, the variance matrix $\boldsymbol{\Sigma_M}$ has diagonal elements $\lambda^2 + \theta$ and off-diagonal elements λ^2. Its inverse, $\boldsymbol{\Sigma_M}^{-1}$, has diagonal elements $\frac{(p-1)\lambda^2+\theta}{\theta(p\lambda^2+\theta)}$ and off-diagonal elements $\frac{-\lambda^2}{\theta(p\lambda^2+\theta)}$. The scaled (to have ones on the diagonal) inverse variance matrix of the manifest variables has off-diagonal elements of the type $\frac{-\lambda^2}{(p-1)\lambda^2+\theta}$. Therefore, all partial correlations between manifest variables are equal, and of the form

$$\rho_{ij.R} = \frac{\lambda^2}{(p-1)\lambda^2 + \theta}.$$

In order to obtain the positive definiteness constraint for the scaled inverse variance matrix of p manifest variables in the case of equal partial correlations, let $\boldsymbol{E}$ be an *equicorrelation matrix* (see Mardia et al. [6, p. 461]). It is a $p \times p$ matrix of the type $(1-\rho)\boldsymbol{I} + \rho\boldsymbol{J}$, with ones on the main diagonal and off-diagonal elements equal to the correlation coefficient ρ. $\boldsymbol{J}$ denotes a $p \times p$ matrix of ones. The eigenvalues of $\boldsymbol{E}$ are $\lambda_1 = 1+(p-1)\rho$, $\lambda_2 = \ldots = \lambda_p = 1-\rho$ and $\boldsymbol{E}$ is positive definite when all eigenvalues are positive, that is, when $\rho \in \left(\frac{-1}{p-1}\,,\ 1\right)$.

Because the focus of this paper is on partial correlations, existing results for an equicorrelation matrix have to be adapted. Indeed, the scaled inverse variance matrix, with ones on the main diagonal and off-diagonal elements being minus the partial correlation coefficients, can be written as

$$\boldsymbol{E}^{\boldsymbol{sc}} = sc(\boldsymbol{E}^{-1}) = (1+\rho_{ij.R})\boldsymbol{I} - \rho_{ij.R}\boldsymbol{J}. \tag{10}$$

The eigenvalues of $\boldsymbol{E}^{\boldsymbol{sc}}$ are $\lambda_1 = 1-(p-1)\rho_{ij.R}$, $\lambda_2 = \ldots = \lambda_p = 1+\rho_{ij.R}$ and $\boldsymbol{E}^{\boldsymbol{sc}}$ is positive definite when $1+\rho_{ij.R} > 0$ and $1-\rho_{ij.R}(p-1) > 0$. Because $\rho_{ij.R} \in (-1\,,\ 1)$, $1+\rho_{ij.R}$ is always positive and $[1-\rho_{ij.R}(p-1)]$ is strictly positive if $\rho_{ij.R} < \frac{1}{p-1}$. In other words, the positive definiteness constraint in the scaled inverse variance matrix with p variables, when all partial correlations $\rho_{ij.R}$ are equal, is that $\rho_{ij.R} \in \left(-1\,,\ \frac{1}{p-1}\right)$.

If the partial correlations arise from marginalizing the single-factor model over the latent variable, the additional constraint that the product of any three partial correlations has to be positive has to be imposed, and therefore $\rho_{ij.R} \in \left(0\,,\ \frac{1}{p-1}\right)$. From Equation 4, all $\rho_{ij.R}$ equal implies all $\tau_{iL.R}$ equal and of the form

$$\tau^2_{iL.R} = \frac{\rho_{ij.R}}{1+\rho_{ij.R}}.$$

The constraint imposed on the partials implies $\tau^2_{iL.R} < 1/p$ with $\tau_{iL.R} \neq 0$, restricting considerably the parameter space for the τs as the number of manifest variables increases.

5 Power

From Salgueiro et al. [10, Result 2] it follows that marginalizing the single-factor graphical Gaussian model over the latent variable induces an independence graph for the manifest variables that is complete, hence none of the population partial correlations should be zero. Taking into account sampling variability, it is of interest to test if all observed partial correlations are “large enough”, in absolute value. In practice the data analyst can perform the first step of a backward elimination model selection procedure, and test for a set of null hypotheses of zero partial correlations (i.e., conditional independencies).

Rejecting all of them implies favoring a model where all edges between manifest variables are present, hence providing support for a single-factor model. On the other hand, the data analyst wants to have adequate power to conclude that if an edge is absent, then the manifest variables could not have arisen from a single-factor model. The current section addresses power issues.

5.1 Theoretical power of selecting a model with a complete graph

Power for single edge exclusion from a saturated graphical Gaussian model has been investigated by Salgueiro et al. [8]. For a two-sided test of size α, the null hypothesis that $\rho_{ij.R}$ equals zero is rejected if the absolute value of the signed square-root test statistic is greater than $z_{1-\alpha/2}$, where $z_{1-\alpha}$ is the upper α quantile of the standard normal distribution. Hence, the power for the two-sided signed square-root score test of excluding edge (i,j) from the saturated graphical Gaussian model can be approximated by

$$P\Big[|S_{ij}| > z_{1-\alpha/2}\,|\,\rho_{ij.R}\Big] \simeq P\left[Z < \frac{z_{\alpha/2} - E(S_{ij})}{\sqrt{\text{var}(S_{ij})}}\right] + P\left[Z > \frac{z_{1-\alpha/2} - E(S_{ij})}{\sqrt{\text{var}(S_{ij})}}\right]. \tag{11}$$

The power for a one-sided test of size $\alpha/2$ is approximated by either the first or the second term on the right-hand side of Equation 11, depending on the direction of the alternative hypothesis.

5.2 Illustration of power calculations

This section illustrates power calculations in the particular case of all partial correlations equal. Theoretical results presented in Section 5.1 are used for the signed square-root score test statistic.

Recall that the positive definiteness constraint on the scaled inverse variance matrix of the manifest variables when all $\rho_{ij.R}$ are equal implies that $\rho_{ij.R} \in \left(-1\,,\, \frac{1}{p-1}\right)$. Additionally, the fact that the association structure between manifest variables arises from marginalizing the single-factor model over the latent factor further requires that $\rho_{ij.R} > 0$.

For all $\rho_{ij.R}$ equal, from Equation 2, the formula for the asymptotic covariance of the signed square-root score test statistics simplifies to $\text{cov}(S_{ij}, S_{kl}) = 2\rho^2_{ij.R}\left(1 + \rho_{ij.R}\right)^2$. Since the $\rho_{ij.R}$ are positive, one-sided tests are considered. Figure 2 presents the estimated overall power curves for association structures between manifest variables with all $\rho_{ij.R}$ equal, from 0 to $\frac{1}{p-1}$. The three, four and five manifest variables cases are considered, with four different sample sizes: 50, 100, 200 and 500.

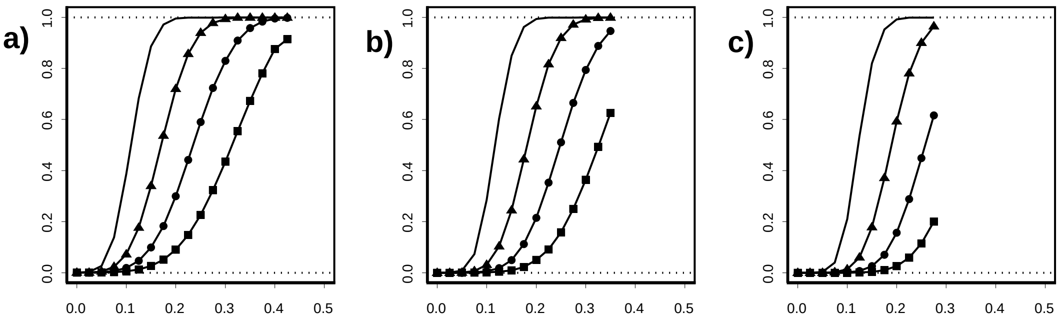

Fig. 2. Overall theoretical power curves for association structures between manifest variables with all $\rho_{ij.R}$ equal, from 0 to $1/(p-1)$. $p = 3$, 4 and 5, respectively in panels a), b) and c). Sample sizes of 50 (squares), 100 (circles), 200 (triangles) and 500 (solid line). The horizontal dotted lines correspond to power values of 0 and 1.

Some conclusions can be drawn from Figure 2:

(1) power increases as partial correlations departure from zero, faster for larger sample sizes;
(2) for $n = 50$, the probability of selecting the saturated model for the three manifest variables (when $\rho_{ij.R} = 0.25$) equals 0.23. This probability goes up to 0.6 for $n = 100$, reaching 0.94 for $n = 200$;
(3) as the number of variables p increases, the probability of selecting a model with a complete graph tends to decrease, for a given sample size.

6 Discussion

The investigation into the relationship between the classical single-factor model with no correlated residuals and graphical Gaussian models has provided some useful results for the practitioner. Inspection of sample partial correlations between manifest variables can provide evidence for and against a single-factor model. Salgueiro et al. [10, Result 2] states that population partial correlations between manifest variables in a single-factor model should be non-zero. Therefore, provided there is adequate overall power, sample partial correlations not significantly different from zero rule out a single-factor model. A further check on the compatibility of the association structure of the manifest variables with a single-factor model can be made by inspecting the pattern of signs in the partial correlation matrix (Salgueiro et al. [10, Result 5]). Hence, provided the data analyst has concluded that all partials are significantly different from zero, this result can also be used to assess compatibility. In the parallel measures case, this result implies that all partial correlations must be positive.

Hence, a single significant negative sample partial correlation would rule out a single-factor model. However, if overall power is small, a single-factor model should still be considered, even if there are non-significant sample partial correlations.

It may be possible to extend some of the results presented in this paper to latent class models by using the results in Salgueiro et al. [9].

References

[1] T.W. Anderson and H. Rubin, Statistical inference in factor analysis, In *Proceedings of the third Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, Vol. V, 1956, 111-150.

[2] D. Edwards, *Introduction to Graphical Modelling*, New York: Springer-Verlag, 2nd edition, 2000.

[3] P. Giudici and E. Stanghellini, Bayesian inference for graphical factor analysis models, *Psychometrika*, 66, 2001, 577–592.

[4] M. Grzebyk, P. Wild, D. Chouanière, On identification of multi-factor models with correlated residuals, *Biometrika*, 91, 2004, 141–151.

[5] S.L. Lauritzen, *Graphical Models*, Oxford: Clarendon Press, 1996.

[6] K. Mardia, J. Kent, J. Bibby, *Multivariate Analysis*, Academic Press, 1979.

[7] C.R. Rao, *Linear Statistical Inference and Its Applications*, John Wiley & Sons, 2nd edition, 1973.

[8] M.F. Salgueiro, P.W.F. Smith, J.W. McDonald, Power of edge exclusion tests in graphical Gaussian models, *Biometrika*, 92, 2005, 173–182.

[9] M.F. Salgueiro, P.W.F. Smith, J.W. McDonald, Power of edge exclusion tests in graphical log-linear models, *Journal of Multivariate Analysis*, 97, 2006, 1691–1701.

[10] M.F. Salgueiro, P.W.F. Smith, J.W. McDonald, The manifest association structure of the single-factor model: insights from partial correlations, *Southampton Statistical Sciences Research Institute, Methodological Working Paper* M06/12, 2006.

[11] P.W.F. Smith and J. Whittaker, Edge exclusion tests for graphical Gaussian models, In *Learning and Inference in Graphical Models*, Jordan, M.I.(Ed.), Dordrecht: Kluwer Academic Press, 1998.

[12] E. Stanghellini, Identification of a single-factor model using graphical Gaussian rules, *Biometrika*, 84, 1997, 241–244.

[13] E. Stanghellini and N. Wermuth, On the identification of path analysis models with one hidden variable, *Biometrika*, 92, 2005, 337–350.

[14] P. Vicard, On the identification of a single-factor model with correlated residuals, *Biometrika*, 87, 1997, 199–205.

[15] J. Whittaker, *Graphical Models in Applied Multivariate Statistics*, John Wiley & Sons, 1990.