Goodness-of-fit for Generalized Linear Latent Variables Models

CONNE, David, RONCHETTI, Elvezio, VICTORIA-FESER, Maria-Pia

Abstract

Generalized Linear Latent Variables Models (GLLVM) enable the modeling of relationships between manifest and latent variables, where the manifest variables are distributed according to a distribution of the exponential family (e.g. binomial or normal) and to the multinomial distribution (for ordinal manifest variables). These models are widely used in social sciences. To test the appropriateness of a particular model, one needs to define a Goodness-of-fit test statistic (GFI). In the normal case, one can use a likelihood ratio test or a modified version proposed by cite{SaBe:01} (S&B GFI) that compares the sample covariance matrix to the estimated covariance matrix induced by the model. In the binary case, Pearson-type test statistics can be used if the number of observations is sufficiently large. In the other cases, including the case of mixed types of manifest variables, there exists GFI based on a comparison between a pseudo sample covariance and the model covariance of the manifest variables. These types of GFI are based on latent variable models that suppose that the manifest variables are themselves induced by [...]

Reference


DOI : 10.1198/jasa.2010.tm09160

Available at:
http://archive-ouverte.unige.ch/unige:6491

Disclaimer: layout of this document may differ from the published version.
Goodness-of-fit for Generalized Linear Latent Variables Models

David Conne
Department of Econometrics
University of Geneva, Switzerland

Elvezio Ronchetti
Department of Econometrics
University of Geneva, Switzerland

Maria-Pia Victoria-Feser
HEC
University of Geneva, Switzerland

Submitted March 2009
Revised January 2010
Abstract

Generalized Linear Latent Variables Models (GLLVM) enable the modeling of relationships between manifest and latent variables, where the manifest variables are distributed according to a distribution of the exponential family (e.g. binomial or normal) and to the multinomial distribution (for ordinal manifest variables). These models are widely used in social sciences. To test the appropriateness of a particular model, one needs to define a Goodness-of-fit test statistic (GFI). In the normal case, one can use a likelihood ratio test or a modified version proposed by Satorra and Bentler (2001) (S&B GFI) that compares the sample covariance matrix to the estimated covariance matrix induced by the model. In the binary case, Pearson-type test statistics can be used if the number of observations is sufficiently large. In the other cases, including the case of mixed types of manifest variables, there exists GFI based on a comparison between a pseudo sample covariance and the model covariance of the manifest variables. These types of GFI are based on latent variable models that suppose that the manifest variables are themselves induced by underlying normal variables (underlying variable approach). The pseudo sample covariance matrices are then made of polychoric, tetrachoric or polyserial correlations. In this article, we propose an alternative GFI that is more generally applicable. It is based on some distance comparison between the latent scores and the original data. This GFI takes into account the nature of each manifest variable and can in principle be applied in various situations and in particular with models with ordinal, and both discrete and continuous manifest variables. To compute the $p$-value associated to our GFI, we propose a consistent resampling technique that can be viewed as a modified parametric bootstrap. A simulation study shows that our GFI has good performance in terms of empirical level and empirical power across different models with different types of manifest variables.
1 Introduction

When theoretical concepts such as intelligence, desirability, welfare, etc. cannot be measured directly, one uses observable indicators (manifest variables) that are supposed to be linked to the unobservable (latent) variables. The data are then analyzed using latent variable models. Jöreskog (1969) proposed such a model which is based on a linear link between normal latent variables and normal manifest variables. Exact maximum likelihood estimators (MLE) can be computed from the sample covariance matrix of the manifest variables. This model is implemented in well-known and widely used software such as LISREL (Jöreskog, 1990) or Mplus (Muthen and Muthen, 2001).

In economics, psychology or social sciences, the measures are very seldom taken on a normal scale, but rather on ordinal or binary scales, or even on mixed normal, ordinal and binary scales. In these cases, the methods implemented in these software suppose that the manifest variables are indirect observations of normal underlying variables (Jöreskog, 1990, Samejima, 1969, Muthen, 1978) so that the maximum likelihood approach for normal manifest variables can by used with the sample covariance matrix replaced by polychoric, tetrachoric or polyserial correlations in the ordinal case (see e.g. Qu, Piedmonte, and Medendorp 1995), the binary case, and the mixed normal ordinal/binary case respectively. Bartholomew (1984) and Moustaki and Knott (2000) proposed a more general framework for latent variable modeling that does not require the underlying normal variable assumption, namely Generalized Linear Latent Variables Model (GLLVM); see Section 2. These models are suitable for manifest variables that are distributed, conditionally on the latent variables, according to a (or a mix of) distribution of the exponential family and this includes the multinomial for ordinal manifest variables. Except in the normal and binary case (with probit link), GLLVM differ from the models assuming the underlying normal variable for
the manifest variables (see e.g. Jöreskog and Moustaki 2001 for the ordinal case).

In a latent variable framework, one works with several unobservable quantities (latent scores, parameters) and it is therefore essential to choose a model that makes the best compromise between parsimony and fit to the data. More specifically, the number of latent variables is clearly unknown, as well as the fact that a particular manifest variable is linked or not to a particular latent variable. To compare models with a different number of latent variables, one could use the Akaike (1973) criterion (see also e.g Vaida and Blanchard (2005) for a corrected Akaike criterion in mixed-effects models) which is a powerful measure of relative fit; see Conne (2003) for some numerical comparison with GLLVM. It is however not a goodness-of-fit criterion and in latent variable models it is difficult or arbitrary to specify an alternative model to the one under investigation, so that a goodness-of-fit criterion (GFI) is needed.

In the normal case (for the manifest variables), one can use the likelihood ratio test (see e.g Bartholomew and Knott 1999) or the modified version proposed by Satorra and Bentler (2001) that compare the sample covariance matrix and the estimated covariance matrix constrained by the latent variables model. In the binary case, the data can be represented by a contingency table and a *Pearson*-type statistic can be derived that compares the empirical frequencies to the estimated frequencies under the model. However, for a moderate sample size and an important number of variables, one is faced with the curse of dimensionality problem and this method becomes infeasible. Moreover, the asymptotic distribution holds only if the number of observations in each cell is large. In the ordinal case, the number of cells can be large and the problem is even worse. Glas (1988), Reiser (1996), Bartholomew and Leung (2002) or Maydeu-Olivares and Joe (2005) propose to use only information on lower (usually first and second) order margins. If there are significant interactions among higher order margins, the GFI may however not reject false models. Moreover,
these methods can only be used if all manifest variables are binary or ordinal, but it is unclear how to extend them to mixed types of manifest variables.

Alternatively, when manifest variables are not normal, the traditional approach is to suppose that they are indirect observations of normal underlying variables, so that a GFI based on the comparison between the estimated covariance between these underlying normal variables (using polychoric, tetrachoric or polyserial correlations) and the estimated covariance matrix constrained by the latent variables model can be used. The S&B GFI is such a GFI. However, if the underlying normal variable assumption does not necessarily hold as in the case of ordinal manifest variables, this approach has important drawbacks. Indeed, reducing the information contained in the sample to an estimated covariance matrix implies a loss of information since the covariance matrix is not a sufficient statistic. Moreover, the estimation of polychoric or polyserial correlations implies the estimation of a number of parameters that increases rapidly with the number of manifest variables. In fact, a GFI based on a function of the covariance matrix of the underlying normally distributed manifest variables in a non-normal setting tests simultaneously the normality assumption of the underlying variable and the fit of the GLLVM structure, except when all manifest variables are normal or binary, see e.g. Muthen (1993).

In this paper, we propose an alternative approach to construct a GFI which is not based on the comparison between the correlation matrices, but rather on the comparison of some distance among the latent scores and the corresponding distance among the manifest variables. The latent scores represent in a way the mapping of the observations on the latent variable space, and therefore, if the model is adequate, to a “distance” between two observations in the original data space should correspond a similar “distance” on the latent space. Let $\mathbf{x}_i$ be the $p \times 1$ vector of observed manifest variables of subject $i$, $i = 1, \ldots, n$ and $\hat{\mathbf{z}}_i$ the $q \times 1$ ($q < p$) vector of corresponding
(estimated) latent scores through the postulated GLLVM. Let also $d_q(\hat{z}_{i1}, \hat{z}_{i2})$ be a distance function on the latent scores space and $\tilde{d}_p(x_{i1}, x_{i2})$ a distance function on the data space. Our proposed GFI is given by

$$\Omega = \frac{\hat{\Lambda}}{n^2 - n} \sum_{i_1 = 1}^{n} \sum_{i_2 = 1}^{n} [d_q(\hat{z}_{i1}, \hat{z}_{i2}) - \tilde{d}_p(x_{i1}, x_{i2})]^2$$

(1)

where $\hat{\Lambda}$ is a correction factor depending on the observations (see Section 3.2). Basically, this GFI is an average squared difference between a general distance on the sample space and its estimated counterpart on the latent space. The choice of the distance functions will depend on the type of manifest variables and will be borrowed from distances used in cluster analysis (see e.g. Kaufman and Rousseeuw 1990).

Our GFI can in principle be applied in various situations and in particular with models with both discrete and continuous manifest variables. The asymptotic distribution of our GFI under the null hypothesis (i.e. the postulated GLLVM) is derived in Section 4.1 with the proof given in an on-line supplemental file (http link). It however appears to be quite difficult to implement numerically. Therefore we propose instead in Section 4.2 to compute associated $p$-values using a modified parametric bootstrap that we modify for computational speed. We show that the estimated $p$-values are consistent (Section 4.3). In a simulation study in Section 5, we consider three different models, one with normal manifest variables, a second one with ordinal manifest variables, and a third one with mixed normal, ordinal and binary manifest variables. We show that our modified parametric bootstrap produces $p$-values with uniform distribution under the null and also that the power of our GFI is quite good. A comparison is made in the normal case with the S&B GFI. The Appendix provides technical details, conditions and proof of Theorem 2, and the simulation parameters while in Section 2 we summarize the GLLVM and the estimation technique based on
the Laplace approximation developed by Huber, Ronchetti, and Victoria-Feser (2004) which provides estimated latent scores.

2 Generalized Linear Latent Variable Models (GLLVM)

The relationship between $p$ manifest variables $x^{(j)}$, $j = 1, \ldots, p$ and $q$ latent variables $z^{(k)}$, $k = 1, \ldots, q$, $q < p$, is formalized in the same manner as in generalized linear models (McCullagh and Nelder, 1989) by means of conditional distributions $g_j(x^{(j)}|z)$ belonging to the exponential family, i.e.

$$g_j(x^{(j)}|z) = \exp \left\{ \frac{x^{(j)} u_j(\alpha^{(j)^T} z) - b(u_j(\alpha^{(j)^T} z))}{\phi_j} + c(x^{(j)}, \phi_j) \right\},$$

where $\alpha^{(j)} = (\alpha_{l}^{(j)})_{l=0,1,\ldots,q} \in \mathbb{R}^{q+1}$ are the loadings, $z = (1, z_{(2)})^T$, $z_{(2)} = (z^{(1)}, \ldots, z^{(q)})^T$, $\phi_j$ is a scale parameter, $u_j(\alpha^{(j)^T} z)$ is the so-called canonical parameter, and the functions $b(u_j(\alpha^{(j)^T} z))$ and $c(x^{(j)}, \phi_j)$ depend on the specific distribution $g_j(x^{(j)}|z)$.

We define a link function between the conditional expectation of $x^{(j)}|z$ and $z$ as $\gamma^{(j)}(E[x^{(j)}|z]) = \alpha^{(j)^T} z$. We give here the specific function $u$, $b$, $c$ and $\gamma$ and the scale parameter $\phi$ for normal and ordinal conditional distributions $g_j(x^{(j)}|z)$.

- Normal manifest variables

Let $x^{(j)}|z$ have a normal distribution with mean $\mu$ and variance $\sigma_j^2$. The link function $\gamma()$ and $u()$ are the identity function, i.e. $u(\alpha^{(j)^T} z) = \alpha^{(j)^T} z$, $\gamma^{(j)}(E[x^{(j)}|z]) = E[x^{(j)}|z] = \mu = \alpha^{(j)^T} z$, and

$$b(u(\alpha^{(j)^T} z)) = \frac{(\alpha^{(j)^T} z)^2}{2}, \quad c(x^{(j)}, \phi_j) = -\frac{1}{2} \left( \frac{x^2}{\phi_j} + \log(2\pi\phi_j) \right), \quad \phi_j = \sigma_j^2.$$
• Ordinal manifest variables

Let \( x^{(j)}|z \) follow an ordered multinomial distribution with categories going from 1 to \( M^{(j)} \). The link function can be chosen as a logit function

\[
\gamma^{(j)}\left( \frac{p_{js}}{1 - p_{js}} \right) = \alpha_s^{(j)^T} z,
\]

where \( p_{js} \) is defined as the cumulative probability of a response \( x^{(j)}|z \) to be \( s \) or less, where \( s = 1, \ldots, M^{(j)} \). The \( s \) index in \( \alpha_s^{(j)} \) indicates that the first component \( \alpha_{0,s}^{(j)} \) of the vector \( \alpha^{(j)} \) is a threshold that is related to each category.

Let

\[
\begin{align*}
u(\alpha^{(j)^T} z) &= \log \left( \frac{p_{js}}{p_{j,s+1} - p_{js}} \right), \\
\beta(u(\alpha^{(j)^T} z)) &= \log \left( \frac{p_{j,s+1}}{p_{j,s+1} - p_{js}} \right),
\end{align*}
\]

then

\[
g_j(x^{(j)}|z) = \exp \left( \sum_{s=1}^{M^{(j)}-1} \left[ u(x^{(j)} \leq s) u(\alpha^{(j)^T} z) - u(x^{(j)} \leq s + 1) \beta (u(\alpha^{(j)^T} z)) \right] \right)
\]

with \( u(x^{(j)} \leq s) = 1 \) if \( x^{(j)} \leq s \) and 0 otherwise. Note that \( c(x^{(j)}, \phi_j) = 0 \) and there is no scale parameter.

The main assumption in the GLLVM is the conditional independence of the manifest variables given the latent ones. Hence, the conditional distribution is \( \prod_{j=1}^p g_j(x^{(j)}|z) \) and the marginal distribution is

\[
f_{\alpha, \phi}(x) = \int \left[ \prod_{j=1}^p g_j(x^{(j)}|z) \right] k_q(z_{(2)}) dz_{(2)}, \tag{2}
\]

where \( k_q(\cdot) \) is the density of the \( q \)-dimensional vector of latent variables and \( \alpha \) is the vector of stacked \( \alpha^{(j)} \). \( k_q(\cdot) \) is assumed to be multivariate standard normal.
correlation matrix between the latent variables can be added, but we do not consider this case in this paper.

Given a sample of \( n \) observations \( x_i = (x_i^{(1)}, \ldots, x_i^{(p)})^T, i = 1, \ldots, n \), the log-likelihood of the loadings \( \alpha \) and the scale parameters \( \phi \) (if any) is \( \ell(\alpha, \phi|x) = \sum_{i=1}^{n} \log f_{\alpha,\phi}(x_i) \). This expression contains a multidimensional integral which cannot be computed explicitly, except when \( x|z \) is multivariate normal. Bock and Liberman (1970) (see also Moustaki and Knott 2000) use a Gauss-Hermite approximation when the number of latent variables is small (at most 2). Huber, Ronchetti, and Victoria-Feser (2004) propose instead a Laplace approximation and define an estimator called LAMLE which can be viewed as an M-estimator (Huber, 1981). This allows consistent estimation and inference even in the presence of a large number of latent and manifest variables. Technical details of the LAMLE are given in Appendix A. The Laplace approximation has been used in several settings, including the Bayesian approach to approximate the posterior distribution (see e.g. Tierney and Kadane, 1986) and in a simplified form in generalized linear mixed models (see Breslow and Clayton, 1993, Pinheiro and Bates 1995). It corresponds to an adaptive Gauss-Hermite approximation with a single node (Rabe-Hesketh, Skrondal, and Pickles 2005).

The Laplace approximation allows to define the h-likelihood scores on each latent variable, the asymptotic properties of which can be found in Lee and Nelder (1996). The h-likelihood is the likelihood based on the joint density between \( x_i \) and \( z_i \) for fixed \( \alpha \) and \( \phi \), so that the estimated latent scores \( \hat{z}_i, \forall i = 1, \ldots, n \) maximize (in \( z_i \))

\[
\sum_{j=1}^{p} \log g_j(x_{i}^{(j)}|z_i) + \log k_q(z_{i(2)})
\]

in which \( \alpha \) and \( \phi \) are replaced by their LAMLE. These estimated scores can also be viewed as penalized quasi likelihood estimators (Green, 1987). In the Bayesian frame-
work, they are the modal of the posterior distribution of the scores with estimated parameters plugged in. The latter are called the Empirical Bayes modal (EBM) by Skrondal and Rabe-Hesketh (2004), or modal a posteriori (MAP) by Bock (1985). For a general overview on latent variable modeling, see also Skrondal and Rabe-Hesketh (2004).

3 Goodness-of-fit for Generalized Linear Latent Variable Models (GLLVM)

3.1 Test Statistic

To compute (1), one needs to define a general distance measure on the latent space and on the data space while taking into account the nature of the different variables. We propose here to use the distances developed in Kaufman and Rousseeuw (1990) for cluster analysis. We suppose that according to the GLLVM, each observation \( x_i \) has a corresponding (unknown) latent score \( z_i \) estimated by \( \hat{z}_i \) which maximizes (3). Let \( d_q(\hat{z}_{i1}, \hat{z}_{i2}) \) be a distance function on the scores space and \( \tilde{d}_p(\mathbf{x}_{i1}, \mathbf{x}_{i2}) \) a distance function on the data space. Since \( \hat{z} \) is continuous, a natural choice for \( d_q(\cdot, \cdot) \) is the Euclidean distance standardized by the standard deviation of \( \hat{z}_i \), i.e.

\[
d_q(\hat{z}_{i1}, \hat{z}_{i2}) = \frac{1}{q} \sqrt{\sum_{j=1}^{q} \left( \frac{\hat{z}_{i1}^{(j)} - \hat{z}_{i2}^{(j)}}{\tilde{\sigma}_z^{(j)}} \right)^2},
\]

where \( \tilde{\sigma}_z^{(j)} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (\hat{z}_i^{(j)} - \bar{z}^{(j)})^2} \) is the empirical standard deviation of the \( \hat{z}_i^{(j)} \). In the sample space, if \( x^{(j)} \) is normally distributed, the Euclidean distance function is also suitable for \( \tilde{d}_p(\cdot, \cdot) \). When \( x^{(j)} \) is ordinal, a standard choice is the Manhattan distance (\( L_1 \) distance) on the ranks \( r_i^{(j)} \) of the observations. Hence for a model with
$p_1$ normal manifest variables and $p_2$ ordinal manifest variables, we have

$$d_p(x_{i1}, x_{i2}) = \frac{1}{p_1} \sum_{j=1}^{p_1} \left( \frac{x_{i1}^{(j)} - x_{i2}^{(j)}}{\hat{\sigma}_x^{(j)}} \right)^2 + \frac{1}{p_2} \sum_{j=p_1+1}^{p_1+p_2} \left| \frac{r_{i1}^{(j)} - r_{i2}^{(j)}}{\frac{n}{2}} \right|,$$

where $\hat{\sigma}_x^{(j)}$ is the empirical standard deviation of $x^{(j)}$ and $\frac{n}{2}$ is a scale factor for the ranks corresponding to the maximum of the differences $r_{i1}^{(j)} - r_{i2}^{(j)}$, which is of the same order as the variance of $r_i^{(j)}$. Other distances could be specified for $d_q(\cdot, \cdot)$ and $\tilde{d}_p(\cdot, \cdot)$. However, Conne (2005) shows in a simulation study that the distances chosen here lead to a GFI with good performance in terms of empirical power compared to the GFI based on other distances.

The distribution of our GFI depends on $\hat{\alpha}$ through $\hat{z}$ and in order to obtain correct inference, it is natural to integrate out $\hat{\alpha}$ using its asymptotic distribution (see Appendix A). As it will be shown in Section 3.2, it turns out that this corresponds to making a correction based on a distance between two estimated asymptotic covariance matrices of $\hat{\alpha}$, namely $\hat{V}(\hat{\alpha})$ and $\hat{V}(\hat{\alpha})$ defined in Section 3.2. Hence, in (1), we have

$$\hat{\Lambda}^2 = 4 \det \left( \hat{V}(\hat{\alpha}) \right) / \det \left( \hat{V}(\hat{\alpha}) \right),$$

and the resulting GFI is

$$\Omega = 2 \left( \frac{\det \left( \hat{V}(\hat{\alpha}) \right)}{\det \left( \hat{V}(\hat{\alpha}) \right)} \right)^{\frac{1}{2}} \cdot S(x, \hat{z} | \hat{\alpha})$$ (4)

where

$$S(x, \hat{z} | \hat{\alpha}) = \frac{1}{n^2 - n} \sum_{i_1=1}^{n} \sum_{i_2=1}^{n} \sum_{i_1 > i_2} \left[ \frac{1}{q} \sqrt{\sum_{j=1}^{q} \left( \frac{\hat{z}^{(j)}_{i_1} - \hat{z}^{(j)}_{i_2}}{\hat{\sigma}_z^{(j)}} \right)^2} \right]^2$$

$$- \frac{1}{p_1} \left[ \sum_{j=1}^{p_1} \left( \frac{x_{i1}^{(j)} - x_{i2}^{(j)}}{\hat{\sigma}_x^{(j)}} \right)^2 - \frac{1}{p_2} \sum_{j=p_1+1}^{p_1+p_2} \left| \frac{r_{i1}^{(j)} - r_{i2}^{(j)}}{\frac{n}{2}} \right| \right]^2.$$ (5)
3.2 Derivation of the correction factor

Let \( \nu_{S|\alpha}(s|\hat{\alpha}) \) be the conditional density of \( S \) in (5) given \( \hat{\alpha} \), associated \( p \)-values depend on \( \hat{\alpha} \). In order to obtain correct unconditional inference, we consider \( \hat{\alpha} \) as a nuisance parameter, and we integrate it out using its asymptotic normal distribution \( N(\alpha, V(\alpha)) \), where \( V(\alpha) \) is given by (15) in Appendix A. We actually estimate \( V(\alpha) \), for a given \( \alpha \), by means of

\[
\hat{V}(\alpha) = \left( \frac{1}{n} \sum_{i=1}^{n} \left[ \frac{\partial \tilde{l}(\alpha, \phi|x_i)}{\partial \alpha} \cdot \frac{\partial \tilde{l}(\alpha, \phi|x_i)}{\partial \alpha} \right] \right)^{-1}
\]

(6)

where \( \tilde{l} \) is the Laplace approximated likelihood function based on (2) and given by (13) in Appendix A. Then the marginal density \( f_S \) of \( S \) is

\[
f_S(s) = \int \nu_{S|\alpha}(s|\hat{\alpha}) \cdot k_{\tilde{p}} \left( V^{-\frac{1}{2}}(\alpha)(\hat{\alpha} - \alpha) \right) \left| \text{det}(V(\alpha)) \right|^{-\frac{1}{2}} d\hat{\alpha}
\]

\[
= \frac{1}{(2\pi)^{\frac{p}{2}} \left| \text{det}(V(\alpha)) \right|^{\frac{1}{2}}} \cdot \int \nu_{S|\alpha}(s|\hat{\alpha}) \cdot \exp \left( \tilde{p} \cdot \kappa(\hat{\alpha}) \right) d\hat{\alpha},
\]

(7)

where \( \kappa(\hat{\alpha}) = -\frac{1}{2\tilde{p}} \cdot (\hat{\alpha} - \alpha)V^{-1}(\alpha)(\hat{\alpha} - \alpha)^T \), \( k_{\tilde{p}}() \) is the density function of the multivariate standard normal with \( \tilde{p} = \text{dim}(\hat{\alpha}) \).

The term outside the integral depends on an unknown matrix \( V(\alpha) \), which will be estimated by \( \hat{V}(\hat{\alpha}) \) with \( \hat{\alpha} \) defined below. Moreover, the maximum of the function \( \kappa(\hat{\alpha}) \) is achieved at \( \hat{\alpha} = \alpha \) with \( \kappa(\alpha) = 0 \). Applying the \( \tilde{p} \)-dimensional Laplace approximation to the integral in (7), we obtain

\[
\int \nu_{S|\alpha}(s|\hat{\alpha}) \cdot \exp(\tilde{p} \cdot \kappa(\hat{\alpha})) d\hat{\alpha} = \frac{1}{\sqrt{\left| \text{det}(\frac{-1}{\tilde{p}} V^{-1}(\alpha)) \right|}} \cdot \nu_{S|\alpha}(s|\alpha) \cdot \left( \frac{2\pi}{\tilde{p}} \right)^{\frac{\tilde{p}}{2}} \{ 1 + O(\tilde{p}^{-1}) \}
\]
and finally

\[
f_S(s) = \frac{1}{(2\pi)^\frac{p}{2} |\det(V(\hat{\alpha}))|^\frac{1}{2}} \cdot \frac{1}{2 \sqrt{|\det(-\frac{1}{p} V^{-1}(\alpha))|}} \nu_{S|\alpha}(s|\alpha) \left( \frac{2\pi}{\tilde{p}} \right)^\frac{p}{2} \{1 + O(\tilde{p}^{-1}) \}
\]

\[
= \frac{1}{2} \left( \frac{|\det(V(\alpha))|}{|\det(V(\hat{\alpha}))|} \right)^{\frac{1}{2}} \cdot \nu_{S|\alpha}(s|\alpha) \{1 + O(\tilde{p}^{-1}) \}.
\]

As \( \alpha \) is unknown, it will be estimated by its LAMLE \( \hat{\alpha} \). Hence we get the correction factor \( \hat{\Lambda} = 2 \det \hat{V}(\hat{\alpha})^{1/2} / \det \hat{V}(\hat{\alpha})^{1/2} \), leading to \( \Omega \) in (4).

Note that if \( \hat{\alpha} = \hat{\alpha} \), the correction factor becomes simply two. Our empirical experience shows that a correction is crucial for having correct inference, and that \( \hat{\alpha} \) should have a smaller variance than the LAMLE \( \hat{\alpha} \). We propose to use the bagging procedure (Breiman, 1996) on the sample \( y = (x_i, \hat{z}_i)_{i=1,...,n} \), i.e. the observations on the manifest variables and the corresponding estimated scores. \( \bar{\alpha} \) is then obtained as follows. For \( k = 1, \ldots, K \), draw a random sample \( y^*_k = (x^*_i, \hat{z}^*_i)_{i=1,...,n} \) of size \( n \) from \( y \) with replacement and compute from \( y^*_k \) the LAMLE \( \bar{\alpha}^*_k \) of the loadings with \( \hat{z}^*_i \) fixed (not estimated). Then \( \bar{\alpha} = \frac{1}{K} \sum_{k=1}^{K} \bar{\alpha}^*_k \). Note that this is different than resampling from \( y = (x_i, )_{i=1,...,n} \) and estimating both \( \alpha \) and the \( z_i \), but our procedure is much faster and stable. Our simulation study shows that this choice is adequate at least for the models we have investigated.
4 Computing the $p$-value

4.1 Asymptotic distribution of the test statistic under the null hypothesis

Under condition C1 in Appendix B.1, $\Omega$ can be written as

$$\Omega = \omega_1 \cdot S(x, \hat{z} | \hat{\alpha}) = \omega_1 \cdot \frac{1}{n^2 - n} \sum_{1 \leq i_1 < i_2} h(x_{i_1}, x_{i_2}),$$

(8)

where $\omega_1 = 2 \left( \frac{\det(\hat{V}(\tilde{\alpha}))}{\det(\hat{V}(\hat{\alpha}))} \right)^{\frac{1}{2}}$. We consider the case of ordinal manifest variables ($p_1 = 0$). This result can be extended to the case of normal manifest variables.

**Theorem 1** Under conditions C1 and C2 in Appendix B.1, we have

$$n^{\frac{1}{2}}(\Omega - \xi) \xrightarrow{D} N(0, 4\zeta_1)$$

as $n \to \infty$, where $\xi = E[h(x_{i_1}, x_{i_2})] = \text{plim}_{n \to \infty}(\Omega)$ and $\zeta_1 = \text{var} \left( E \left[ h(x_{i_1}, x_{i_2}) | x_{i_1} \right] \right)$.

The proof and explicit expressions for $\xi$ and $\zeta_1$ are provided in a supplementary file (http link). Estimators for $\xi$ and $\zeta_1$ are given in Conne (2008).

We investigate the accuracy of the (estimated) asymptotic distribution in a simulation study. We consider models containing two and three latent variables. 10'000 samples of size 100 (for the model with 2 latent variables) and of size 200 (for the model with 3 latent variables) were simulated. They contain 5 and 10 ordinal manifest variables respectively. In both cases, we observe a large bias under the null hypothesis. More precisely, the estimators of the expectation and the variance in the two latent variables model underestimate 3 times the mean and 5 times the variance.
of $\Omega$ respectively. In the three latent variables model, it is even worse. We found the biases to be persistent, though to a lesser extent, with samples of size up to 1000. Clearly, the problem lies in the estimators of $\xi$ and $\zeta_i$ that are not accurate enough to be used in practice. Indeed, their computation needs numerical approximations which makes inference quite unstable even infeasible. We propose instead in the next section to approximate the distribution of $\Omega$ under the null hypothesis by means of resampling methods.

4.2 Resampling method for computing the $p$-value

To compute $p$-values associated to $\Omega$ in (4), we use a resampling procedure that is close to a parametric bootstrap. Traditionally, a parametric bootstrap procedure implies the generation of samples (of manifest variables realizations) from the model under the null hypothesis (here a GLLVM) estimated by means of the estimated parameters (here the LAMLE) on the observed sample, i.e. $\hat{\alpha}$ and $\hat{\phi}$ (if there are normal manifest variables). On each bootstrapped sample, the model’s parameters are estimated (here the LAMLE of $\alpha$ and $\phi$ as well as the latent scores $z_i$) and the bootstrapped test statistic is computed using these and the bootstrapped sample. We found that a straightforward parametric bootstrap is too computer intensive because of the simultaneous estimation of $\alpha$, $\phi$ and of the latent scores $z_i$ on each bootstrapped sample. Moreover, it does not provide uniformly distributed $p$-values under the null, at least in the examples considered in the simulation study. Hence we simplified the procedure by “skipping” the estimation of $\alpha$ and $\phi$ in the bootstrapped samples. However, in order to take into account at least some of the variability of $\hat{\alpha}$ (we treat $\phi$ as a nuisance parameter), we change the value of $\hat{\alpha}$ when generating the bootstrapped samples according to its asymptotic normal distribution with mean and covariance estimated using the LAMLE on the observed sample. Namely, we generate $\alpha^*_b$ from
that is not only used to generate a bootstrap sample but also as an “estimate” of $\alpha$ from the bootstrapped samples, while for $\phi$ we use all along $\hat{\phi}$.

Note that the latent scores are estimated in the conventional way, i.e. from each bootstrapped sample. This resampling procedure uses a similar idea as in the fast bootstrap of Salibian-Barrera and Zamar (2002). In Section 4.3, we show that this procedure provides consistent $p$-values.

To be more precise, the $p$-values are computed according to the following procedure. Repeat for $b = 1, \ldots, B$:

1. Generate one $\tilde{p}$-vector $\alpha_b^\ast$ from $N(\hat{\alpha}, \frac{1}{n} \hat{V}(\hat{\alpha}))$.

2. Generate $q$ independent standard normal vectors of size $n$ to construct the $n \times q$ matrix $z$.

3. Generate $\forall j = 1, \ldots, p$ the $n$-vectors of conditional means of all responses

   $$\mu^{(j)} = (\gamma^{(j)})^{-1}(\alpha_b^{(j)T} z). \tag{9}$$

   Note that in the ordinal case, one generates $M^{(j)} - 1$ of each with $\alpha_{sb}^{(j)}$ to get the cumulative probabilities of an ordered multinomial distribution.

4. Generate the bootstrapped sample of manifest variable $x_b^\ast$ based upon the mean that were calculated in (9) as well as from the LAMLE of the scale parameters $\hat{\phi}$ for the normal responses.

5. Given $x_b^\ast$, estimate $\hat{z}_b^\ast$ conditionally on $\alpha_b^\ast$ as the maximum of (3).

6. Evaluate $\hat{V}(\alpha_b^\ast)$ with (6), using $x_b^\ast$ and $\hat{z}_b^\ast$.

7. Evaluate $\Omega_b^\ast = 2 \left( \frac{\det(\hat{V}(\alpha_b^\ast))}{\det(V(\hat{\alpha}))} \right)^\frac{1}{2} \cdot S(x_b^\ast, \hat{z}_b^\ast | \alpha_b^\ast)$. 

14
Finally, we estimate the $p$-values by $p-value = \frac{1}{B} \#\{\Omega_0^* > \Omega_{obs}\}$, where $\Omega_{obs}$ is the observed value of $\Omega$ computed on the original sample. One could also use $\hat{V}(\tilde{\alpha})/n$ as a covariance matrix estimate to simulate $\alpha_0^*$ in step 1. However, in our experience, the associated $p$-values are not as close to uniform than the ones associated with the procedure presented above.

It should be stressed that this statistic and the procedure to evaluate its $p$-value is widely applicable. Indeed, if one uses other consistent estimators of $z$, $\alpha$ with a corresponding asymptotic covariance matrix, one could apply the same procedure to define a goodness-of-fit index and its corresponding $p$-value.

### 4.3 Consistency of the resampling procedure

Consistency of the $p$-value computed using the resampling procedure proposed in section 4.2, is provided by the following theorem.

**Theorem 2** Under conditions C1 and C2 in Appendix B.1, given $x_1, \ldots, x_n$,

$$n^{\frac{1}{2}}(\Omega^* - \hat{\Omega}) \overset{D}{\rightarrow} N(0, 4\zeta_1)$$

as $n \rightarrow \infty$, where $\Omega^*$ is the bootstrapped test statistic and $\hat{\Omega}$ is the test statistic computed on the original sample $x_1, \ldots, x_n$.

The proof is given in Appendix B.2.

### 5 Simulation study

In this Section, we study the behavior in terms of level and power of $\Omega$. We consider three different models, one with normal manifest variables, a second one with ordinal manifest variables, and a third one with mixed normal, ordinal and binary manifest
variables. The models with normal and ordinal variables contain each 2 latent variables, while the third one has 3 latent variables. The parameter’s values under the null hypothesis and several alternatives are given in Appendix C.

For the first and second models, we have 2 latent variables and 5 normal or ordinal (5 levels) variables. For the models to be uniquely estimated, one of the loadings (in each model) must be fixed (e.g. to 0), so that a “full” 2 latent variables model with 5 manifest variables would include $2 \cdot 5 - 1 = 9$ loadings, and 5 intercepts in the normal case, and $5 \cdot (5 - 1) = 20$ thresholds in the ordinal case. However, under the null hypothesis, we choose a more constrained model by first setting 2 more loadings to a fixed value (i.e. 0) in both models, and the thresholds to be the same across manifest variables in the ordinal case. This makes a total of $9 - 2 = 7$ loadings and 5 intercepts in the normal case and 4 thresholds in the ordinal case, i.e. respectively 12 and 11 parameters to estimate. As alternative models, only more general models can be considered (i.e. so that the model under $H_0$ is nested in the alternative model). One can consider models with the same number of latent variables but with more non-null loadings, or models with more latent variables. We choose, compared to the null model, one alternative model that has one additional non-null loading (hence 13, respectively 12, non null parameters), a second one that has two additional non-null loadings (hence 14, respectively 13, non-null parameters). A third alternative model has a third latent variable with 5 different loadings with respect to the null model (see Appendix C). In terms of model size (number of different parameters), the alternative models differ from the null model by 1,2 and 5 parameters respectively.

For the third model, we have 3 latent variables and 10 manifest variables, of which 3 are ordinal (5 levels), 2 are Bernoulli and 5 are normal. For the model to be uniquely estimated, three of the loadings must be fixed (e.g. to 0), one on one latent variable and 2 on another latent variable with one corresponding to the same manifest
variable, so that a “full” model with 3 latent variables and 10 manifest variables would include $3 \cdot 10 - 3 = 27$ loadings and $3 \cdot 4 + 2 = 14$ thresholds and 5 intercepts, i.e. 46 parameters. However we choose a more constrained model by setting 2 more loading to a fixed value (i.e. 0). This makes a total of $46 - 2 = 44$ parameters to estimate. As alternative models we choose, compared to the null model, one alternative model with one additional non-null loading (hence 45 non-null parameters), a second one with two additional non-null loadings (hence 46 non-null parameters) and a third one with a forth latent variable with 3 non-null loading, which compared to the null model has 5 additional non-null parameters.

To compute the empirical levels, for the first an second model, we simulate 10,000 samples of 100 observations and for the third model 5,000 samples of 200 observations. To compute the $p$-value of each sample, we use 100 bootstrapped samples in all cases and for $\tilde{\alpha}$, we use 30 random samples in the bagging procedure. To compute the empirical powers, we make use of the extrapolation method given in Boos and Zhang (2000). Namely, we generate 200 samples of 100 observations under the alternatives for the first and second model and 200 samples of 200 observations under the alternatives for the third model. We then compute the empirical levels using 19, 39, and 59 bootstrapped samples for the calculation of the $p$-values. The power under one alternative of one model is then extrapolated from equation (5) in Boos and Zhang (2000) using the empirical levels found with respectively 19, 39, and 59 bootstrapped samples.

For a given model with $p$ manifest variables, $q$ latent variables and parameters $\alpha$ and $\phi$, the samples of size $n$ are generated in S-Plus using the steps 2 to 4 in the procedure given in Section 4.2, in which $\alpha^*_b^{(j)}$, $x^*_b$ and $\hat{\phi}$ are replaced by respectively $\alpha^{(j)}$, $x$ and $\phi$.

As argued in Section 1, other GFI such as S&B GFI are designed for generalized
latent variable models with the underlying normal variable hypothesis. To compute the latter, one can use a standard software such as *Mplus* (Muthen and Muthen 2001), but the data must be simulated under the underlying normal hypothesis. If the data are generated according to the procedure above, for many samples the test statistic cannot be computed (up to 40% at the null model and more at alternative models). S&B GFI can however be compared to our GFI in the normal model.

Table 1 presents the empirical levels for \( \Omega \) and the S&B GFI for the three GLLVM null models. At the normal manifest variables model (null model 1), the empirical levels of both statistics are comparable and both match the nominal levels, S&B GFI being slightly liberal at the 5% nominal level. However, since the \( p \)-values for S&B GFI are based on the asymptotic distribution, the empirical level of the S&B GFI could in principle be improved by using for instance a parametric bootstrap. At the ordinal and mixed ordinal, binary and normal manifest variables (null models 2 and 3), the empirical levels for \( \Omega \) still match the nominal ones, being perhaps slightly liberal at the 1% nominal level. This simulation shows that the resampling procedure we propose provides consistent \( p \)-values for \( \Omega \).

Table 2 presents the empirical power at the 5% level for \( \Omega \) and the S&B GFI for the same three GLLVM null models. Not surprisingly, at the normal manifest variables model (null model 1), the power of the S&B GFI is larger than the power of \( \Omega \). Indeed, the S&B GFI makes fully use of the normality assumption. The power of \( \Omega \) is reasonable, especially given that the models under the alternative are not very different from the model under the null and that the number of parameters to estimate is relatively large. Indeed, one can see that in smaller models (e.g. null model 2) the powers are overall larger than in the larger model (null model 3), and that the power increases rapidly with the size difference between the null and alternative models.
Table 1: Empirical levels for Ω and the S&B GFI for different GLLVM null models. MCV stands for the Monte-Carlo variation (±2·σ) of the estimated p-values.

<table>
<thead>
<tr>
<th>Alternatives to</th>
<th>Null model 1</th>
<th>Null model 2</th>
<th>Null model 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size difference</td>
<td>1 2 5</td>
<td>1 2 5</td>
<td>1 2 5</td>
</tr>
<tr>
<td>Ω</td>
<td>0.47 0.59 0.73</td>
<td>0.10 0.43 0.77</td>
<td>0.11 0.21 0.42</td>
</tr>
<tr>
<td>S&amp;B GFI</td>
<td>0.82 0.98 1.00</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Empirical power at the 5% level for Ω and the S&B GFI for different GLLVM null models.

6 Outlook

The GFI and the corresponding resampling procedure proposed here is to the best of our knowledge the only GFI for GLLVM models that is not based on the underlying normal variable hypothesis. Moreover, it can in principle be used with other estimators than the LAMLE. What is needed is their (asymptotic) distribution and the resampling procedure can be adapted accordingly. This procedure can also be naturally extended to GLLVM with covariates. In that case, the estimated latent scores \( \hat{z} \) in (5) are replaced by \( \hat{z} - \hat{\beta}^T \xi \), where \( \xi \) is the vector of covariates and \( \hat{\beta} \) the vector of estimated parameters. In the resampling procedure, both \( \hat{\alpha} \) and \( \hat{\beta} \) are generated from their estimated asymptotic distribution. This extension however needs further investigation. The GFI and the procedure to evaluate its p-value can also in principle be extended to multilevel models, structural equations models or combined measurement models, using the generalized factor formulation of Skrondal and Rabe-Hesketh (2004), but more work needs to be done to extend the estimation procedure and their resampling techniques to these cases.
References


Vaida, F. and S. Blanchard (2005). Conditional Akaike information for mixed-
effects models. *Biometrika* 92(2), 351–370.

David Conne is Quantitative Analyst at Life Store SA, Geneva, Switzerland. Elvezio Ronchetti is Professor in the Department of Econometrics and Maria-Pia Victoria-Feser is Professor at HEC, University of Geneva, Blvd. Pont d’Arve 40, CH-1211 Geneva (Switzerland). The authors thank the Editor, the Associate Editor, two referees, E. Cantoni, and S. Rabe-Hesketh for very constructive comments which greatly improved the original manuscript. The research was partially carried out while the first author was a Ph. D. student at the University of Geneva. The third author was partially supported by the Swiss National Science Foundation (grant no PP001-106465).
Appendix

A Definition and properties of LAMLE

For completeness, we give here the definition and the properties of the LAMLE. For more details, see Huber, Ronchetti, and Victoria-Feser (2004). By rewriting the density of $x_i$ in (2) as

$$f_{\alpha, \phi}(x_i) = \int e^{pQ(\alpha, z, x_i)} d\mathbf{z}(2),$$

(10)

where

$$Q(\alpha, z, x_i) = \frac{1}{p} \left[ \sum_{j=1}^{p} \left[ \frac{x_i^{(j)} u_j(\alpha^{(j)T} z) - b_j(u_j(\alpha^{(j)T} z))}{\phi_j} + c_j(x_i^{(j)}, \phi_j) \right] ight] - \frac{z_{(2)}^T z_{(2)}}{2} - \frac{q}{2} \log(2\pi),$$

(11)

and by applying the Laplace approximation to (10), we obtain

$$f_{\alpha, \phi}(x_i) = \left( \frac{2\pi}{p} \right)^{\frac{q}{2}} \det(-W(\hat{z}_{i(2)}))^{-\frac{1}{2}} e^{pQ(\alpha, \hat{z}_i, x_i)} (1 + O(p^{-1})),$$

(12)

where

$$W(z) = \frac{\partial^2 Q(\alpha, z, x_i)}{\partial z_{(2)}^T z_{(2)}} = -\frac{1}{p} \Gamma(\alpha, \phi, z),$$

$$\Gamma(\alpha, \phi, z) = \sum_{j=1}^{p} \frac{1}{\phi_j} \frac{\partial^2 b_j(u_j(\alpha^{(j)T} z))}{\partial z^T z} + I_q.$$  

(12) depends on the unknown quantity $\hat{z}_i = (1, \hat{z}_{i(2)})$, the maximum of the function $Q$. Lee and Nelder (1996) show that as $n \to \infty$, $n^{\frac{1}{2}}(\hat{z}_{i(2)} - z_{i(2)}) \xrightarrow{D} N(0, \Gamma(\alpha, \phi, z_i)^{-1}).$
From (11) and (12), we obtain the Laplace approximated log-likelihood function

\[
\tilde{l}(\alpha, \phi, |x) = \sum_{i=1}^{n} \left( \frac{-1}{2} \log \det \left\{ \Gamma(\alpha, \phi, \hat{z}_i) \right\} - \frac{\hat{z}_i^{T}(2) \hat{z}_i^{(2)}}{2} + \sum_{j=1}^{p} \left\{ x_i^j u_j(\alpha^{(j)T} \hat{z}_i) - b_j(\alpha^{(j)T} \hat{z}_i) + c_j(x_i^j, \phi_j) \right\} \right). 
\] (13)

The resulting loadings estimator of \( \alpha \) called the LAMLE is the solution of

\[
\frac{\partial \tilde{l}(\alpha, \phi| x)}{\partial \alpha_i^{(j)}} = \sum_{i=1}^{n} \psi_{ji}(x_i; \alpha, \hat{z}_i) = 0, \quad j = 1, ..., p, \quad l = 0, ..., q, \tag{14}
\]

where

\[
\psi_{ji}(x_i; \alpha, \hat{z}_i) = -\frac{1}{2} \operatorname{tr} \left( \Gamma(\alpha, \phi, \hat{z}_i)^{-1} \frac{\partial \Gamma(\alpha, \phi, \hat{z}_i)}{\partial \alpha_i^{(j)}} \right) + \frac{1}{\phi_j} \left( x_i^j - \frac{\partial b_j(\alpha^{(j)T} \hat{z}_i))}{\partial \alpha_i^{(j)T} \hat{z}_i} \right) \hat{z}_i \hat{z}_i^T .
\]

Equation (14) may have multiple solutions. If \( q > 1 \), it is necessary to impose \( \frac{q(q-1)}{2} \) constraints on the parameters \( \alpha \) to obtain a unique solution.

The LAMLE \( \hat{\alpha} \) belongs to the class of M-estimators (Huber, 1981), and under the conditions given in Huber (1981), \( \hat{\alpha} \) is consistent and asymptotically normal, that is as \( n \to \infty \), \( n^{\frac{1}{2}} (\hat{\alpha} - \alpha) \overset{D}{\to} N(0, V(\alpha)) \), with

\[
V(\alpha) = B(\alpha)^{-1} A(\alpha) B(\alpha)^{-T},
\] (15)

\[
A(\alpha) = E \left[ \Psi(x; \alpha, \hat{z}) \Psi^T(x; \alpha, \hat{z}) \right], \quad B(\alpha) = -E \left[ \frac{\partial \Psi(x; \alpha, \hat{z})}{\partial \alpha} \right], \quad \Psi = (\psi_{ji})_{i=1,...,p}
\]

and the expectations are taken under the GLLVM model. The same argument holds for \( \hat{\phi} \), the LAMLE of \( \phi \). For more details, in particular for the specific estimation
equations in the normal and ordinal cases, see Huber, Ronchetti, and Victoria-Feser (2004), Huber (2004) and Huber, Scaillet, and Victoria-Feser (2009).

B Conditions and proofs

B.1 Conditions for Theorems 1 and 2

C1: For $i = 1, \ldots, n$, $r_i = (r_i^{(1)}, \ldots, r_i^{(p)})$ is a function of $x_i$ only.

C2: $E[h^2(x_{i1}, x_{i2})] < \infty$

Condition C1 is satisfied e.g. when $n$ is sufficiently large. It ensures that $S(x, z|\hat{\alpha})$ can be viewed as a U-statistic of order 2. Condition C2 is a standard condition to ensure the asymptotic normality of U-statistics.

B.2 Proof of Theorem 2

Let $x \sim F_\theta$, where $\theta = (\alpha, \phi)$. Then, according to subsection 4.2, bootstrapped data are distributed according to $x^* \sim F_{\hat{\theta}^*}$, where $\hat{\theta} = (\hat{\alpha}, \hat{\phi})$ and $\hat{\alpha} \sim N(\hat{\alpha}, \hat{V}(\hat{\alpha})/n)$. Now we can apply Theorem 3.1 in Bickel and Freedman (1981). Their condition (3.15) is our condition C2 and their condition (3.16) is satisfied because in our case $h(x_{i1}, x_{i2}) = 0$. This provides the consistency of the bootstrap procedure when resampling under the empirical distribution function. Finally, we need to check condition (2.4) in Bickel and Freedman (1981) which guarantees that the same result holds when resampling is performed under $F_{\hat{\theta}^*}$. This condition is satisfied in our case because

(a) the mapping $\theta \rightarrow \int v(x)dF_\theta(x)$ is continuous whenever

$$\int |v(x)|dF_\theta(x) < \infty;$$
(b) $\hat{\alpha}$ and $\hat{\phi}$ are consistent for $\alpha$ and $\phi$ respectively (see Appendix A) and conditional on $\hat{\alpha}$, $n\hat{x}(\alpha - \hat{\alpha})$ is asymptotically $N(0, \hat{V}(\alpha))$.

C Simulation Parameters

C.1 Normal model

The normal model contains 5 normal manifest variables and 2 latent variables. The intercepts are

$$\alpha_0 = \begin{bmatrix} 1 & 1.5 & 2 & 3 & 0.5 \end{bmatrix}^T$$

Under the null hypothesis, the loadings matrix $\lambda$ is

$$\lambda_0 = \begin{bmatrix}
1.25 & 0 & -1.75 & -3.25 \\
0 & 1.9 & 3.75 & 0.95 \\
0 & -0.5 &
\end{bmatrix}$$  \hspace{1cm} (16)
Note that the loadings set to 0 are not estimated. Under the three different alternatives, the loading matrices are respectively

\[
\begin{align*}
\lambda_{A1} &= \begin{bmatrix}
1.25 & 0 \\
-1.75 & -3.25 \\
2.25 & 1.9 \\
3.75 & 0.95 \\
0 & -0.5 
\end{bmatrix} \\
\lambda_{A2} &= \begin{bmatrix}
1.25 & 0 \\
-1.75 & -3.25 \\
2.25 & 1.9 \\
3.75 & 0.95 \\
-2.25 & -0.5 
\end{bmatrix} \\
\lambda_{A3} &= \begin{bmatrix}
1.25 & 0 & 0 \\
-1.75 & -3.25 & 0 \\
0 & 1.9 & 2.75 \\
3.75 & 0 & 1.5 \\
-2.25 & -0.5 & -2.75 
\end{bmatrix}
\end{align*}
\]

(17)

C.2 Ordinal model

The ordinal model contains 5 ordinal manifest variables with 5 levels and 2 latent variables. The thresholds are the same across manifest variables and are

\[
\alpha_0^{(j)} = [\alpha_{0s}^{(j)}]_{s=1,\ldots,4}^T = \begin{bmatrix}
-0.2 & 0.4 & 1.4 & 2
\end{bmatrix}^T, \forall j = 1, \ldots, p
\]

Under the null hypothesis, the loadings matrix \(\lambda\) is given by (16) and under the three different alternatives, the loading matrices are given respectively by (17)

C.3 Mixed ordinal, Bernoulli and normal model

This model has 3 ordinal (5 levels), 2 Bernoulli and 5 normal manifest variables and 3 latent variables. The intercepts of the normal manifest variables, are \(\alpha_0^{(j)} = \begin{bmatrix}
5 & -2 & 3 & 0 & -8
\end{bmatrix}^T, \forall j = 1, \ldots, 5\) and the scale parameters are set to \(\phi_j = 1, \forall j = \ldots, 5\)
1, \ldots, 5. The thresholds for the ordinal variables are

\[
\mathbf{\alpha}_0 = [\mathbf{\alpha}_0^{(j)}]_{j=1,\ldots,3,s=1,\ldots,4} = \begin{bmatrix}
-2 & -1 & 0 & 1 \\
-1 & -0.3 & 0.4 & 0.5 \\
0.4 & 0.6 & 0.9 & 1.5
\end{bmatrix}
\]

and the ones of the binary variables are \( \mathbf{\alpha}_0 = [\mathbf{\alpha}_0^{(j)}]_{j=1,2} = [ -1 \ 0.9 ]^T \). The 10 \times 3 loadings matrix under the null hypothesis is

\[
\mathbf{\lambda}_0 = \begin{bmatrix}
4 & 0 & 0 \\
0 & 6 & 0 \\
-6 & 4 & 2.75 \\
1 & 8 & 1.5 \\
-3 & -2 & -2.75 \\
-1 & 1.4 & -3.25 \\
0.5 & 0.7 & 0.9 \\
2 & -1.4 & 2.4 \\
-0.3 & -0.5 & -0.75 \\
0 & -0.2 & 1.2
\end{bmatrix}
\]
Note that the loadings set to 0 are not estimated. Under the three different alternatives, the loading matrices are respectively

\[
\lambda_{A1} = \begin{bmatrix}
4 & 0 & 0 & 0 \\
0 & 6 & 0 & 0 \\
-6 & 4 & 2.75 & 0 \\
1 & 8 & 1.5 & 0 \\
-3 & -2 & -2.75 & 0 \\
-1 & 1.4 & -3.25 & 0 \\
0.5 & 0.7 & 0.9 & 0 \\
2 & -1.4 & 2.4 & 0 \\
-0.3 & -0.5 & -0.75 & 0 \\
-1.8 & -0.2 & 1.2 & 0
\end{bmatrix}
\]

\[
\lambda_{A2} = \begin{bmatrix}
4 & 0 & 0 & 0 \\
2 & 6 & 0 & 0 \\
-6 & 4 & 2.75 & 0 \\
1 & 8 & 1.5 & 0 \\
-3 & -2 & -2.75 & 0 \\
-1 & 1.4 & -3.25 & 0 \\
0.5 & 0.7 & 0.9 & 0 \\
2 & -1.4 & 2.4 & 0 \\
-0.3 & -0.5 & -0.75 & 0 \\
-1.8 & -0.2 & 1.2 & 0
\end{bmatrix}
\]

\[
\lambda_{A3} = \begin{bmatrix}
4 & 0 & 0 & 0 \\
2 & 6 & 0 & 0 \\
-6 & 4 & 2.75 & 0 \\
1 & 8 & 1.5 & 0 \\
-3 & -2 & -2.75 & 0 \\
-1 & 1.4 & -3.25 & 0 \\
0.5 & 0.7 & 0.9 & 0 \\
2 & -1.4 & 2.4 & 2.75 \\
-0.3 & -0.5 & -0.75 & 1.5 \\
-1.8 & -0.2 & 1.2 & -2.75
\end{bmatrix}
\]
Proof of Theorem 1

Since $\hat{V}(\hat{\alpha})$ and $\tilde{V}(\hat{\alpha})$ are two different consistent estimators of the asymptotic variance of the loadings $\alpha$, when $n$ tends to infinity, these matrices converge to the same matrix $V(\alpha)$ and $\omega_1 \to 2$.

Under condition C1, $S(x, \hat{z} | \hat{\alpha})$ can be viewed as a U-statistic of order 2. Indeed, by construction, $h(x_{i_1}, x_{i_2})$ is a symmetric function. Moreover, $\hat{z}_i$ is function of $x_i$, $\alpha$, and $\phi$. Since $\phi$ is just a scale parameter and $S$ is conditional on $\hat{\alpha}$, $h$ is defined as a function of $x_{i_1}$ and $x_{i_2}$ only.

Then, the asymptotic distribution of $S$ is given by Serfling (1980) and requires the calculation of the first and second moments. The calculation and estimation of the second moment is straightforward, since the second moment is conditional on $x_{i_1}$. For the first moment, we have

$$\xi = E[h(x_{i_1}, x_{i_2})] = E[d_q(\hat{z}_{i_1}, \hat{z}_{i_2})^2] + E[\tilde{d}_{p_2}(x_{i_1}, x_{i_2})^2]$$

$$- 2E[d_q(\hat{z}_{i_1}, \hat{z}_{i_2}) \cdot \tilde{d}_{p_2}(x_{i_1}, x_{i_2})].$$

(18)

To compute $\xi$ explicitly, we need the two additional conditions

C3: $r^{(j)}_{i_1}$ is independent of $r^{(j)}_{i_2}$ for all $i_1 \neq i_2$ and for all $j = 1, ..., p_2$.

C4: $corr[d_q(z_{i_1}, \hat{z}_{i_2}); \tilde{d}_p(x_{i_1}, x_{i_2})] = 1$, when the model is correctly specified.

Condition C3 is satisfied e.g. when $n$ is sufficiently large. Condition C4 can be justified with the idea which is under the specification of our GFI: "to a large distance (comparatively to the others) between two observations should correspond a large
distance on the latent scores scales if the model is well specified". However, it is much more difficult to justify that the relationship between those distances is perfectly linear, especially in the discrete case. This strong assumption has to be made because no feasible alternatives seem to be available.

Under condition C4, we have that

$$2E \left[ d_q(\hat{z}_{i1}, \hat{z}_{i2}) \cdot \tilde{d}_p^2(x_{i1}, x_{i2}) \right] = 2 \left\{ \sqrt{\text{var} \left( d_q(\hat{z}_{i1}, \hat{z}_{i2}) \right)} \sqrt{\text{var} \left( \tilde{d}_p^2(x_{i1}, x_{i2}) \right)} ight. \\
\left. + E \left[ d_q(\hat{z}_{i1}, \hat{z}_{i2}) \right] E \left[ \tilde{d}_p^2(x_{i1}, x_{i2}) \right] \right\} = 2 \left\{ \sqrt{E \left[ d_q^2(x_{i1}, x_{i2}) \right] - E \left[ d_q^2(\hat{z}_{i1}, \hat{z}_{i2}) \right]^2} \right. \\
\left. \cdot \sqrt{E \left[ \tilde{d}_p^2(x_{i1}, x_{i2}) \right] - E \left[ \tilde{d}_p^2(x_{i1}, x_{i2}) \right]^2} \\+ E \left[ d_q(\hat{z}_{i1}, \hat{z}_{i2}) \right] E \left[ \tilde{d}_p^2(x_{i1}, x_{i2}) \right] \right\}. \quad (19)$$

We have that

$$E \left[ d_q^2(\hat{z}_{i1}, \hat{z}_{i2}) \right] \overset{n \to \infty}{\longrightarrow} \frac{1}{q^2} \sum_{j=1}^{q} \left[ 2 + \left( \frac{z_{i1}^{(j)} - z_{i2}^{(j)}}{\sigma_z^{(j)}} \right)^2 \right]$$

To calculate the other terms, we use the distribution of $r^{(j)}$ as a function of the empirical distribution $F^{n(j)}$ of $x^{(j)}$. Suppose that each ordinal manifest variable has $M^{(j)}$ category, then

$$k(m^{(j)}) = \frac{n}{2} \left[ F^{n(j)}(m^{(j)}) + F^{n(j)}(m^{(j)} - 1) \right] + \frac{1}{2}$$

and

$$P \left( r^{(j)} = k(m^{(j)}) \right) = F^{n(j)}(m^{(j)}) - F^{n(j)}(m^{(j)} - 1),$$
where \( m^{(j)} = 1, \ldots, M^{(j)} \). We have

\[
E \left[ \tilde{d}_{p_2}(x_{i_1}, x_{i_2})^2 \right] = \frac{1}{p_2^2} \sum_{j=1}^{p_2} E \left[ \frac{(r_{i_1}^{(j)} - r_{i_2}^{(j)})^2}{n^2} \right] + \frac{8}{p_2^2 n^2} \sum_{j_1=1}^{p_2} \sum_{j_2=1}^{p_2} E \left[ \left| r_{i_1}^{(j_1)} - r_{i_1}^{(j_2)} \right| \cdot \left| r_{i_1}^{(j_2)} - r_{i_2}^{(j_2)} \right| \right].
\] (20)

and under condition C3, we get

\[
E \left[ \tilde{d}_{p_2}(x_{i_1}, x_{i_2})^2 \right] = \frac{4}{p_2^2 n^2} \sum_{j_1=1}^{p_2} \sum_{j_2=1}^{p_2} \sum_{m_1=1}^{M^{(j)}} \sum_{m_2=1}^{M^{(j)}} \left[ \frac{n}{2} \left( F^{(j)}(m_1 - 1) - F^{(j)}(m_2 - 1) \right) \right]^2 \\
- \left( F^{(j)}(m_1) - F^{(j)}(m_2) \right) \left( F^{(j)}(m_1 - 1) - F^{(j)}(m_2 - 1) \right) \\
+ \frac{8}{p_2^2 n^2} \sum_{j_1=1}^{p_2} \sum_{j_2=1}^{p_2} \sum_{m_1=1}^{M^{(j)}} \sum_{m_2=1}^{M^{(j)}} \left[ \text{cov} \left[ r_{i_1}^{(j_1)} - r_{i_2}^{(j_1)}, r_{i_1}^{(j_2)} - r_{i_2}^{(j_2)} \right] \right] \\
+ E \left[ \left| r_{i_1}^{(j_1)} - r_{i_2}^{(j_1)} \right| \right] \cdot E \left[ \left| r_{i_1}^{(j_2)} - r_{i_2}^{(j_2)} \right| \right],
\] (21)

where

\[
\text{cov} \left[ r_{i_1}^{(j_1)} - r_{i_2}^{(j_1)}, r_{i_1}^{(j_2)} - r_{i_2}^{(j_2)} \right] = 2 \text{cov}[r_{i_1}^{(j_1)}, r_{i_2}^{(j_2)}] \\
\cdot \left[ 2P[r_{i_1}^{(j_1)} \geq r_{i_2}^{(j_1)}, r_{i_1}^{(j_2)} \geq r_{i_2}^{(j_2)}] \\
+ 2P[r_{i_1}^{(j_1)} < r_{i_2}^{(j_1)}, r_{i_1}^{(j_2)} < r_{i_2}^{(j_2)}] - 1 \right].
\] (22)
and for \( j = j_1, j_2 \)

\[
E \left[ \left| r_{i_1}^{(j)} - r_{i_2}^{(j)} \right| \right] = \sum_{m_1=1}^{M^{(j)}} \sum_{m_2=1}^{M^{(j)}} \sum_{m_1 \neq m_2}^{m_1} \left\{ \frac{n}{2} \left( F^{n(j)}(m_1 - 1) - F^{n(j)}(m_2 - 1) \right) 
+ \left( F^{n(j)}(m_1) - F^{n(j)}(m_2) \right) \right\} \cdot \left( F^{n(j)}(m_1 - 1) - F^{n(j)}(m_1 - 1) \right) 
\cdot \left( F^{n(j)}(m_2) - F^{n(j)}(m_2 - 1) \right),
\]

so that

\[
E \left[ \tilde{d}_{p_2}(\mathbf{x}_{i_1}, \mathbf{x}_{i_2})^2 \right] = \frac{4}{p_2^2 n^2} \sum_{j=1}^{p_2} \sum_{m_1=1}^{M^{(j)}} \sum_{m_2=1}^{M^{(j)}} \sum_{m_1 \neq m_2}^{M^{(j)}} \left\{ \frac{n}{2} \left( F^{n(j)}(m_1 - 1) - F^{n(j)}(m_2 - 1) \right) 
+ \left( F^{n(j)}(m_1) - F^{n(j)}(m_2) \right) \right\}^2 \cdot \left( F^{n(j)}(m_1 - 1) - F^{n(j)}(m_1 - 1) \right) 
\cdot \left( F^{n(j)}(m_2) - F^{n(j)}(m_2 - 1) \right) + \frac{8}{p_2^2 n^2} \sum_{j_1=1}^{p_2} \sum_{j_2=1}^{p_2} \sum_{j_1 \neq j_2} \sum_{j_i > j_2} \left[ \sum_{m_1=1}^{M^{(j_1)}} \sum_{m_2=1}^{M^{(j_1)}} \sum_{m_1 \neq m_2}^{M^{(j_1)}} \frac{n}{2} \left( F^{n(j_1)}(m_1 - 1) - F^{n(j_1)}(m_2 - 1) \right) 
+ \left( F^{n(j_1)}(m_1) - F^{n(j_1)}(m_2) \right) \right) \cdot \left( F^{n(j_1)}(m_1 - 1) - F^{n(j_1)}(m_1 - 1) \right) 
\cdot \left( F^{n(j_1)}(m_2) - F^{n(j_1)}(m_2 - 1) \right) \cdot \sum_{m_1=1}^{M^{(j_2)}} \sum_{m_2=1}^{M^{(j_2)}} \sum_{m_1 \neq m_2}^{M^{(j_2)}} \frac{n}{2} \left( F^{n(j_2)}(m_1 - 1) - F^{n(j_2)}(m_2 - 1) \right) 
+ \left( F^{n(j_2)}(m_1) - F^{n(j_2)}(m_2) \right) \right\} \cdot \left( F^{n(j_2)}(m_1 - 1) - F^{n(j_2)}(m_1 - 1) \right) 
\cdot \left( F^{n(j_2)}(m_2) - F^{n(j_2)}(m_2 - 1) \right) \right\} \cdot \left( F^{n(j_2)}(m_2) - F^{n(j_2)}(m_2 - 1) \right). \]
Finally, from (23), we have

\[
E \left[ \bar{d}_{p2}(x_{i1}, x_{i2}) \right] = E \left[ \frac{2}{np} \sum_{j=1}^{p} \left| r_{i1}^{(j)} - r_{i2}^{(j)} \right| \right] = \\
\frac{2}{np} \sum_{j=1}^{p} \sum_{m_1=1}^{M^{(j)}} \sum_{m_2=1}^{M^{(j)}} \sum_{m_1 \neq m_2} \frac{n}{2} \left( F^{n(j)}(m_1 - 1) - F^{n(j)}(m_2 - 1) \right) \\
+ F^{n(j)}(m_1) - F^{n(j)}(m_2) \right| \cdot (F^{n(j)}(m_1) - F^{n(j)}(m_1 - 1)) \\
\cdot (F^{n(j)}(m_2) - F^{n(j)}(m_2 - 1)) \quad (25)
\]

Note that the quantity \( E [d_q(\hat{z}_{i1}, \hat{z}_{i2})] \) needs to be estimated; see Conne (2008).

References


*Ph. D. Thesis, Department of Econometrics, University of Geneva.*