A simulation study to compare competing estimators in structural equation models with ordinal variables

Elefant-Yanni, Véronique Rica; Huber, Philippe; Victoria-Feser, Maria-Pia

2004.11

FACULTE DES SCIENCES ECONOMIQUES ET SOCIALES

HAUTES ETUDES COMMERCIALES

A simulation study to compare competing estimators in structural equation models with ordinal variables

Veronique ELEFANT-YANNI
Philippe HUBER
Maria-Pia VICTORIA-FESER

HEC GENÈVE

UNIVERSITÉ DE GENÈVE

A simulation study to compare competing estimators in structural equation models with ordinal variables

Veronique Elefant-Yanni, Philippe Huber and Maria-Pia Victoria-Feser*

University of Geneva, Switzerland

July 2004

Abstract

Structural equation models have been around for now a long time. They are intensively used to analyze data from different fields such as psychology, social sciences, economics, management, etc. Their estimation can be performed using standard statistical packages such as LISREL. However, these implementations suffer from an important drawback: they are not suited for cases in which the variables are far from the normal distribution. This happens in particular with ordinal data that have a non symmetric distribution, a situation often encountered in practice. An alternative approach would be to use generalized linear latent variable models (GLLVM) as defined for example in Bartholomew and Knott 1999

*Corresponding author (address: HEC, University of Geneva, 40 bd du Pont d'Arve, 1211 Geneva, Switzerland. Tel: +41 22 379 8807. Email: maria-pia.victoriafeser@hec.unige.ch). All authors aknowledge the support of the Swiss National Science Fundation (grant no 610-057883.99)

and Moustaki and Knott (2000). These models consider the data as they are, i.e. binary or ordinal but the loglikelihood function is intractable and needs numerical approximations to compute it. Several approaches exist such as Gauss-Hermite quadratures or simulation based methods, as well as the Laplace approximation, i.e. the Laplace approximated maximum likelihood estimator (LAMLE) proposed by Huber, Ronchetti, and Victoria-Feser (2004) for these models. The advantage of the later is that it is very fast and hence can cope with relatively complicated models. In this paper, we perform a simulation study to compare the parameters' estimators provided by LISREL which is taken as a benchmark, and the LAMLE when the data are generated from a confirmatory factor analysis model with normal variables which are then transformed into ordinal ones. We will show that while the LISREL estimators can provide seriously biased estimators, the LAMLE not only is unbiased, but one can also recover an unbiased estimator of the correlation matrix of the original normal variables.

Keywords: Confirmatory factor analysis, Laplace approximation, covariance structure, LISREL, Generalized Linear Latent Variable Models, LAMLE.

1 Introduction

In many scientific fields researchers use models based on theoretical concepts that cannot be observed directly. These theoretical concepts include for example the standard of living or welfare in economics, intelligence or anxiety, etc., in psychology, marketing orientation in management, etc. These concepts are very important within the framework of theoretical models, but when these models are validated by means of observed data, the problem of measurement arises. In these situations, observable quantities (manifest variables) that are proxies for the concepts of interest are used to build up the theoretical concepts (latent variables). For this kind of problems statistical methods have long been available. Principal component analysis, factor analysis and structural equation modelling (see e.g. Jöreskog 1969 and Arminger and Küsters 1988) are suitable methods. For the later, analysis can be used that are available in now standard software such as LISREL (Jöreskog and Sörbom 1993).

Although LISREL incorporates methods dealing with a wide variety of applied problems, it is based on the assumption that the manifest variables are multivariate normal. When this is obviously not the case (as in the case of binary or ordinal variables), the manifest variables are taken as indirect observations of multivariate normal variables and standard inference based on the maximum likelihood estimator (MLE) is then performed (see Jöreskog 1990). This approach, as implemented in LISREL and other packages, suffers from an important drawback: the resulting estimators can be seriously biased when the data generating distribution is far from the normal one, especially when it is non symmetric. Several simulation studies have shown this feature; see e.g. DiStefano (2002).

In our opinion, it is essential that the manifest variables are treated as they are, i.e. binary, ordinal or continuous, and that the model that formalizes the relationship between the manifest and the latent variables should take the type of data into account. Such models were first investigated by Bartholomew (1984a, 1984b) who considered the case of binary data. More recently, Moustaki (1996) and Moustaki and Knott (2000) considered mixtures of manifest variables. They proposed a generalized linear latent variable model (GLLVM) that allows one to link latent variables to manifest variables of different type (see also Bartholomew and Knott 1999).

The statistical analysis of GLLVM presents a difficulty: since the latent variables are not observed, they must be integrated out from the likelihood function. One could consider several approaches to solve this problem. Moustaki (1996) proposes using a simple Gauss-Hermite quadrature as a numerical approximation method. However, it is known that a simple Gauss-Hermite quadrature can lead to a poor approximation. Moreover, it is often infeasible when the number of latent variables is large. A possible improvement is provided by an adaptive Gauss-Hermite quadrature which appropriately centers and rescales the quadrature nodes. This technique is implemented in the function `gllamm` in `Stata` (see Rabe-Hesketh, Skrondal, and Pickles 2002) to fit generalized latent and mixed models (Skrondal and Rabe-Hesketh 2004) and can be used to fit our models. However, the procedure is at the moment extremely slow. We propose instead using the Laplace approximation of the likelihood function, resulting in the so-called Laplace approximated MLE, i.e. LAMLE (see Huber, Ronchetti, and Victoria-Feser 2004) which is implemented in a software called LCube. In the case of generalized linear mixed models (GLLAMM), which can be seen as a generalization of GLLVM, a simplified version of the Laplace approximation is used by Breslow and Clayton (1993) and Lin and Breslow (1996) which results in the same estimator as that proposed by McGilchrist (1994) and Lee and Nelder (1996) (see also Huber, Ronchetti, and Victoria-Feser 2004). Laplace approximation of the likelihood has the important advantage with respect to quadrature that it allows one to estimate more complex models in an efficient and fast way. Alternative estimation methods include methods based on stochastic approximations such as MCMC and MCEM; see e.g. Yau and McGilchrist (1996). While these methods have been applied successfully in many complex situations, there are potential drawbacks such as long computation times and stopping rules.

In this paper, we compare the performance in terms of bias and variance of the LAMLE versus the estimator provided by LISREL which is taken as a benchmark. We do that through a simulation study, using a typical and quite important factor analysis model for ordinal data. In particular, we will show that while the LISREL estimates can be seriously biased when the ordinal data are non symmetric, the LAMLE is never biased. We will start by briefly describing the proposed estimators for GLLVM and then present the design and the results of the simulation study.

2 Alternative estimators for the GLLVM

2.1 The underlying variable approach of LISREL

The underlying variable approach assumes that all the manifest variables are multivariate normal. If a variable is not normal, it is assumed to be an indirect observation of an underlying normal variable. This approach can be formulated as follows. Let $x^{(j)}$, $j = 1, \ldots, p$, one of the p manifest variables, be a Bernoulli variable, $\mathbf{z} = [1, z_1, \ldots, z_q]^T = [1, \mathbf{z}_{(2)}^T]^T$ a vector of latent variables with $q < p$, and $\boldsymbol{\alpha}_j = [\alpha_{j0}, \ldots, \alpha_{jq}]^T$ a vector of parameters (also called loadings). Let the conditional distribution of $y^{(j)}$ given $\mathbf{z}$ be normal with mean $\boldsymbol{\alpha}^T\mathbf{z}$ and unit variance. Given $\mathbf{z}$, a link is then established between $x^{(j)}$ and $y^{(j)}$ in that it is assumed that $x^{(j)}$ takes the value 1 if $y^{(j)}$ is positive and 0 otherwise. Then,

$$E\left[x^{(j)} \mid \mathbf{z})\right] = P(y^{(j)} > 0 \mid \mathbf{z}) = \Phi(\boldsymbol{\alpha}_j^T\mathbf{z}),$$

where $\Phi(\cdot)$ is the normal cumulative distribution function. We obtain from the last equation that

$$\text{probit}\left(E\left[x^{(j)} \mid \mathbf{z}\right]\right) = \Phi^{-1}\left(E\left[x^{(j)} \mid \mathbf{z}\right]\right) = \boldsymbol{\alpha}_j^T\mathbf{z}.$$

Consequently, the assumption of an underlying normal variable in the LISREL approach can be compared to the one with the GLLVM (see below), except that the link function is a probit instead of a logit. These two link functions are very close (see e.g. Lord and Novick 1968), so that in our simulations the estimators provided by LISREL can be compared to the LAMLE.

In practice, the model parameters are estimated in three steps (Jöreskog, 1969, 1990). First, the thresholds of the underlying variables are estimated from the univariate means of the manifest variables. In a second step, the correlation matrix between manifest and underlying variables is estimated using polychoric, polyserial and Pearson correlations depending on the type of manifest variables, and finally, the model parameters are obtained from a factor analysis. For the later, several methods are available (see Jöreskog and Sörbom 1993), and a popular one is the weighted least squares estimator (WLSE).

2.2 The LAMLE

The LAMLE of Huber, Ronchetti, and Victoria-Feser (2004) has been design to estimate the parameters of a GLLVM. The later describes the relationship between p manifest variables $x^{(j)}$and the q latent variables z_k by means of the conditional distributions $g_j(x^{(j)}|\mathbf{z})$, which belong to the exponential family (with canonical link)

$$g_j(x^{(j)}|\mathbf{z}) = \exp\left\{ \frac{x^{(j)}\boldsymbol{\alpha}_j^T\mathbf{z} - b_j(\boldsymbol{\alpha}_j^T\mathbf{z})}{\phi_j} + c_j(x^{(j)}, \phi_j) \right\}, \tag{1}$$

where b_j and c_j are known functions that depend on the chosen distribution g_j and ϕ_j is a scale parameter (McCullagh and Nelder 1989). Note that the canonical link for ordinal variables is the logit function. The essential assumption in GLLVM is that, given the latent variables, the manifest variables are conditionally independent, and therefore the joint conditional distribution of the manifest variable is

$$\prod_{j=1}^{p} g_j(x^{(j)}|\mathbf{z})h(\mathbf{z}_{(2)}^T). \tag{2}$$

where $h(\mathbf{z}_{(2)}^T)$ is the density of the latent variables which is assumed to be the standard normal with covariance $\mathbf{I}_q$. The last assumption of independence can actually be relaxed. Since the latent variables are not observed, their realizations are treated as missing, and are integrated out, giving the marginal density of the manifest variables

$$f_{\boldsymbol{\alpha},\boldsymbol{\phi}}(\mathbf{x}) = \int \left\{ \prod_{j=1}^{p} g_j(x^{(j)}|\mathbf{z}) \right\} h(\mathbf{z}_{(2)}^T)\mathbf{d}\mathbf{z}_{(2)}^T. \tag{3}$$

with $\boldsymbol{\alpha} = [\boldsymbol{\alpha}_1 \ldots \boldsymbol{\alpha}_p]^T$ and $\boldsymbol{\phi} = [\phi_1 \ldots \phi_p]^T$.

Given a sample $\mathbf{x}_i = [x_i^{(1)}, \ldots, x_i^{(p)}]$, $i = 1, \ldots, n$, one can use the log-likelihood function to estimate the parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\phi}$. The later contains a multidimensional integral which cannot be computed explicitly, except when all the g_j are normal. Huber, Ronchetti, and Victoria-Feser (2004) propose to use a Laplace approximation for the integrals (see also Tierney and Kadane 1986) leading to implicit estimators that we do not present here. The error rate is of order p^{-1}, where p is the number of manifest variables and hence the approximation improves as the number of latent variables grows (because with more latent variables one needs more manifest variables). In addition, the Laplace approximation yields automatically estimates of individual latent scores $\widehat{z}_j$ (see Huber, Ronchetti, and Victoria-Feser 2004). Finally, Huber, Ronchetti, and Victoria-Feser (2004) show that the LAMLE belongs to the class of M-estimators from which they derive the asymptotic normality and inference.

It should also be noted, that the model in (3) is not identifiable unless one imposes constraints on the loadings $\boldsymbol{\alpha}$. Huber, Ronchetti, and Victoria-Feser (2004) provide the necessary conditions for the LAMLE of $\boldsymbol{\alpha}$ to be unique. Alternatively, one can use a rotation such as a varimax rotation.

3 Simulation study

3.1 Design

The aim of the simulation study is to compare two different estimators for a confirmatory factor analysis model, namely the LAMLE (i.e. provided by LCube) and the one based on a normal factor analysis with polychoric correlations as input and with the WLSE (i.e. provided by LISREL). The population model we consider here is presented in Figure 1. We have 10 manifest ordinal variables and through a factor analysis (unconstrained but with varimax rotation) we try to recover the model. The factors Z_1 and Z_2 are independent and normally distributed with zero mean and unit variance. Given these factors, the manifest variables are (in a first step) normally distributed with means computed by means of the factor loadings and unit variances. These variables are then transformed in ordinal variables (5 categories) following two methods:

- Ordinal-symmetric: the first 5% on the normal values were assigned the value of 1 on the ordinal scale, the following 21% the value of 2, the following 48% the value of 3, the following 21% the value of 4 and the 5% remaining the value of 5.
- Ordinal-non-symmetric: the cutoff points were chosen to correspond to 75%, 15%, 5%, 3% and 2% of the normal values.

Half of the manifest variables were transformed in ordinal-symmetric variables, the others in ordinal-non-symmetric (see model in Figure 1). We consider 2 sample sizes, namely 100 and 300. We made 100 replications for each sample size.

From the estimates distribution, we are able to check two important aspects of the model. The first and most natural one is the estimator's bias in estimating the factor loadings. We do so by estimating a unconstrained two factors model and apply a varimax rotation. The second aspect, is the correlation matrix of the underlying normal manifest variables. In other words we ask the following question: given that the true generating process for the manifest variables is the normal distribution with mean $\mathbf{0}$ and correlation matrix

$$\mathbf{R} = \boldsymbol{\alpha}^T\boldsymbol{\alpha} + \boldsymbol{\Psi} \tag{4}$$

where $\boldsymbol{\Psi}$ is the diagonal matrix of so-called residual variances (or uniqueness) and is such that $\text{diag}\big(\boldsymbol{\alpha}^T\boldsymbol{\alpha} + \boldsymbol{\Psi}\big) = \mathbf{1}$, can the estimators, after transforming the normal variables into ordinal ones, recover the original correlation matrix? This will be done by analyzing the bias distribution of the constructed correlation matrices using the LAMLE and of the polychoric correlation matrices provided by LISREL

3.2 Loadings estimates

We present the results in the form of boxplots of the estimators bias distributions. These are given in Figures 2 and 3 for the loadings of respectively the first and second latent variable, with $n = 100$. The 10 graphs in each Figure corresponds to the 10 loadings of the unconstrained two factors model. The estimates are those obtained after varimax rotation and centered at the true value of the corresponding loading. In each graph, the left boxplot is for the estimators provided by LISREL (LIS) and the right boxplot is for the LAMLE provided by LCube (LCu).

For the first latent variable Z_1 (Figure 2), one notices that both estimators are unbiased for the loadings corresponding to a nil value (i.e. for X_5 to X_{10}). However, for the loadings of X_1 to X_4, the LISREL estimator is clearly biased for the ordinal-non-symmetric variables (i.e. X_2 and X_4) and even slightly biased for the ordinal-symmetric ones. On the other hand, the LAMLE is unbiased in all situations. For the second latent variable (Figure 3), the same conclusion can be drawn.

We also performed the same analysis with $n = 300$ but do not present the results here. We found the same features, that is to say that the LISREL estimator is biased, whereas the LAMLE is not. In fact, the bias of the LISREL estimator doesn't disappear as n grows.

3.3 Correlation estimates

We also present the results in the form of boxplots of the estimators' bias distributions. For the LAMLE, we use the relationship (4) to estimate the correlation matrix of the original manifest variables. For LISREL, we use the provided polychoric correlations. Figure 4 is for the LAMLE and Figure 5 is for LISREL. Each graph in each Figure corresponds to one row of $\mathbf{R}$. In each row, one of the correlation is by definition equal to 1, and therefore the corresponding boxplot is just a line at 0.

For the LAMLE (Figure 4), all correlation estimators are unbiased. For the polychoric correlations provided by LISREL (Figure 5), one can notice that about half of the correlations estimators are biased. The polychoric correlations actually underestimate the true correlations when both variables load on one of the latent variables. This explains why the resulting loadings, obtained by a factor analysis based on the polychoric correlation are themselves biased.

4 Conclusion

Latent variable models are important in many disciplines and have recently attracted a considerable attention (see e.g. Bartholomew and Knott 1999 and Skrondal and Rabe-Hesketh 2004). When the manifest variables are not normal but instead ordinal for example, the estimation problem is challenging because of the complicated form of the likelihood function. Several estimators have been proposed sofar like the one based on polychoric, polyserial and Pearson correlations as implemented in most statistical packages, LISREL being one of them. Other estimators are based on numerical approximations of the integrals in the likelihood function which include simple or adaptive Gauss-Hermite quadratures, stochastic approximations such as MCMC and MCEM and the Laplace approximation (as proposed by Huber, Ronchetti, and Victoria-Feser 2004). We have argued that an adaptive Gauss-Hermite quadrature or stochastic approximations are very computationally intensive so that complicated models are almost impossible to estimate, whereas the LAMLE is quite fast and can be used with relatively complex models. The question we have addressed in this paper is to what extent and in practice, the LAMLE can improve a statistical analysis compared to a traditional approach such as LISREL. For that, we have performed a simulation study involving a quite common model with ordinal manifest variables, some of which are non symmetrically distributed. It should be stressed that this type of situation is very common in practice. We have concluded that the LISREL estimates can be seriously biased, that even if the resulting estimator is based on the underlying (normal) variable approach, it cannot recover the original normal correlation structure, whereas on the other hand, the LAMLE is unbiased and is able to recover the original correlation structure.

References

Arminger, G. and U. Küsters (1988). *Latent Trait Models*. New York: Plenum Press.

Bartholomew, D. J. (1984a). The foundations of factor analysis. *Biometrika 71*, 221–232.

Bartholomew, D. J. (1984b). Scaling binary data using a factor model. *Journal of the Royal Statistical Society, Series B 46*, 120–123.

Bartholomew, D. J. and M. Knott (1999). *Latent Variable Models and Factor Analysis*. Kendall's Library of Statistics 7. London: Arnold.

Breslow, N. E. and D. G. Clayton (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association 88*, 9–25.

DiStefano, C. (2002). The impact of categorization with confirmatory factor analysis. *Structural Equation Modeling 9*(3), 327–346.

Huber, P., E. Ronchetti, and M.-P. Victoria-Feser (2004). Estimation of generalized latent trait models. *Journal of the Royal Statistical Society, Serie B*. To appear.

Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika 34*, 183–203.

Jöreskog, K. G. (1990). New developments in LISREL: Analysis of ordinal variables using polychoric correlations and weighted least squares. *Quality and Quantity 24*, 387–404.

Jöreskog, K. G. and D. Sörbom (1993). *LISREL 8: Structural Equation Modeling with SIMPLIS Command Language*. London: Lawrence Erlbaum.

Lee, Y. and J. A. Nelder (1996). Hierarchical generalized linear models. *Journal of the*

Royal Statistical Society, Serie B 58, 619–678.

Lin, X. and N. E. Breslow (1996). Bias correction in generalized linear mixed models with multiple components of dispersion. *Journal of the American Statistical Association 91*, 1007–1016.

Lord, F. M. and M. E. Novick (1968). *Statistical Theories of Mental Test Scores*. Addison-Wesley.

McCullagh, P. and J. A. Nelder (1989). *Generalized Linear Models*. London: Chapman and Hall. Second edition.

McGilchrist, C. A. (1994). Estimation in generalized mixed models. *Journal of the Royal Statistical Society, Serie B 56*(1), 61–69.

Moustaki, I. (1996). A latent trait and a latent class model for mixed observed variables. *British Journal of Mathematical and Statistical Psychology 49*, 313–334.

Moustaki, I. and M. Knott (2000). Generalized latent trait models. *Psychometrika 65*, 391–411.

Rabe-Hesketh, S., A. Skrondal, and A. Pickles (2002). Reliable estimation of generalized linear mixed models using adaptive quadrature. *The Stata Journal 2*, 1–21.

Skrondal, A. and S. Rabe-Hesketh (2004). *Generalized Latent Variable Modeling*. London: Chapman and Hall.

Tierney, L. and J. B. Kadane (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association 81*, 82–86.

Yau, K. K. W. and C. A. McGilchrist (1996). Simulation study of the GLMM method applied to the analysis of clustered survival data. *Journal of Statistical Computation and Simulation 55*, 189–200.

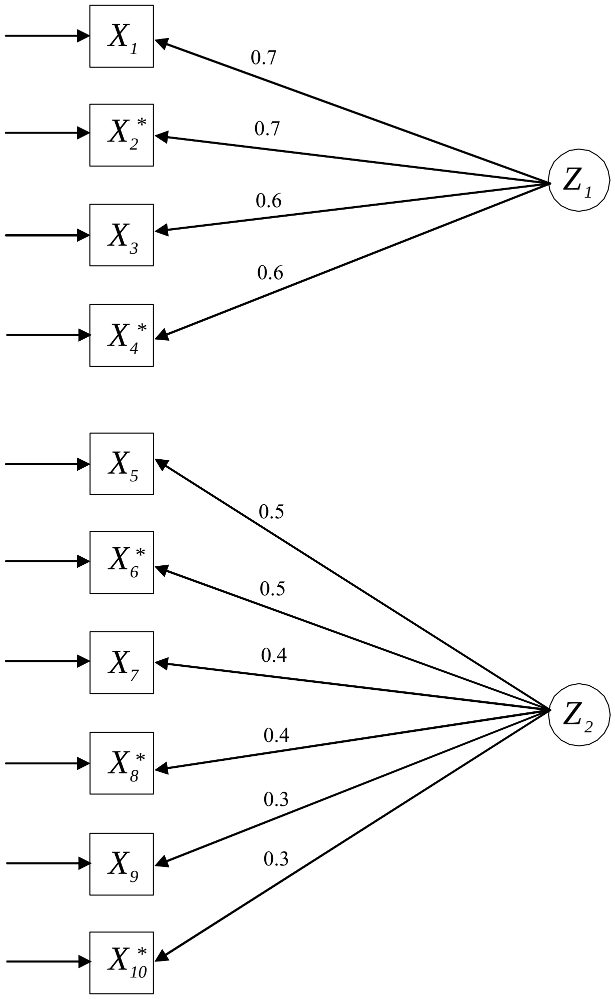

Figure 1: Factor analysis population model, starred items for the non-symmetric ordinal items

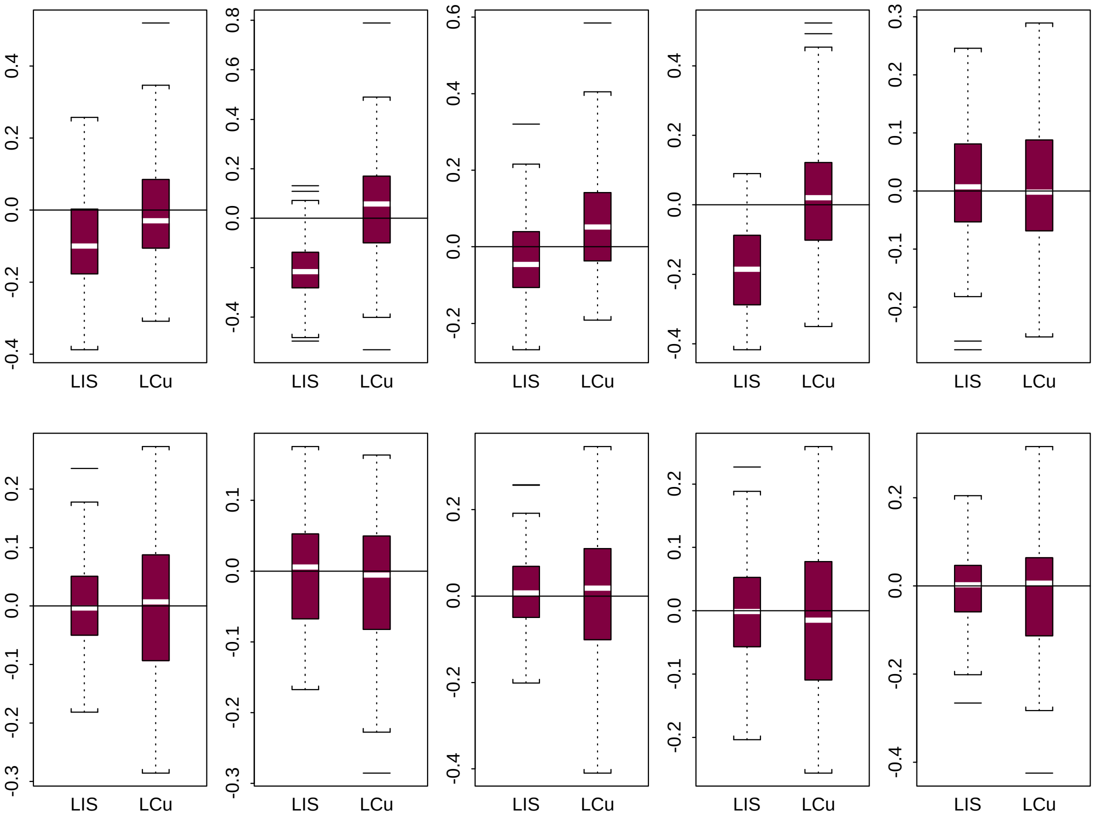

Figure 2: Bias distribution of loadings' estimates for Z_1, $n = 100$

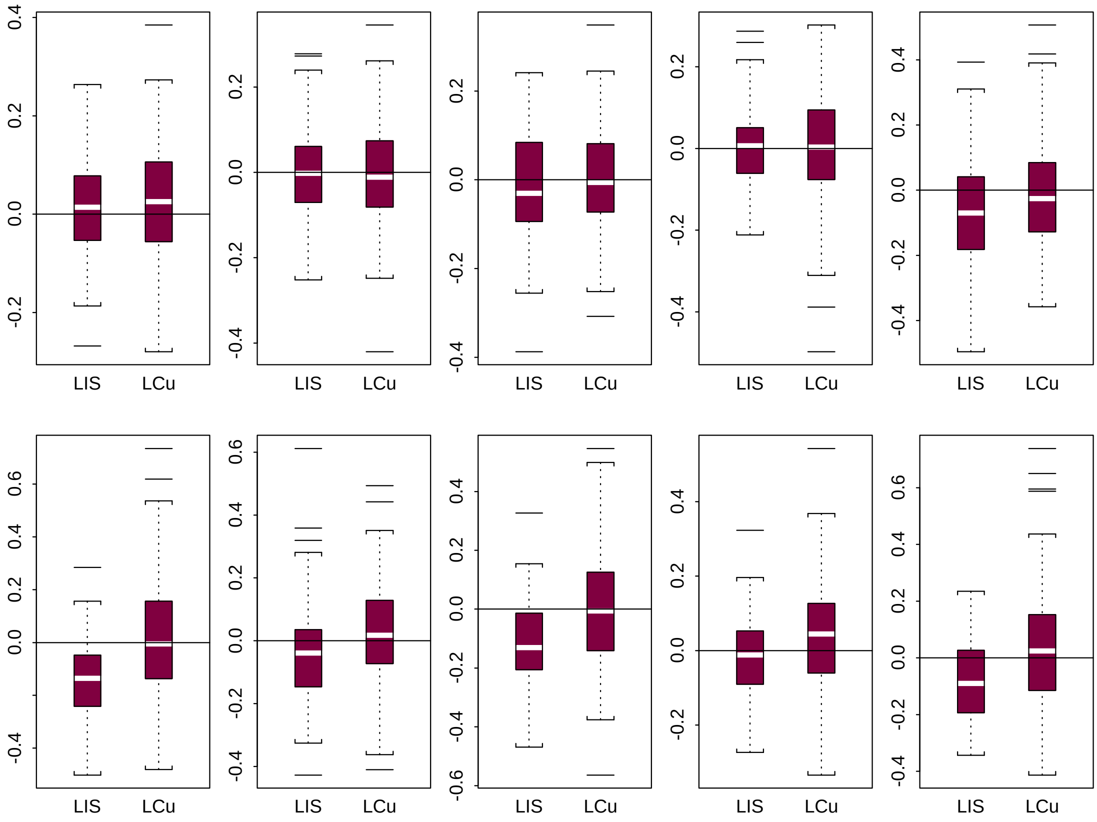

Figure 3: Bias distribution of loadings' estimates for Z_2, $n = 100$

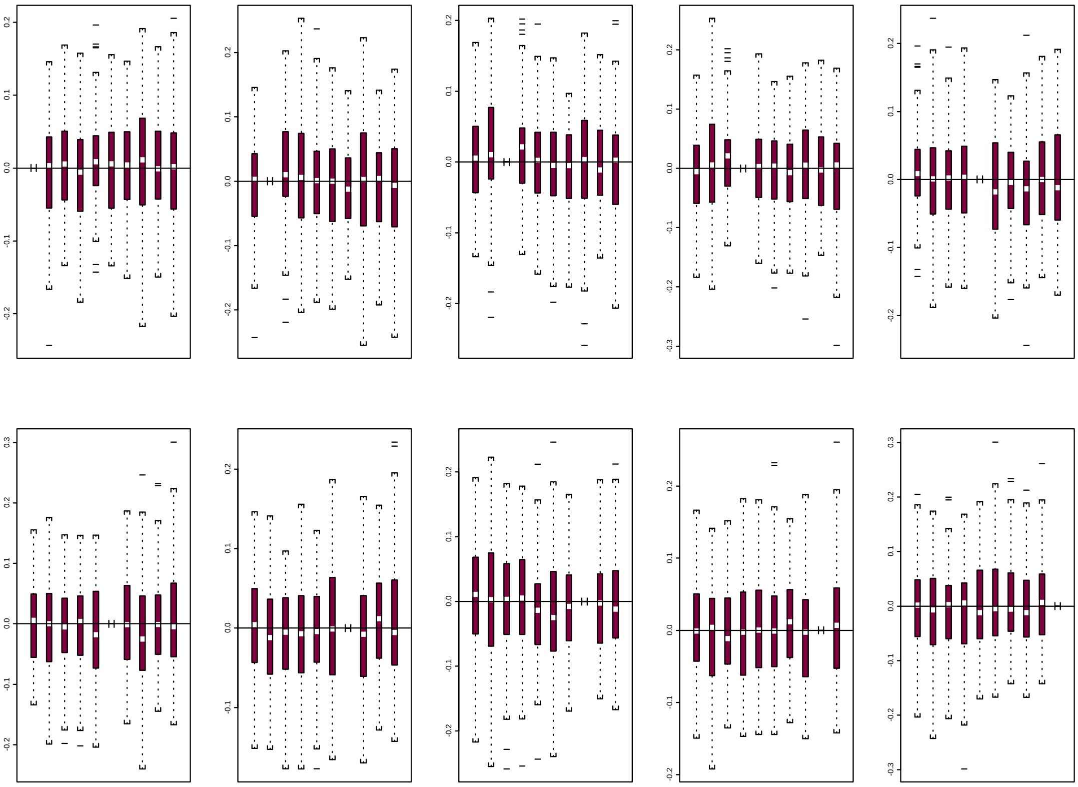

Figure 4: Bias distribution of the underlying normal variables correlations' estimates constructed from the LAMLE, $n = 100$

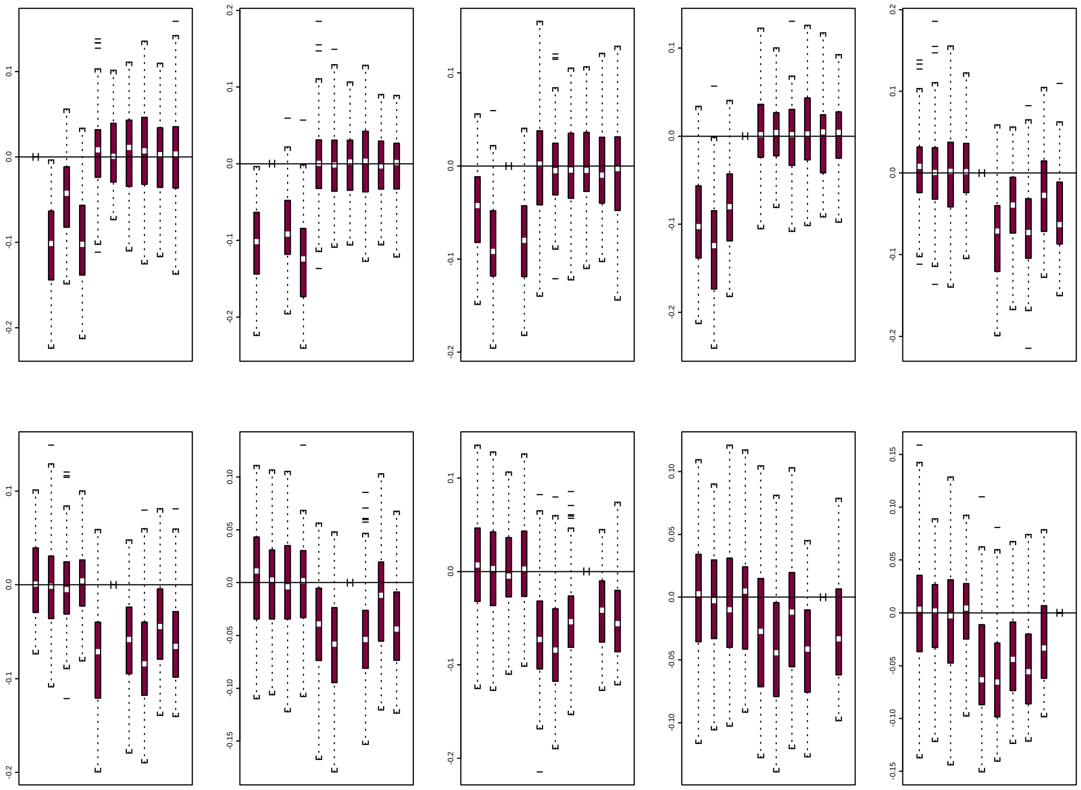

Figure 5: Bias distribution of the underlying normal variables polychoric correlations' estimates in LISREL, $n = 100$