Ioulia Papageorgiou and Irini Moustaki

Sampling of pairs in pairwise likelihood estimation for latent variable models with categorical observed variables

Article (Accepted version) (Refereed)

Sampling of pairs in pairwise likelihood estimation for latent variable models with categorical observed variables

Ioulia Papageorgiou*and Irini Moustaki†

12 April 2018

Abstract

Pairwise likelihood is a limited information estimation method that has also been used for estimating the parameters of latent variable and structural equation models. Pairwise likelihood is a special case of composite likelihood methods that uses lower order conditional or marginal log-likelihoods instead of the full log-likelihood. The composite likelihood to be maximized is a weighted sum of marginal or conditional log-likelihoods. Weighting has been proposed for increasing efficiency but the choice of weights is not straightforward in most applications. Furthermore, the importance of leaving out higher order scores to avoid duplicating lower order marginal information has been pointed out. In this paper, we approach the problem of weighting from a sampling perspective. More specifically, we propose a sampling method for selecting pairs based on their contribution to the total variance from all pairs. The sampling approach does not aim to increase efficiency but to decrease the estimation time, especially in models with a large number of observed categorical variables. We demonstrate the performance of the proposed methodology using simulated examples and a real application.

keywords Principal Component Analysis, Structural equation models, Factor analysis, Composite likelihood.

*Athens University of Economics and Business, Department of Statistics, 76 Patission, 104 34 Athens, Greece, ioulia@aueb.gr

†London School of Economics, Houghton Street, London WC2A 2AE, U.K., i.moustaki@lse.ac.uk

1 Introduction

Latent variable models (LVM) and structural equation models (SEM) are widely used for analyzing multiple observed variables from social surveys and administrative data. Our models treat jointly the observed variables as measures of unobserved (latent) constructs. Latent variable models are mainly used to explain associations between p observed variables, denoted by $Y_1, \ldots, Y_p$, that are assumed to be indicators of q ($q < p$) latent variables denoted by $\xi_1, \ldots, \xi_q$ (exploratory factor analysis), such as attitudes, beliefs, and abilities, but also to test specific hypotheses driven by social or economic theory (confirmatory factor analysis and SEM). Confirmatory factor analysis (CFA) postulates certain relationships among the observed and the latent variables by assuming a pre-specified pattern for the model parameters (factor loadings). CFA is mainly used for testing a hypothesis arising from theory and therefore, the number of latent variables and the variables that will be used to measure each latent variable are known in advance (Bartholomew, Steele, Moustaki, and Galbraith, 2008). Substantive research questions typically centre on associations involving the latent variables, for example how an individual's attitude depends on covariates such as age and education.

Questionnaire items in social surveys are often of a categorical nature (ordinal or nominal). In the SEM literature, one common approach for the analysis of categorical variables with factor models is the underlying variable approach (UVA) in which categorical variables are assumed to be generated by underlying continuous variables (e.g., see Jöreskog, 1990, 1994; Lee, Poon, and Bentler, 1990, 1992; Muthén, 1984). Under the UVA, full maximum likelihood is not feasible for a large number of p and limited information estimation methods have been proposed instead, such as the three-stage least squares estimation method (Jöreskog, 1990, 1994; Muthén, 1984) and composite estimation methods.

Composite likelihood estimation methods (CLM) (see e.g., Besag, 1974; Lindsay, 1988; Cox and Reid, 2004; Varin, 2008; Varin, Reid, and Firth, 2011) have been developed for when the full likelihood is too expensive or intractable to compute. The main idea behind CLM is to construct a pseudolikelihood from marginal or conditional densities of lower dimension than the original data. Usually, the composite likelihood is given as a weighted sum of lower-order marginal or conditional log-likelihoods. Composite likelihood estimators subject to regularity conditions have the desired properties of being asymptotically consistent, and normally distributed. de Leon (2005) used the pairwise maximum likelihood approach to estimate thresholds and polychoric correlations of ordinal data. Jöreskog and Moustaki (2001) proposed an underlying bivariate normal method that maximised the sum of all the univariate and bivariate log-likelihoods. Liu (2007) proposed a multistage estimation method for SEM as an alternative to the commonly used three-stage methods. Pairwise likelihood has also been found to work well and to be computationally attractive over full information maximum likelihood (FIML) for SEM for binary, ordinal and ranking variables (Katsikatsou, Moustaki, Yang-Wallentin, and Jöreskog, 2012; Katsikatsou, 2013) and for factor analysis models for longitudinal data where both latent variables and random effects are used to account for item depen-

dencies (Vasdekis, Cagnone, and Moustaki, 2012). In all the above papers, composite likelihood is defined as the sum of all pairwise log-likelihoods. Furthermore, Chan and Bentler (1998) and Fieuws and Verbeke (2006) used the composite likelihood for a covariance structure analysis for ranking data and for estimating mixed effects models for multivariate longitudinal outcomes respectively. In their implementation of the composite likelihood, each pairwise likelihood is maximized separately and the final parameter estimates are obtained as a simple average of the estimates produced by the separate bivariate maximizations. Recently, Vasdekis, Rizopoulos, and Moustaki (2014) proposed a weighted estimator instead that improves parameter efficiency. Their method is also based on separate maximizations of pairwise likelihoods and a weighted average of the individual estimates with weights obtained by minimizing the variance of the estimates.

Pairwise likelihood estimation can be computationally demanding in problems with a large number of variables. In this paper, we propose a sampling method for estimating the model parameters from a sub-set of pairs selected from the population of all possible bivariate densities to be included in the pairwise log-likelihood function to be maximized. The methodology developed here can be also extended to other multivariate models for which their model parameters can be identified by bivariate likelihoods, but the details are illustrated here only for a latent variable model for categorical variables. Pairwise likelihood estimation has been developed for the multivariate probit model with random effects, a poisson-lognormal mixture model and a multi-normal copula model (Zhao and Joe, 2005), for the spatial generalized linear mixed models (Varin, Host, and Skare, 2005), for ordinal categorical time series (Varin and Vidoni, 2006), for generalized linear models with crossed random effects (Bellio and Varin, 2005), Poisson regression models with time-varying Gamma frailty for modeling longitudinal count data (Henderson and Shimakura, 2003) and for binary spatial data (Heagerty and Lele, 1998b). This list of papers is not by any means exhaustive but provides an idea on how popular pairwise likelihood estimation is for estimating complex multivariate models and the advantages of developing methods that could potentially reduce even further the computational complexity and time.

The paper is organized as follows. Section 2 presents some general results on pairwise estimation, Section 3 provides a brief presentation of the estimation of the factor analysis model for categorical variables using pairwise likelihood estimation, Section 4 discusses the proposed sampling method of pairs, Section 5 studies the performance of the proposed methods on estimation and section 6 on inference using simulations, Section 7 applies the proposed sampling methods to a real data set and Section 8 concludes.

2 General results on pairwise likelihood estimation

Suppose there is a $p \times 1$ vector of observed variables: $\mathbf{Y}' = (Y_1, Y_2, \ldots, Y_p)$. Let $f_{\mathbf{y}}(.;\boldsymbol{\theta})$ denote the joint density of a parametric model for the data and $f_{y_i,y_j}(.;\boldsymbol{\theta})$ the bivariate marginal density, with parameter vector $\boldsymbol{\theta}$. Following Lindsay (1988), the pairwise log-

likelihood for a random sample of size n is

$$pl = \sum_{h=1}^{n} \sum_{i<j} w_{ij,h} \log f_{y_{ih},y_{jh}}(y_{ih}, y_{jh}; \boldsymbol{\theta}), \tag{1}$$

where $w_{ij,h}$ are non-negative weights and the parameter vector $\boldsymbol{\theta}$ needs to be identifiable from the set of bivariate margins. Unequal weights can be chosen to increase efficiency and therefore depend on some measure related to asymptotic relative efficiency. Weighting the likelihood components has been proposed in time series and spatial analysis (Lindsay, 1988; Varin et al., 2005; Heagerty and Lele, 1998a; Varin and Czado, 2010) and in clustered data (Joe and Lee, 2009). For example, for clustered data the weights could take into account the degree of cluster homogeneity where, in longitudinal or spatial data, the weights could downweight pairs that are far apart in time or space. For a lengthy discussion on the topic see also Varin et al. (2011) and references therein. Lindsay has also mentioned in seminars the need to use fewer of the higher order moments in an attempt to avoid duplicating lower order marginal information. It is expected that CLM can also be more robust under possible misspecification of the higher order dimensional distributions and they can allow a less complex structure on the parameter space that might lead to a smoother likelihood surface.

The central limit theorem for the composite likelihood score statistic implies that the distribution of $\hat{\boldsymbol{\theta}}_{PL}$ can be approximated by the Normal with mean $\boldsymbol{\theta}$ and variance-covariance matrix $G^{-1}(\boldsymbol{\theta})$ where $G(\boldsymbol{\theta})$ is the Godambe information matrix (also known as sandwich information) (Varin, 2008; Varin et al., 2011). In particular, $G(\boldsymbol{\theta}) = H(\boldsymbol{\theta})J^{-1}(\boldsymbol{\theta})H(\boldsymbol{\theta})$, where $H(\boldsymbol{\theta})$ is the sensitivity matrix,

$H(\boldsymbol{\theta}) = E\left\{-\nabla^2 pl(\boldsymbol{\theta}; \mathbf{x})\right\}$, and $J(\boldsymbol{\theta})$ is the variability matrix, $J(\boldsymbol{\theta}) = Var\left\{\nabla pl(\boldsymbol{\theta}; \mathbf{x})\right\}$. In general, the identity $H(\boldsymbol{\theta}) = J(\boldsymbol{\theta})$ does not hold in the case of composite likelihoods. The assumed independence among the likelihood components forming the composite likelihood is usually not valid when the full likelihood is considered. The sample estimates of $H(\boldsymbol{\theta})$ and $J(\boldsymbol{\theta})$ are

$\hat{H}(\hat{\boldsymbol{\theta}}_{PML}) = \nabla^2 pl(\hat{\boldsymbol{\theta}}_{PML}; \mathbf{x})$, and

$\hat{J}(\hat{\boldsymbol{\theta}}_{PML}) = \frac{1}{n}\sum_{h=1}^{n} \left(\nabla pl(\boldsymbol{\theta}; \mathbf{x}_h)\right)\left(\nabla pl(\boldsymbol{\theta}; \mathbf{x}_h)\right)^T$ respectively. The ratio of $G(\boldsymbol{\theta})$ to the expected Fisher information $I(\boldsymbol{\theta})$ determines the asymptotic efficiency of $\hat{\boldsymbol{\theta}}_{PL}$.

Furthermore, Wald, score and LR test statistics under CLM are available (Pace, Salvan, and Sartori, 2011) as well as AIC (Varin and Vidoni, 2005) and BIC model (Gao and Song, 2010) selection criteria. Recently, LR test statistics and AIC and BIC have been also developed for SEM under CLM (Katsikatsou and Moustaki, 2016).

3 Factor analysis models for categorical data

Let $\mathbf{Y}' = (Y_1, Y_2, \ldots, Y_p)$ denote a vector of p ordinal/binary observed variables (items) for a single respondent, where Y_i has m_i categories, $i = 1, \ldots, p$. There are $R = \prod_{i=1}^{p} m_i$ possible response patterns. Categorical variables Y are generated by underlying unobserved continuous variables Y^*, assumed to be normally distributed. The connection

between an observed categorical (ordinal) variable Y_i and the underlying continuous variable $Y_i^\star$ is

$$Y_i = c_i \Longleftrightarrow \tau_{c_i-1}^{(i)} < Y_i^\star < \tau_{c_i}^{(i)} \,, \tag{2}$$

where $\tau_{c_i}^{(i)}$ is the c_i^{th} threshold of variable Y_i and $-\infty = \tau_0^{(i)} < \tau_1^{(i)} < \ldots < \tau_{m_i-1}^{(i)} < \tau_{m_i}^{(i)} = +\infty$. In the binary case, there is only one threshold for each observed variable. Since only ordinal information is available, the distribution of $Y_i^\star$ is determined only up to a monotonic transformation. In practice it is convenient to assume a standard normal distribution. The Y variables are regarded as measures of a vector of latent variables, representing some constructs of interest. The classical factor analysis model is of the form

$$\mathbf{Y}^\star = \Lambda\boldsymbol{\xi} + \boldsymbol{\delta} \,, \tag{3}$$

where Λ is the $p \times q$ matrix of factor loadings, $\boldsymbol{\xi}$ is the $q \times 1$ vector of latent variables, $\boldsymbol{\delta}$ is the p-dimensional vector of unique variables. The elements of Λ play a key role in interpreting the factors. In addition, it is assumed that $\boldsymbol{\xi} \sim N_q(\mathbf{0}, \Phi)$ where Φ is a $q \times q$ matrix that has ones on its main diagonal being this way, the correlation matrix of latent factors, and $\boldsymbol{\delta} \sim N_p(\mathbf{0}, \Theta)$ with Θ a $p \times p$ diagonal matrix, $\Theta = I - \text{diag}(\Lambda\Phi\Lambda')$, and $Cov(\boldsymbol{\xi}, \boldsymbol{\delta}) = \mathbf{0}$. The diagonal elements of the matrix Φ are set to 1 for identifying the scale of the latent variables. The parameter vector $\boldsymbol{\theta}' = (\boldsymbol{\lambda}, \boldsymbol{\varphi}, \boldsymbol{\tau})$ contains $\boldsymbol{\lambda}$ and $\boldsymbol{\varphi}$ which are the free non-redundant parameters in matrices Λ and Φ, respectively, and $\boldsymbol{\tau}$ is a vector of all free thresholds.

For a random sample of size n the pairwise log-likelihood, $pl(\boldsymbol{\theta}; \mathbf{y})$, is maximized with respect to the parameter vector $\boldsymbol{\theta}' = (\boldsymbol{\lambda}, \boldsymbol{\varphi}, \boldsymbol{\tau})$ given by

$$\begin{aligned} pl(\boldsymbol{\theta}; \mathbf{y}) = \sum_{i<j} \ln L\left(\boldsymbol{\theta}; (y_i, y_j)\right) \quad = \\ \sum_{i<j} \sum_{c_i=1}^{m_i} \sum_{c_j=1}^{m_j} n_{c_i c_j}^{(ij)} \ln \pi_{c_i c_j}^{(ij)}(\boldsymbol{\theta}) \end{aligned} \tag{4}$$

where $n_{c_i c_j}^{(ij)}$ is the observed frequency of a response in category c_i and c_j for variables Y_i and Y_j, respectively and

$$\begin{aligned} \pi_{c_i c_j}^{(ij)}(\boldsymbol{\theta}) &= \pi\left(Y_i = c_i, Y_j = c_j; \boldsymbol{\theta}\right) \\ &= \Phi_2\left(\tau_{c_i}^{(i)}, \tau_{c_j}^{(j)}; \rho_{ij}\right) - \Phi_2\left(\tau_{c_i}^{(i)}, \tau_{c_j-1}^{(j)}; \rho_{ij}\right) \\ &- \Phi_2\left(\tau_{c_i-1}^{(i)}, \tau_{c_j}^{(j)}; \rho_{ij}\right) + \Phi_2\left(\tau_{c_i-1}^{(i)}, \tau_{c_j-1}^{(j)}; \rho_{ij}\right) \,, \end{aligned} \tag{5}$$

where $\Phi_2\left(a, b; \rho\right)$ is the bivariate cumulative normal distribution with correlation ρ evaluated at the point (a, b),

$$\rho_{ij}(\boldsymbol{\theta}) = \boldsymbol{\lambda}_i \Phi \boldsymbol{\lambda}_j^{'} \,,$$

and $\boldsymbol{\lambda}_i$ are the factor loadings and Φ the $q \times q$ covariance matrix of the latent variables. Details can be found in Katsikatsou et al. (2012). The factor analysis model for binary observed variables is a special case with $m_i = 2$. Furthermore, a mixture of binary and ordinal variables can also be considered under this framework. The PML estimation for SEM models developed in Katsikatsou et al. (2012) set the weights $w_{ij,h}$ for all i, j, h in Equation (1) equal to one. PML estimation and testing for SEM have been implemented in the R package `lavaan` (Rosseel, 2012; Rosseel, Oberski, Byrnes, Vanbrabant, Savalei, and Merkle, 2012). The advantage of PML over FIML estimation is that the former requires the evaluation of up to two-dimensional normal probabilities, regardless of the number of observed or latent variables.

4 Selecting a subset of pairs in pairwise likelihood

As shown in Section 3, for the construction of the PML in (4), all the possible bivariate marginal densities out of p are being used. There are two potential drawbacks with the use of all the possible pairwise likelihoods. The first relates to the increase of computational time needed to estimate models with a large number of observed variables. For example, the number of pairs needed to evaluate the pairwise likelihood for $p = 20, 30$ or 40 variables are $190, 435$ and 780 respectively. The second issue relates more to literature findings that pruning of higher order moments will result in less repetition of the same variable occurring in many pairs and ultimately to an increase in efficiency. To tackle the computational demands we propose to take a subset/ sample from the population of bivariate densities that produce unbiased estimates. The selection of pairs is done first and the model is estimated on the selected pairs only. This can be seen as a weighted PML where the weights are fixed to zero or one prior to estimation.

We need to make a distinction between two different sample and population sizes, namely, the sample size n that defines the number of sampled respondents from a population of N respondents and the sample size n^* that defines the number of selected pairs of variables selected from all the possible pairs of variables given by $N^* = \binom{p}{2}$. In our case, the population is defined as all the possible pairs of variables. The population elements are not independent since they share common and correlated variables. For the selection of pairs, a model-based sampling approach, in which a statistical model is assumed for the population elements is considered first (Sarndal, 1978; Bolfarine and Zacks, 1992). Our proposed approach consists of two stages. At stage 1, a subset of pairs is selected using an appropriate sampling scheme and at stage 2, the factor model given in (3) is estimated using PML (see Section 3) on the subset obtained from stage 1. Therefore, the inference part is provided through the PML estimation framework rather than the sampling scheme used to select the sample. An alternative to model-based sampling is the design-based sampling approach which is a probability design and no statistical model is assumed for the population.

Generally by adopting the model-based approach for sampling, we consider a population of size N^* distributed with a multivariate density with some population mean $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$. Usually, the population mean is assumed unknown

and the population quantity of interest depends on $\boldsymbol{\mu}$, whereas $\boldsymbol{\Sigma}$ is assumed known.

Let us denote with $\mathbf{X} = (X_1, \cdots, X_{N^*})$ the population random vector with elements representing the complete set of all possible pairs of variables (Y_i, Y_j) and $i \neq j$. Under the assumed model, the mean and covariance matrix of $\mathbf{X}$ are denoted by $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ respectively. In our case, the parameter vector $\boldsymbol{\mu}$ is not relevant and can be disregarded. Typically finding the optimal sampling design requires a minimization of the MSE which can be infeasible in many applications, including ours. We adopt instead a method proposed by Chao (2004) in which the sampling design results from an iterative algorithm rather than a minimization. The idea behind this is to use the population covariance matrix $\boldsymbol{\Sigma}$ and select those population elements that account for as much as possible of the total population variance. The proposed algorithm is based on the spectral decomposition of matrix $\boldsymbol{\Sigma}$ that corresponds to the known data reduction method of principal component analysis (PCA). The selected sample, although not optimal according to the strict definition of minimizing the MSE of the estimator, is proven to have good properties and it is an improvement on simple random sampling (SRS) where no particular model for the population is being assumed. The objective is to select, for a fixed sample size n^*, the units: $s = (X_{i_1}, \ldots, X_{i_{n^*}})$ from N^*, for $i_j \neq i_{j'}, \forall j \neq j'$. We discuss below the steps of the proposed sampling scheme.

Let us denote with $\lambda_1 > \lambda_2 > \cdots > \lambda_{N^*}$ and $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_{N^*}$ the eigenvalues and their corresponding eigenvectors of $\boldsymbol{\Sigma}$. According to PCA, the i^{th} principal component is written as a linear combination of the X's:

$$\mathbf{e}_i'\mathbf{X} = e_{i1}X_1 + e_{i2}X_2 + \cdots + e_{iN^*}X_{N^*}$$

where e_{ij} is the j^{th} component in the i^{th} eigenvector $\mathbf{e}_i$ and $i = 1, \ldots, N^*$. Each X_i corresponds to a pair of variables. The sampling units are selected based on the magnitude and the sign of their corresponding components in the leading k eigenvectors.

The sample size n^* is taken to be equivalent to the number of eigenvalues that are greater than one. Therefore, the sample size is chosen through the sampling procedure and depends on the correlation among the pairs. In case of confirmatory factor analysis with more than one factor, the number of eigenvalues that are greater than one from the total matrix is augmented by the number of eigenvalues that are greater than one for the sub-matrices which correspond to each factor. The algorithm as described in Chao (2004) is given below:

Step 1 The first step defines a set of elements denoted by s' from which the first element will be selected. However, the element to be selected is decided after all the remaining elements have been selected.

Let $s' = (j_1, j_2, ..., j_m)$, where $m < N^*$,

$$|e_{1j_1}| \geq |e_{1j_2}| \geq \cdots \geq |e_{1j_m}| \geq \cdots \geq |e_{1j_{N^*}}|$$

and m is an integer that indicates the number of units in s'. m can be appropriately specified before the survey according to the population size N^*.

$\vdots$

Step k Let $s_{tmp} = \{l_1, l_2\}$ where l_1 and l_2, satisfy:

l_1, l_2 not having been selected into s.

$$|e_{kl_1}| = \max_i |e_{ki}|$$

$$|e_{kl_2}| = \max_{\substack{j \\ e_{kj}e_{kl_1}<0}} |e_{kj}|$$

Units l_1 and l_2 will be added into s by

$$i_{2(k-1)} = l_1, i_{2k-1} = l_2 \quad \text{if} \quad n \geq 2k-1$$

$$i_{2(k-1)} = l_1 \quad \text{if} \quad n = 2(k-1)$$

- Repeat step k from $k = 2$ till $n = 2k-1$ or $n = 2(k-1)$.
- Final adjustment for selecting the first element: Let $s_{-i_1} = \{i_2, \ldots, i_{n^*}\}$ and $i_1 = j_p$, $j_p \in s'$ such that j_p satisfies

$$mcor(j_p, s_{-i_1}) = \min_{j_k \in s', j_k \notin s_{-i_1}} mcor(j_k, s_{-i_1})$$

where $mcor(j_k, s_{-i_1})$ is the multiple correlation coefficient between unit j_k and the set s_{-i_1}.

It becomes apparent that the sampling units are selected based not only on the magnitude but also the sign of the corresponding components of the largest eigenvectors. The algorithm is fast, and the only requirement is the population variance-covariance matrix, $\mathbf{\Sigma}$.

In our case, the population units are the pairs of variables and therefore, $\mathbf{\Sigma}$ refers to the covariance matrix of those units. In the computation of the covariance matrix among the population units we need to distinguish between two types of pairs; pairs that share a common variable and pairs that do not share a common variable. We propose two methods for obtaining the covariance matrix of the pairs, $\mathbf{\Sigma}$. The first method is based on the covariance matrix of the observed proportions. The total number of bivariate probabilities is $\binom{p}{2} \times (m-1)^2$ where $m = 2$ for binary variables.

For binary variables, let us define with

$\pi_{11}^{(ij)} = P(Y_i = 1, Y_j = 1)$ the population probability that (Y_i, Y_j) falls in the cell $(1, 1)$ or in other words of giving a positive response to variables i and j and with $P_{11}^{(ij)}$ the corresponding bivariate observed (sample) proportion. The covariances are given by

$$Cov(P_{11}^{(ij)}, P_{11}^{(ik)}) = \pi_{111}^{(ijk)} - \pi_{11}^{(ij)}\pi_{11}^{(ik)} \tag{6}$$

$$Cov(P_{11}^{(ij)}, P_{11}^{(kl)}) = \pi_{1111}^{(ijkl)} - \pi_{11}^{(ij)}\pi_{11}^{(kl)} \tag{7}$$

where $\pi_{111}^{(ijk)} = P(Y_i = 1, Y_j = 1, Y_k = 1)$ and
$\pi_{1111}^{(ijkl)} = P(Y_i = 1, Y_j = 1, Y_k = 1, Y_l = 1)$
For ordinal variables with m categories where

$c = 0, 1, \ldots, m-1$, the category 0 is not included in the covariance matrix but all the other combinations of variables and categories are given by

$$Cov(P_{c_i,c_j}^{(ij)}, P_{c_i,c_k}^{(ik)}) = \pi_{c_i,c_j,c_k}^{(ijk)} - \pi_{c_i,c_j}^{(ij)} \pi_{c_i,c_k}^{(ik)} \tag{8}$$

$$Cov(P_{c_i,c_j}^{(ij)}, P_{c_k,c_l}^{(kl)}) = \pi_{c_i,c_j,c_k,c_l}^{(ijkl)} - \pi_{c_i,c_j}^{(ij)} \pi_{c_k,c_l}^{(kl)} \tag{9}$$

where $i, j, k, l = 1, \ldots, p$, and $c_i, c_j, c_k, c_l = 1, \ldots, m-1$. $\pi_{c_i,c_j}^{(ij)}$ is the population probability that (Y_i, Y_j) falls in the cell (c_i, c_j) and $P_{c_i,c_j}^{(ij)}$ is the corresponding observed probability. Similarly we define the population probabilities $\pi_{c_i,c_j,c_k}^{(ijk)}$ and $\pi_{c_i,c_j,c_k,c_l}^{(ijkl)}$. For the calculation of $\mathbf{\Sigma}$, the population probabilities are replaced by their corresponding observed (sample) probabilities.

The second method is based on the RV-coefficient derived by Escoufier (1973) and Robert and Escoufier (1976). The RV-coefficient is introduced as a measure of similarity between two data matrices. If W and V are two data matrices the RV-coefficient is given by

$$RV(W, V) = \frac{tr(S_{12}S_{21})}{(trS_{11}^2 trS_{22}^2)^{1/2}} \tag{10}$$

where S_{11}, S_{22} are the covariance matrices for W and V respectively and S_{12} is the cross-covariance matrix.

In our context, W and V stand for two distinct pairs of variables, i.e. $W = (Y_i, Y_j)$ and $V = (Y_k, Y_l)$. The RV-coefficient takes values in $[0, 1]$ where a value close to 1 indicates a close pattern or co-structure for data sets W and V. Similarly, the higher the RV value is, the better W can substitute V and vice versa. The RV-coefficient can be seen as a measure of closeness of two configurations $C(W)$ and $C(V)$. It provides a single number for measuring the relationship between two data sets with different dimensions and it has been a useful tool in genetics, bioinformatics, ecology and high-dimensional data in general.

Several modifications of the RV-coefficient have been suggested in the literature. Most of them aim to improve its asymptotic properties. We mention here: the RV-adjusted (Mayer and Horgan, 2011), where the r-adjusted Pearson sample correlation coefficient instead of the Pearson coefficient is used in the calculations of S_{11}, S_{22} and S_{12}; the RLS coefficient (Lingoes and Schönemann, 1974) given by

$$RLS(W, V) = \frac{(tr(S_{12}S_{21}))^{1/2}}{(trS_{11}trS_{22})^{1/2}}$$

and, lastly, the Lg coefficient defined as the RV but for the computation of S_{11} and S_{22} the two principal eigen values are used in the denominator instead of the trace of the matrices.

4.1 Optimal controlled sampling designs

In this section, we discuss two design-based sampling schemes that can be implemented in addition to, or in combination with, the model-based design proposed in the previous section aiming to improve the efficiency of the derived estimators. The first sampling design, known as *Sampford's design* (Sampford, 1967), belongs to the class of fixed size sampling designs without replacement and inclusion probabilities proportional to size. The inclusion probabilities are calculated with the help of an auxiliary variable z. More specifically, if z_i is the value of unit i measured on z, the probability of the unit i to be selected in the sample, π_i, is calculated as $n^* z_i/Z$, where $Z = \sum z_i$, the sum of $z-$values over all population units. $p_i = z_i/Z$ is the probability of selecting unit i at any draw of the n^* in total. The Sampford's design is an iterative rejection procedure. More specifically, the first step is to select unit i with probability p_i and for the subsequent $n^* - 1$ draws to select unit i with probability proportional to $\frac{p_i}{1-n^* p_i}$ and with replacement. When a unit is selected a second time, it is rejected and the sampling starts again. The process ends when the sample size reaches the preassigned sampling size. Sampford's design has several theoretical advantages compared to other inclusion probability proportional to size sampling designs (e.g. see Haziza, Mecatti, and Rao, 2008). We adopted Sampford's plan by assuming as population the set of all possible pairs of variables and defining their sizes using as an auxiliary variable their corresponding magnitude, in absolute value, defined by the elements of the first eigenvector of the covariance matrix $\mathbf{\Sigma}$. The sample size is assumed to be the same as in Chao's implementation.

The second design-based sampling method is known as the *optimal control sampling method*. Control sampling is a methodology implemented to a sampling design with the aim of reducing the probability of selecting non-desired samples among the possible samples of the assumed sampling design. The associated sampling design before the use of control is called the uncontrolled sampling design. The first and second order inclusion probabilities of the uncontrolled sampling design, i.e. the probability of a population unit being included in the sample or the probability of two distinct population units both being included in the sample, are retained during the controlled sampling procedure (see, Rao and Nigam, 1992, 1990). The samples of a sampling plan are characterized as desired or undesired based on the sampling cost or the precision of estimation. Optimality is defined by minimizing the probability of selecting non-desired samples or, more generally, by optimizing a linear objective function such as a weighted sum of probabilities of samples selection. Optimal sampling can be implemented by using standard linear programming techniques.

We implement optimal control sampling as an attempt to combine Chao's method of sampling, which takes into account the correlation among population units, and the Sampford's sampling design, which is a proportional to size plan. More specifically, we assume as our uncontrolled sampling plan the Sampford's design with population the set of variables and with possible samples all comprising possible pairs of variables. To control the design we use Chao's sampling method as implemented in the previous section. A pair of variables is characterized as desired or undesired based on the fact

whether it has been selected by Chao's method or not. More analytically, the steps for the control Sampford design are:

1. Assume as population the set of observed variables $(Y_1, \ldots, Y_p)$.

2. The uncontrolled sampling plan is Sampford's plan for the above population and sample size $n' = 2$, so that each sample s corresponds to a pair of variables. Let $\mathcal{S}$ be the set of all possible samples of size two each, $\mathcal{S} = \{s_1, s_2, \cdots s_{\binom{p}{2}}\}$

3. If V defines the $p \times p$ covariance matrix of the observed variables and $\mathbf{e}_1$ is the eigenvector that corresponds to the largest eigenvalue of V, the inclusion probabilities π_i for implementing Sampford's sampling are defined according to the largest, in absolute magnitude, of each variable in $\mathbf{e}_1$. Second order probabilities π_{ij} can consequently be calculated. The *R* package *pps* has been used for the calculation of the inclusion probabilities.

4. Using inclusion probabilities π_i we calculate the probability of selection, $p(s)$, for every sample $s \in \mathcal{S}$ according to Sampford's plan.

5. Separate the set of possible samples (possible pairs of variables) into two categories: the desired pairs, S_D, and the non-desired ones, S_{ND}. The pairs selected from the Chao's sampling method will be the ones that belong in S_D, and $S_{ND} = \mathcal{S} - S_D$.

6. The control optimal sampling plan is defined as the sampling plan with sample probabilities $p_c(s)$, $s \in \mathcal{S}$ which result from the minimization problem of the objective function $\phi = \sum_{s \in S_{ND}} p(s)$, with constraints (i) $\sum_{s \ni i,j} p(s) = \pi_{ij}, (i < j = 1, 2, \cdots, p)$ and (ii) $p(s) \geq 0$ for all $s \in \mathcal{S}$.
   The linear programming technique is used for the minimization problem.

7. We draw n^* samples (pairs of variables) from $\mathcal{S}$ according to probabilities $p_c(s)$.

5 Simulation study on the performance of the proposed sampling methods and a comparison with the PML estimator

5.1 Simulation study set-up

A simulation study has been conducted to study and compare the performance of the proposed sampling schemes with the standard pairwise likelihood estimation (PML) where all pairs are taken into account. All the simulations we conducted involve exploratory and confirmatory factor analysis models for binary (simulations 1 to 4) and ordinal (simulations 5 and 6) observed variables. We also compare our proposed methods with simple random sampling (SRS), but the performance of SRS is rather poor in simulations with more than one factor and therefore we do not show any of the results here. A possible explanation is that the sample size specified by Chao's method is not sufficient for SRS for models with more than one factor.

Our experimental conditions vary with respect to the number of factors ($q = 1, 2$), the number of variables ($p = 15, 20, 30, 50$), the sample size of individuals ($n = 200, 500$ and $1,000$) and the number of item categories ($m = 2, 4$). The number of replications is set to $r = 500$ for each experimental condition. We fitted both one-factor model and a confirmatory factor analysis model with two correlated factors (following the rationale of a CFA discussed in the introduction, a number of loadings is constrained to be zero on each factor, e.g. $\lambda_{ij} = 0$ implies that the observed variable i does not measure the latent variable j or that the latent variable j does not have a direct influence on variable i). For simulations 1-4 the factor loadings of the assumed model have been selected from a *Uniform*$(0.6, 0.9)$ distribution and the thresholds for all variables are chosen to be -0.5. For simulations 5 and 6 the factor loadings are in the range from 0.6 to 0.9 and the thresholds for all variables are chosen to be $-1.25, 0$ and 1.25. In all experimental conditions the variances of the latent variables have been set to 1 for identification purposes.

Chao's sampling scheme is implemented on the estimated covariance matrix (cov) of pairs (6,7, 8 and 9) and on the similarity matrix computed using the RV-coefficient (RV) (10) as well as the other variants of the RV-coefficient such as the RV-adjusted (RV-adj) and the Lg coefficient (Lg). The estimated covariance matrix method is the most computationally demanding and therefore it has only been used in the small example with 15 variables. All those methods are discussed in Section 4. Furthermore, we also compare Chao's sampling scheme with the design-based sampling schemes of Sampford (S) and Controlled Sampford (C-S) discussed in Section 4.1. The design based schemes are implemented in combination with Chao's method and we expect them to have the best performance.

5.2 Performance criteria

We are interested in studying the properties of the estimators and their corresponding standard errors obtained under the proposed sampling schemes in terms of bias and mean

square error. When the number of parameters is large, instead of presenting results for each parameter estimate we compute the average relative bias (ARB) or percentage bias and the average root mean square error (ARMSE) across all estimated parameters. Previous simulations studies (Yang-Wallentin, Jöreskog, and Luo, 2010; Kaplan, 1989) considered relative bias values less than 5% as trivial bias, values between 5% and 10% as moderate bias and values greater than 10% as substantial bias. The average bias and average root mean squared error are given below.

$$ARB = 100 \times \frac{1}{r}\sum_{i=1}^{r}\frac{1}{k}\sum_{j=1}^{k}\left(\frac{\hat{\theta}_{ij}-\theta_j}{\theta_j}\right),$$

and

$$ARMSE = \frac{1}{r}\sum_{i=1}^{r}\sqrt{\frac{1}{k}\sum_{j=1}^{k}\left(\frac{\hat{\theta}_{ij}-\theta_j}{\theta_j}\right)^2},$$

where r here is the number of valid replicates, k is the total number of parameters, $\hat{\theta}_{ij}$ is the estimate of the j^{th} parameter or of its asymptotic standard error at the i^{th} valid replication, and θ_j is the corresponding true value. In the case of standard errors, where the true value θ_j is unknown, the standard deviation of parameter estimates across valid replications is used. However, for simulation studies 1, 2, and 6 which have up to 20 variables, we also report bias and mean square error for each estimated parameter in addition to the ARB and ARMSE. We only report the ARB and ARMSE for the estimated factor loadings, the estimated factor correlations and their corresponding standard errors because the bias for the estimated thresholds in all simulations is negligible and of similar magnitude across all methods including the PML applied to all pairs.

We also report the computational time efficiency across methods.

5.3 Results

In simulation studies 1 and 2, we generate data from a one-factor model with 15 observed binary variables and sample sizes of 200 and 500 respectively. The average relative bias and average root mean square error for the estimated factor loadings and their estimated standard errors under all proposed methods are included in Table 1. For the estimated factor loadings (top panel of Table 1), the ARB shows trivial bias (less than 5%) for all methods and for both sample sizes. The ARMSE decreases with the increase of the sample size and it is smaller for the PML, as expected, and of similar magnitude for the proposed methods at both sample sizes. The results for the estimated standard errors (bottom panel of Table 1) show trivial bias for most methods under $n = 200$ and trivial for all methods under $n = 500$. The RV and RV-adj were found to perform the same and both of them perform similar to the Lg method. The Cov method is computationally intensive and it is not recommended for larger models. Figures 1 and 2 give the bias and mean square error for each estimated factor loadings and its corresponding estimated asymptotic standard errors for PML, RV, S, and C-S and for sample sizes 200 and 500

Table 1: Simulations 1 and 2: Average Relative Bias (ARB) and Average Root MSE (ARMSE) for estimated factor loadings (top panel) and estimated asymptotic standard errors (bottom panel) under PML, RV-Chao (RV), RV-Chao adjusted (RV-adj), Lg-Chao (Lg), Sampford (S), Controlled Sampford (C-S) and Covariance-Chao (Cov). One-factor model, $p = 15, q = 1, n = 200$ and $n = 500, N^* = 105, r = 500$

	Average Relative Bias		**Average Root MSE**	
	$N = 200$	$N = 500$	$N = 200$	$N = 500$
Factor loadings				
PML	-0.51	-0.10	0.08	0.05
RV	-1.38	-0.49	0.11	0.07
RV-adj	-1.38	-0.49	0.11	0.07
Lg	-2.44	-0.91	0.12	0.07
S	-0.76	-0.17	0.11	0.07
C-S	-0.51	-0.12	0.11	0.07
Cov	-1.09	-0.15	0.11	0.07
Standard errors				
PML	0.81	0.002	0.14	0.10
RV	6.00	0.06	0.22	0.17
RV-adj	6.00	0.06	0.22	0.17
Lg	2.25	-0.78	0.22	0.18
S	1.31	-2.09	0.26	0.20
C-S	2.99	-3.13	0.25	0.21
Cov	4.71	1.70	0.22	0.18

respectively. In terms of bias the C-S and S methods were found to perform similarly, close to PML and better than the RV. The MSE of the estimated factor loadings were comparable among the three proposed sampling methods and PML gave the smallest as expected. With the increase of the sample size from 200 to 500 both the bias and the MSE decreased for all estimated factor loadings and their estimated standard errors. The average number of pairs selected across the 500 replications for the different sampling methods is given in rows 1 and 2 of Table 2. For simulations 1 and 2, PML uses all the 105 possible pairs where all the other methods select between 27 to 29 pairs.

In simulation study 3, we generate data from a one-factor model with 50 observed binary variables and sample size of 1,000. Since the RV behaved similarly to the RV-adj and Lg, and the C-S was among the best performing in Simulations 1 and 2, we only give the results here for the PML, the RV-Chao and the Controlled Sampford. In Table 3, the ARB shows trivial bias (less than 5%) for both the estimated factor loadings and the estimated standard errors for the RV and C-S methods. The ARMSE for the estimated factor loadings is comparable for all methods (PML, RV and C-S) and the PML gives the smallest ARMSE for the estimated standard errors. The average number of pairs selected across the 500 replications for the two different sampling methods is given in the third row of Table 2. PML uses all the 1,225 possible pairs whilst the other two methods select 99 pairs each.

In simulation study 4, we generate data from a two-factor confirmatory model with 30 binary observed variables and sample size of 500. Factor 1 loads on 20 variables and

Figure 1: Simulation 1: Bias and mean square error for the estimated factor loadings and their corresponding estimated asymptotic standard errors under PML, RV-Chao (RV), Sampford (S) and Controlled Sampford (C-S). One-factor model, $p = 15, q = 1, n = 200, N^* = 105, r = 500$.

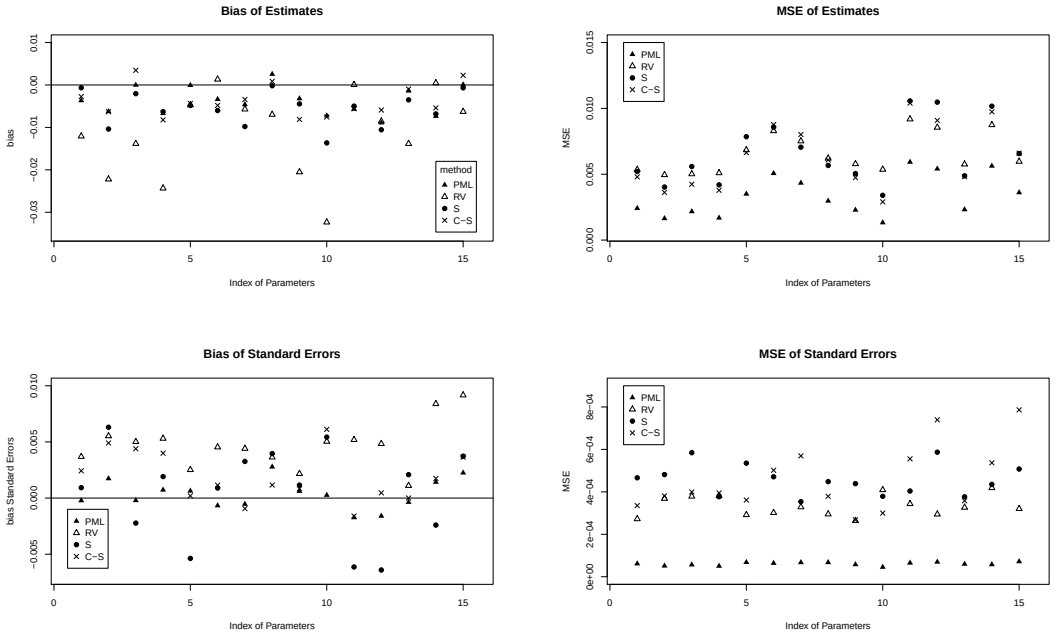

Table 2: Average sample size of pairs (rounded) selected across the 500 replications under RV-Chao (RV), RV-Chao adjusted (RV-adj), Lg-Chao (Lg), Sampford (S) and Controlled Sampford (C-S) for the six simulation scenarios.

	PML	RV	RV-adj	Lg	S	C-S	Cov
Simulation 1	105	27	27	27	27	27	29
Simulation 2	105	29	29	29	29	29	29
Simulation 3	1225	99				99	
Simulation 4	435	120				120	
Simulation 5	190	39				39	
Simulation 6	190	39				39	

Figure 2: Simulation 2: Bias and Mean Square Error for the estimated factor loadings and their corresponding estimated asymptotic standard errors under PML, RV-Chao (RV), Sampford (S), and Controlled Sampford (C-S)). One-factor model, $p = 15, q = 1, n = 500, N^* = 105, r = 500$.

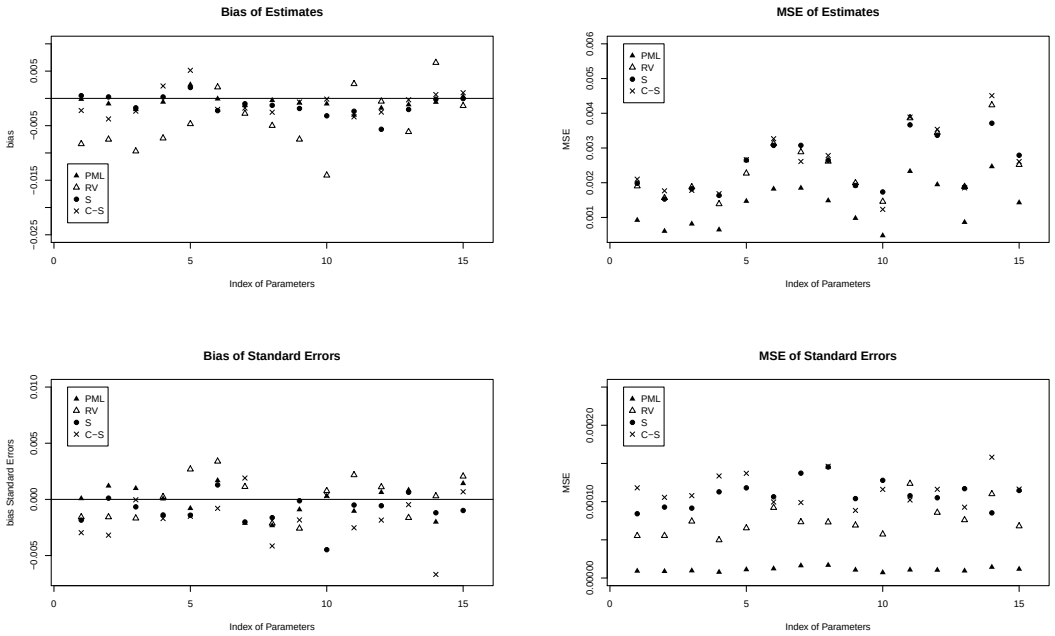

Table 3: Simulation 3: Average Relative Bias (ARB) and Average Root MSE (ARMSE) for factor loadings and for their standard errors under PML, RV-Chao (RV) and Controlled Sampford (C-S).One-factor model, $p = 50, q = 1, n = 1,000, N^* = 1,225, r = 500$

	Factor Loadings		Standard Errors	
	ARB	ARMSE	ARB	ARMSE
PML	0.06	0.04	-0.28	0.07
RV	-0.42	0.05	-1.02	0.15
C-S	0.09	0.05	-2.98	0.23

Table 4: Simulation 4: Average Relative Bias (ARB) and Average Root MSE (ARMSE) for factor loadings and for Standard Errors under PML, RV-Chao (RV) and Controlled Sampford (C-S). Two-factor confirmatory model, $p = 30, q = 2, \phi = 0.4, n = 500, N^* = 435, r = 500$

	Factor Loadings		Standard Errors	
	ARB	ARMSE	ARB	ARMSE
PML	-0.08	0.05	-0.95	0.10
RV	-0.56	0.09	4.66	0.23
C-S	-0.37	0.08	-0.22	0.26

factor 2 on 10 variables. The true correlation between the two factors is set equal to 0.4. The specified model is identified according to the three-indicator rule which is a sufficient but not necessary condition for identification. The rule requires that each factor has at least three indicators, one non-zero element per row of Λ, no correlated errors (Θ diagonal), and no restrictions on the covariance matrix Φ of the latent variables. For further information on model identifiability see Bollen (1989). Again, here we present the results for the PML, the RV-Chao and the Controlled Sampford. The ARB and ARMSE for the estimated loadings and the estimated asymptotic standard errors are given in Table 4. The C-S outperforms the RV in terms of ARB for both the loadings and the standard errors, and in terms of ARMSE both proposed methods perform similarly. Furthermore, the average estimated correlation between the two factors was found to be $0.415, 0.413$ and 0.4 for PML, RV and C-S respectively. The average number of pairs selected across the 500 replications for the two different sampling methods is given in the fourth row of Table 2. PML uses all the 435 possible pairs whilst the other two methods select 120 pairs.

In simulation studies 5 and 6, we generate data from a one-factor model with 20 observed ordinal variables with four response categories each and sample sizes of 200 and 500 respectively. We only show results for the PML, RV-Chao and Controlled Sampford. In Table 5, ARB shows trivial bias (less than 5%) for both the estimated factor loadings and the estimated standard errors for the PML, RV and C-S methods for both samples sizes. The ARMSE for the estimated factor loadings is comparable for all methods (PML, RV and C-S) and again here the PML gives the smallest ARMSE for the estimated standard errors and even smaller for $n = 500$. The average number of pairs selected across the 500 replications for the two different sampling methods is given in the fifth and sixth row of Table 2. PML uses all the 190 possible pairs whilst the other two methods select 39 pairs each. Figure 3 gives the bias and mean square error for each estimated factor loading and its corresponding estimated asymptotic standard errors for the three different methods and for sample size 500. For most parameters, C-S sampling shows less bias for the loadings and their estimated standard errors compared to RV but bigger MSE.

Table 5: Simulations 5 and 6: Average Relative Bias (ARB) and Average Root MSE (ARMSE) for factor loadings (top panel) and for their standard errors (bottom panel) under PML, RV-Chao (RV) and Controlled Sampford (C-S). One-factor model, ordinal data, $p = 20, q = 1, n = 200$ and $n = 500, N^* = 190, r = 500$

	Average Relative Bias		**Average Root MSE**	
	$N = 200$	$N = 500$	$N = 200$	$N = 500$
Factor loadings				
PML	0.18	-0.11	0.06	0.03
RV	-0.26	-0.33	0.07	0.05
C-S	0.22	-0.11	0.07	0.05
Standard errors				
PML	-2.26	0.76	0.13	0.09
RV	0.02	-1.65	0.19	0.16
C-S	-3.78	-1.12	0.22	0.19

Figure 3: Simulation 6: Bias and Mean Square Error for the estimated factor loadings and their corresponding estimated asymptotic standard errors under PML, RV-Chao (RV), and Controlled Sampford (C-S). One-factor model, ordinal variables, $p = 20, q = 1, n = 500, N^* = 190, r = 500$.

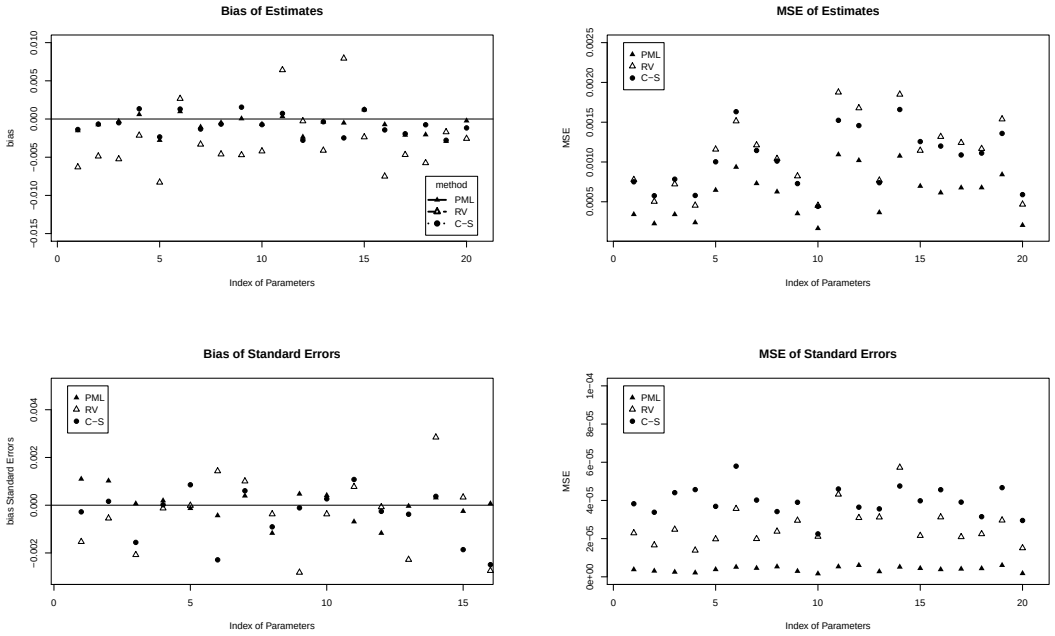

Table 6: Exact average time of estimation per replication (averaged across 500 replications) of proposed sampling methods and relative time compared under PML (in parentheses), RV-Chao (RV), RV-Chao adjusted (RV-adj), Lg-Chao (Lg), Sampford (S) and Controlled Sampford (C-S) for the six simulation scenarios.

	PML	RV	RV-adj	Lg	S	C-S	Cov
1	24.77	19.33	19.35	21.37	18.57	18.47	527.6
	(1.00)	(0.78)	(0.78)	(0.86)	(0.75)	(0.74)	(21.3)
2	17.80	14.97	14.97	16.54	15.09	15.13	338.5
	(1.00)	(0.84)	(0.84)	(0.93)	(0.85)	(0.85)	(19.02)
3	718.2	446.8				507.1	
	(1.00)	(0.62)				(0.71)	
4	108.9	83.65				84.71	
	(1.00)	(0.77)				(0.78)	
5	60.64	46.74				48.02	
	(1.00)	(0.77)				(0.79)	
6	38.69	28.44				28.37	
	(1.00)	(0.73)				(0.73)	

The main purpose for introducing the sampling of pairs in pairwise likelihood estimation for latent variable models is to improve the computational time for large-scale problems. Table 6 gives the average time in seconds per replication and relative time of estimation (in brackets) of all the proposed methods compared to PML in the six simulation studies. The computational time for each method has been averaged across the 500 replications for each experimental condition. The times for each replication include the estimation of the model parameters and their corresponding estimated asymptotic standard errors. For the proposed sampling methods, the times also include the time required to select the pairs. It is worth noting that the actual average computational time decrease for all simulations and all methods (including the PML) with the increase of sample size. The Chao method, which, uses the sampling covariance to compute covariances among pairs, is very intensive and should be avoided. The RV and the C-S have very similar performances and reduce overall computational time by approximately 20% to 35%. Table 6 also shows that computational gains of the RV method compared to PML depend on model complexity and sample size. In particular, in simulations 1 and 2 (one-factor model with 15 binary items) the relative gains of RV over PML decreased from 22% to 16% for sample sizes 200 and 500 respectively. On the contrary, in simulations 5 and 6 (one-factor model with 20 ordinal variables of four response categories) the corresponding relative gains increased from 23% to 27% with the increase of the sample size from 200 to 500. The simulations have shown that both the absolute and relative computational gains of the proposed sampling methods over PML are bigger for larger and more complex models (e.g. variables with more response categories, multidimensional factor models). Moreover, estimation of more complex models requires larger sample sizes and convergence can be achieved faster with the increase of the sample size. In those cases, the computational gains of using sampling methods such as the RV become even more evident.

The estimation time varies across the different methods, however, the number of iterations needed does not. For example, in Simulation 2, the average number of iterations (across the 500 simulations) required to achieve convergence are 188.0, 184.8, 186.8 and 185.0 for PML, RV, S and C-S respectively. For Simulation 4, the numbers are 379.3, 378.4 and 383.1 for PML, RV and C-S respectively. The time per iteration of the optimization procedure takes less time under the sampling methods, but the reduction in time is not proportional to the reduction in the number of bivariate components included in the likelihood.

6 A simulation study on the performance of the sampled pairwise on fit statistics and model selection criteria

Katsikatsou and Moustaki (2016) have developed pairwise likelihood ratio test statistics (PLRT) under the PML estimation for testing the overall fit of a model and for comparing nested models. They have shown that asymptotically the PLRT statistic both for the overall fit and for testing nested models, is a weighted sum of independent chi-squared variables. To determine the asymptotic distribution of PLRT the Satterthwaite approximation is used which leads to the mean-and-variance adjusted PLRT. Details can be found in their paper. They found that the type I error and power of the PLRT statistics are satisfactory for various experimental conditions and their performances improve with the sample size.

Furthermore, model selection criteria have been developed under the composite likelihood framework. Varin and Vidoni (2005) proposed the Akaike information criterion, AIC_{PL}:

$$AIC_{PL} = -pl\left(\hat{\boldsymbol{\theta}}; \mathbf{y}\right) + tr(\hat{J}(\hat{\boldsymbol{\theta}})\hat{H}^{-1}(\hat{\boldsymbol{\theta}})), \tag{11}$$

and, Gao and Song (2010), the Bayesian information criterion, BIC_{PL}:

$$BIC_{PL} = -2pl\left(\hat{\boldsymbol{\theta}}; \mathbf{y}\right) + tr(\hat{J}(\hat{\boldsymbol{\theta}})\hat{H}^{-1}(\hat{\boldsymbol{\theta}})) \times \log n\,, \tag{12}$$

where $\hat{\boldsymbol{\theta}}$ is the PML estimate under the hypothesized model, and $tr(\hat{J}(\hat{\boldsymbol{\theta}})\hat{H}^{-1}(\hat{\boldsymbol{\theta}}))$ defines the number of effective parameters. The model with the smallest AIC_{PL} or BIC_{PL} is selected.

Since the controlled-sampford method was found to outperform the other sampling methods, we conducted a small simulation to study the effect of the C-S pairwise likelihood on inference such as the type I error of the likelihood ratio test statistic for overall fit and on the AIC_{PL} and BIC_{PL} model selection criteria. The PLRT has been adjusted to take into account that a subset of all the possible pairs is selected. The simulation scenario is the same with simulation 2 (15 binary items, 1 factor) and sample sizes 500 and 1000. The number of replications is 500. Table 7 gives the empirical type I error rates for the PLRT for overall fit of the one-factor model under the PML and the C-S for two nominal significance levels. The results indicate that the PML gives empirical type I error rates close to the nominal ones for both sample sizes where the C-S improves

with the increase of the sample size from 500 to 1000. The confidence intervals are only given when the nominal significance level is not included.

Table 7: Empirical type I error rates for the overall-fit test statistic under PML and Controlled Sampford (C-S) for nominal significance levels 5% and 1%; in parenthesis 95% confidence intervals are provided only when the nominal value of type I error is not included. One-factor model, $p = 15, q = 1, n = 500$ and $n = 1000$, $N^* = 105, r = 500$.

Nominal level	$n = 500$		$n = 1000$	
	PML	C-S	PML	C-S
5%	0.064	0.082 (0.058,0.110)	0.042	0.054
1%	0.020	0.036 (0.0197,0.052)	0.010	0.016

For studying the performance of the AIC_{PL} and BIC_{PL}, we assumed the same two candidate models as in Katsikatsou and Moustaki (2016). More specifically, the data generator model (Model 1) is a confirmatory two-factor model with 20 items with four response categories, where the first 10 items have non-zero loadings in the first factor and items from 11 to 20 have non-zero loadings in factor 2. The two models are nested due to parameter constraints (some factor loadings are set equal to zero) The loadings are 0.3, 0.4, 0.4, 0.5, 0.5, 0.6, 0.6, 0.7, 0.8, 0.9 for both factors and the correlation between the factors is 0.4. Model 2, is also a confirmatory two-factor model with the only difference that items 11, 12, and 13 load also to the first factor. Again we investigate two sample sizes of 500 and 1000. The number of replications is set to 500 for both experiments.

Table 8: Rates of AIC_{PL} and BIC_{PL}, confirmatory factor model, $p = 20, q = 2, n = 500$ and $n = 1000, r = 500$.

	$n = 500$		$n = 1000$	
	PML	C-S	PML	C-S
AIC_{PL}	92.75%	78.88%	86.85%	79.34%
BIC_{PL}	100%	89.85%	100%	91.07%

As already reported in Katsikatsou and Moustaki (2016), for both sample sizes, BIC_{PL} selects the right model with 100% success and performs better than AIC_{PL}. The C-S method also shows higher rates for the BIC_{PL} but lower than the ones produced under the PML.

Further investigations are needed to explore in detail the effect of the sampling methods to inference under different experimental conditions that will vary the number of items, the sample size and the number of response categories. The small simulation we conducted here shows that the asymptotic results for the PLRT hold with the increase of the sample size and the BIC_{PL} has a a 92% success rate in selecting the true model when the sample size is 1000.

7 Comparison of PML and Controlled Sampford using empirical data

We use the same data that were analyzed in Katsikatsou et al. (2012) using PML. Originally the data were collected by Selnes and Sallis (2003) who aimed to study whether specific factors affect the learning capabilities of targeted customer-supplier relationships. The 18 variables analyzed here serve as indicators of four factors. The four factors, as named in their paper, are: collaborative commitment (ξ_1), internal complexity (ξ_2), relational trust (ξ_3), and environmental uncertainty (ξ_4). The observed variables used to measure each factor are given in the Appendix. All indicators were measured on a seven-point scale; with 1 referring to "strongly disagree" or "low" and 7 to "strongly agree" or "high" depending on the form of the question. The sample size is 286 after listwise deletion. This is a confirmatory factor analysis model with no cross-loadings (i.e. that each item is loaded on just one of the four factors). The correlation matrix of the four latent variables is unrestricted. It should be noted that pairwise likelihood does not guarantee a positive definite matrix and that this should be checked or a positive definite constraint should be added to the numerical optimization. Along with the thresholds, which are six for each variable, there are a total of 132 free parameters to be estimated (6×18=108 thresholds, 18 factor loadings and 6 factor correlations). Since the Controlled Sampford was found to be the best method and similar to RV in the simulations we only compare it with the PML, which takes into account all the 135 possible pairs. The sample of pairs selected based on the Chao method is 57. The 'internal complexity' factor has only three indicators. For that factor only, we used two pairs out of three because it has a small number of indicators and it is also the factor with the smallest correlation with the other three factors. The estimated factor loadings and factor correlations together with their estimated asymptotic standard errors for the PML and C-S are given in Table 9. Overall, the PML and C-S give similar results for the loadings and the correlations except for the estimated correlation between the factors 'collaborative commitment' and 'internal complexity', which has been overestimated by the C-S method. The estimated standard errors are higher for the C-S method. The computing time using the C-S method is reduced by 32% compared to the corresponding time using PML.

8 Conclusions

In this paper, we propose a weighted pairwise likelihood estimation for estimating the parameters of latent variable models for categorical observed variables in which weights are computed using a sampling scheme. We studied both exploratory factor analysis models with 15, 20 and 50 binary and ordinal observed variables and the confirmatory factor analysis model with two factors and 30 observed binary variables. The weights are either 1 or 0 and are decided by employing sampling of pairs of variables from the population of all possible pairs. We propose three sampling methods, of which

Table 9: Estimated factor loadings, correlations and standard errors (in brackets) under PML and C-S method, Relationship learning data.

Loadings	$\hat{\lambda}_{ij}$	PML	C-S
	cc1	0.878 (0.025)	0.964 (0.058)
	cc2	0.900 (0.018)	0.942 (0.025)
	cc3	0.894 (0.018)	0.888 (0.021)
	cc4	0.909 (0.016)	0.962 (0.016)
	cc5	0.881 (0.021)	0.867 (0.025)
	ic1	0.597 (0.108)	0.619 (0.176)
	ic2	0.817 (0.082)	0.711 (0.175)
	ic3	0.771 (0.084)	0.969 (0.226)
	rt1	0.840 (0.025)	0.846 (0.031)
	rt2	0.848 (0.027)	0.854 (0.034)
	rt3	0.875 (0.025)	0.852 (0.037)
	rt4	0.912 (0.017)	0.890 (0.035)
	rt5	0.864 (0.022)	0.864 (0.050)
	eu1	0.779 (0.035)	0.719 (0.050)
	eu2	0.841 (0.030)	0.855 (0.030)
	eu3	0.769 (0.042)	0.807 (0.043)
	eu4	0.728 (0.045)	0.679 (0.054)
	eu5	0.732 (0.044)	0.610 (0.057)
Correlation	$\phi_{cc,ic}$	0.229 (0.093)	0.456 (0.158)
	$\phi_{cc,rt}$	0.652 (0.049)	0.648 (0.053)
	$\phi_{cc,eu}$	0.681 (0.052)	0.667 (0.053)
	$\phi_{ic,rt}$	0.117 (0.085)	0.118 (0.088)
	$\phi_{ic,eu}$	0.227 (0.092)	0.233 (0.083)
	$\phi_{rt,eu}$	0.655 (0.053)	0.709 (0.048)

one is model-based sampling and two are design-based sampling schemes but they also incorporate the information from the model-based sampling procedure. Various methods have been investigated for computing the covariance matrix of the pairs required in the model-based sampling scheme. The simulation studies have shown that the model-based sampling, and in particular Chao's method with RV and the Controlled Sampford method combined with the RV method, behave very similar, to the standard PML, which utilizes all pairwise likelihoods. More specifically, in all simulations for the estimated factor loadings, the Controlled Sampford is found to have the smallest bias and the MSE is very similar for both methods. For the estimated standard errors, the two methods gave acceptable bias but neither of the two methods behaved consistently better than the other. Again the MSE was of similar magnitude in all simulations for both methods. The simulations have shown that applying the PML on the selected subset of pairs gives estimated parameters and standard errors with trivial bias. However, the MSE's under the sampling methods compared to the ones obtained under PML with all pairs, are larger since the proposed methods are using a subset of the whole population of pairs and efficiency will be reduced. The simulation studies have shown that the proposed sampling schemes improve computational time by at least 20%. The simulation studies on the performance of the PLRT for overall fit and on the AIC_{PL} and BIC_{PL} show that the asymptotic results hold with the increase of the sample size and that the BIC_{PL} has a large rate of selecting the true model.

To the best of our knowledge, the paper is the first attempt to introduce sampling

methods in composite likelihood estimation. The proposed methods are easy to implement and have been found to work very similarly to the standard PML. Further simulations are needed for more complex SEM (e.g. multigroup models, longitudinal models). Ways to also improve the efficiency of the estimates need to be investigated within the sampling framework.

Appendix: The indicators for the Relationship Learning data

Collaborative Commitment

cc1 To what degree do you discuss company goals with the other party in this relationship?

cc2 To what degree are these goals developed through joint analysis of potentials?

cc3 To what degree are these goals formalized in a joint agreement or contract?

cc4 To what degree are these goals implemented in day-to-day work?

cc5 To what degree have you developed measures that capture performance related to these goals?

Internal Complexity

ic1 The products we exchange are generally very complex.

ic2 There are many operating units involved from both organizations.

ic3 There are many contract points between different departments or professions between the two organizations.

Relational Trust

rt1 I believe the other organization will respond with understanding in the event of problems.

rt2 I trust that the other organization is able to fulfill contractual agreements.

rt3 We trust that the other organization is competent at what they are doing.

rt4 There is a general agreement in my organization that the other organization is trustworthy.

rt5 There is a general agreement in my organization that the contact people on the other organization are trustworthy.

Environmental Uncertainty

eu1 End-users needs and preferences change rapidly in our industry.

eu2 The competitors in our industry frequently make aggressive moves to capture market share.

eu3 Crises have caused some of our competitors to shut down or radically change the way they operate.

eu4 It is very difficult to forecast where the technology will be in the next 2-3 years in our industry.

eu5 In recent years, a large number of new product ideas have been made possible through technological breakthroughs in our industry.

References

Bartholomew, D., F. Steele, I. Moustaki, and J. Galbraith (2008). *Analysis of Multivariate Social Science Data* (2nd ed.). Chapman and Hall/CRC.

Bellio, R. and C. Varin (2005). A pairwise likelihood approach to generalized linear models with crossed random effects. *Statistical Modelling 5*, 217–227.

Besag, J. E. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society, Series B 36*, 192–236.

Bolfarine, H. and S. Zacks (1992). *Prediction Theory for Finite Population*. New York: Springer Verlag.

Bollen, K. A. (1989). *Structural Equations With Latent Variables*. New York: Wiley and Sons.

Chan, W. and P. Bentler (1998). Covariance structure analysis of ordinal ipsative data. *Psychometrika 63*, 369–399.

Chao, C.-T. (2004). Selection of sampling units under a correlated population based on the eigensystem of the population matrix. *Environmetrics 15*(8), 757–775.

Cox, D. R. and N. Reid (2004). A note on pseudolikelihood constructed from marginal densities. *Biometrika 91*, 729–737.

de Leon, A. R. (2005). Pairwise likelihood approach to grouped continuous model and its extension. *Statistics & Probability Letters 75*, 49–57.

Escoufier, Y. (1973). Le traitement des variables vectorielles. *Biometrics 29*(4), 751–760.

Fieuws, S. and G. Verbeke (2006). Pairwise fitting of mixed models for the joint modeling of multivariate longitudinal profiles. *Biometrics 62*, 424–431.

Gao, X. and P. X. Song (2010). Composite likelihood Bayesian information criteria for model selection in high dimensional data. *Journal of the American Statistical Association 105*(492), 1531–1540.

Haziza, D., F. Mecatti, and J. Rao (2008). Evaluation of some aproximate variance estimators under the Rao-Sampford unequal probability sampling design. *Metron-International Journal of Statistics LXVI*, 91–108.

Heagerty, P. and S. Lele (1998a). A composite likelihood approach to binary spatial data. *Journal of the American Statistical Association 93*, 1099–1111.

Heagerty, P. J. and S. Lele (1998b). A composite likelihood approach to binary spatial data. *Journal of the American Statistical Association 93*, 1099–1111.

Henderson, R. and S. Shimakura (2003). A serially correlated gamma frailty model for longitudinal count data. *Biometrika 90*, 355–366.

Joe, H. and Y. Lee (2009). On weighting of bivariate margins in pairwise likelihood. *Journal of Multivariate Analysis 100*, 670–685.

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. (1994). On the estimation of polychoric correlations and their asymptotic covariance matrix. *Psychometrika 59*, 381–389.

Jöreskog, K. G. and I. Moustaki (2001). Factor analysis of ordinal variables: a comparison of three approaches. *Multivariate Behavioral Research 36*, 347–387.

Kaplan, D. (1989). A study of the sampling variability and z-values of parameter estimates from misspecified structural equation models. *Multivariate Behavioral Research 24*, 41–57.

Katsikatsou, M. (2013). *Composite Likelihood Estimation for Latent Variable Models with Ordinal and Continuous or Ranking Variables.* Ph. D. thesis, Uppsala University.

Katsikatsou, M. and I. Moustaki (2016). Pairwise likelihood ratio tests and model selection criteria for structural equation models with ordinal variables. *Psychometrika 81*, 1046–1068.

Katsikatsou, M., I. Moustaki, F. Yang-Wallentin, and K. G. Jöreskog (2012). Pairwise likelihood estimation for factor analysis models with ordinal data. *Computational Statistics and Data Analysis 56*, 4243–4258.

Lee, S. Y., W. Y. Poon, and P. M. Bentler (1990). Full maximum likelihood analysis of structural equation models with polytomous variables. *Statistics and Probability letters 9*, 91–97.

Lee, S. Y., W. Y. Poon, and P. M. Bentler (1992). Structural equation models with continuous and polytomous variables. *Psychometrika 57*, 89–105.

Lindsay, B. (1988). Composite likelihood methods. In N. U. Prabhu (Ed.), *Statistical Inference from Stochastic Processes*, pp. 221–239. Providence, RI: American Mathematical Society.

Lingoes, J. and P. Schönemann (1974). Alternative measures of fit for the Schnemann-Carrol matrix fitting algorithm. *Psychometrika 39*, 423–427.

Liu, J. (2007). *Multivariate ordinal data analysis with pairwise likelihood and its extension to SEM.* Ph. D. thesis, University of California.

Mayer, C-D. Lorent, J. and G. Horgan (2011). Exploratory analysis of multiple omics datasets using the adjusted RV coefficient. *Statistical Applications in Genetics and Molecular Biology 10*, Article 14.

Muthén, B. (1984). A general structural model with dichotomous, ordered categorical and continuous latent variable indicators. *Psychometrika 49*(1), 115–132.

Pace, L., A. Salvan, and N. Sartori (2011). Adjusting composite likelihood ratio statistics. *Statistica Sinica 21*, 129–148.

Rao, J. and A. Nigam (1990). Optimal controlled sampling designs. *Biometrika 77*, 807–814.

Rao, J. and A. Nigam (1992). Optimal controlled sampling: a unified approach. *International Statistical Review 60*, 89–98.

Robert, P. and Y. Escoufier (1976). A unifying tool for linear multivariate statistical methods: The RV-coefficient. *Journal of the Royal Statistical Society Series C (Applied Statistics) 25*(3), 257–265.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software 48 (2)*, 1–36.

Rosseel, Y., D. Oberski, J. Byrnes, L. Vanbrabant, V. Savalei, and E. Merkle (2012,

September). *Package lavaan*.

Sampford, M. R. (1967). On sampling without replacement with unequal probabilities. *Biometrika 54*, 499–513.

Sarndal, C.-E. (1978). Design-based and model-based inference in survey sampling. *Scandinavian Journal of Statistics 5*(1), 27–52.

Selnes, F. and J. Sallis (2003). Promoting relationship learning. *Journal of Marketing 67*, 80–95.

Varin, C. (2008). On composite marginal likelihoods. *Advances in Statistical Analysis 92*, 1–28.

Varin, C. and C. Czado (2010). A mixed autoregressive probit model for ordinal longitudinal data. *Biostatistics 11*, 127–138.

Varin, C., G. Host, and O. Skare (2005). Pairwise likelihood inference in spatial generalized linear mixed models. *Computational Statistics and Data Analysis 49*, 1173–1191.

Varin, C., N. Reid, and D. Firth (2011). An overview of composite likelihood methods. *Statistica Sinica 21*, 5–42.

Varin, C. and P. Vidoni (2005). A note on composite likelihood inference and model selection. *Biometrika 92*, 519–528.

Varin, C. and P. Vidoni (2006). Pairwise likelihood inference for ordinal categorical time series. *Computational Statistics and Data Analysis 51*, 2365–2373.

Vasdekis, V., S. Cagnone, and I. Moustaki (2012). A pairwise likelihood inference in latent variable models for ordinal longitudinal responses. *Psychometrika 77*, 425–441.

Vasdekis, V., D. Rizopoulos, and I. Moustaki (2014). Weighted pairwise likelihood estimation for a general class of random effects models. *Biostatistics 15*, 677–689.

Yang-Wallentin, F., K. G. Jöreskog, and H. Luo (2010). Confirmatory factor analysis of ordinal variables with misspecified models. *Structural Equation Modeling 17*, 392–423.

Zhao, Y. and H. Joe (2005). Composite likelihood estimation in multivariate data analysis. *The Canadian Journal of Statistics 33*, 335–356.