Estimation of Models in a Rasch Family for Polytomous Items and Multiple Latent Variables

Carolyn J. Anderson
University of Illinois

Zhushan Li
University of Illinois

Jeroen K. Vermunt
Tilburg University

Abstract

The Rasch family of models considered in this paper includes models for polytomous items and multiple correlated latent traits, as well as for dichotomous items and a single latent variable. An R package is described that computes estimates of parameters and robust standard errors of a class of log-linear-by-linear association (LLLA) models, which are derived from a Rasch family of models. The LLLA models are special cases of log-linear models with bivariate interactions. Maximum likelihood estimation of LLLA models in this form is limited to relatively small problems; however, pseudo-likelihood estimation overcomes this limitation. Maximizing the pseudo-likelihood function is achieved by maximizing the likelihood of a single conditional multinomial logistic regression model. The parameter estimates are asymptotically normal and consistent. Based on our simulation studies, the pseudo-likelihood and maximum likelihood estimates of the parameters of LLLA models are nearly identical and the loss of efficiency is negligible. Recovery of parameters of Rasch models fit to simulated data is excellent.

Keywords: pseudo-likelihood estimation, log-linear-by-linear association models, logistic regression, multinomial logistic regression, conditionally specified models, R.

1. Introduction

Multiple choice items on tests, questionnaires and surveys often have more than two response options (e.g., "strongly agree," "agree," "disagree," "strongly disagree"). Responses to items may reflect one or more underlying or unobserved trait, ability, or attitude. The goal in modeling responses to items may be to study the items for use on future measurement instruments, measure individuals' values on latent trait(s), or study the relationship between latent variables. Although our emphasis is on estimation of parameters, all three goals can be achieved for a Rasch family of models using the package that we present in this paper for the R programing environment (R Development Core Team 2007) .

The application of multidimensional item response theory models has been hindered by estimation difficulties. Marginal maximum likelihood estimation of a model involves numerically integrating the latent variable out of the model. This method becomes problematic for multiple latent variables, because it requires multiple numerical integration. For example, item response theory models can be estimated using nonlinear mixed models software, which is advantageous because of the flexibility to fit a variety of models (Rijmen, Tuerlinckx, de Boeck, and Kuppens 2003; de Boeck and Wilson 2004; Shieu, Chen, Su, and Wang 2005). Unfortunately fitting IRT models as non-linear mixed models using marginal maximum likelihood may become problematic because computation time increases exponentially with the number of dimensions (Wang, Chen, and Cheng 2004; Glas 2005). Markov chain Monte Carlo (MCMC) is another method that has been used to obtain estimates of parameters of multidimensional item response theory models, but this method is difficult to adjust to different models and specific problems. The computations in MCMC can be extremely time consuming and require relatively sophisticated programing skills. In this paper, we propose estimating multidimensional (or unidimensional) Rasch models for polytomous (or binary) items using a model for observed responses derived by conditionally specifying the response models using rest-scores (i.e., test total minus one item). A limitation of this approach is that current maximum likelihood estimation algorithms are limited to small numbers of items because the problem size increases exponentially with the number of items; however, this is a problem with an easily implemented solution that is theoretically sound, numerically well behaved, computationally fast and very flexible.

The package that we present for the R programing environment (R Development Core Team 2007) fits Rasch models to data as log-linear-by-linear association (LLLA) models, which are models for observed response patterns. The LLLA models that we derive from a Rasch family of models are extensions of log-linear models with bi-variate interactions. The interaction terms in log-linear models are replaced by products of category scores (e.g., integers) and a parameter that represents the strength of the relationship between the items. The cases of LLLA models we use are special cases and extensions of the uniform association model for ordinal data (Goodman 1985; Agresti 2002).

Log-linear-by-linear association and closely related models have been derived from Rasch models in at least four different ways (Anderson and Yu 2007). In the mid-1960's, Rasch proposed a latent structure model for polytomous items that was equivalent to Goodman's (Goodman 1979, 1985) row-column (RC) association model, a model related to LLLA models (Andersen 1995). A second way uses the fact that models in the Rasch family are equivalent to quasi-symmetric models (Tjur 1982; Agresti 2002). Log-linear-by-linear association models are limited information quasi-symmetric models (limited in the sense that only 2-way associations are modeled). The third derivation, which is based on item response theory, was made by Holland (1990) using the Dutch Identity and two critical assumptions. The assumptions made by Holland are equivalent to those made by Anderson and Vermunt (2000) (Anderson and Böckenholt 2000; Anderson 2002) who started with statistical graphical models for discrete and continuous data (Anderson and Yu 2007). The fourth derivation of LLLA models uses the fact that LLLA models are uniquely implied by conditionally specifying models for each item conditioning on rest-scores, which are test totals less the score for an item (Anderson and Yu 2007). In this paper, we generalized the conditional specification method presented in Anderson and Yu (2007) who only derived models for binary items and uni-dimensional models. Generalizing the conditional specification approach is advantageous because fewer distribu-

tional assumptions are required, justification exists for using rest-scores, and the conditional models themselves are used in the solution to an estimation problem for LLLA models.

Maximum likelihood estimation (MLE) of LLLA models for small problems is straightforward and can be done using standard software for Poisson (log-linear) regression models. For estimation algorithms such as Newton-Raphson, the MLE of LLLA models requires iteratively computing probabilities for all possible response patterns. The number of possible response patterns increases exponentially with the addition of items and response options per item, which makes MLE prohibitive for moderate to large numbers of items. For large problems, we propose using pseudo-likelihood estimation.

Pseudo-likelihood estimation (PLE) was first proposed by Besag (1974) to solve the problem of fitting spatial interaction models, which have complex normalization constants. Pseudo-likelihood estimation has solved estimation problems in a number of settings including models for social networks (Strauss and Ikeda 1990; Wasserman and Pattison 1990), multivariate clustered data (Geys, Molenberghs, and Ryan 1999; Molenberghs and Verbeke 2005), computer network tomography (Liang and Yu 2003), learning language comprehension (Johnson and Riezler 2002), and nonlinear measurement error models (Huwang and Huwang 2002). Pseudo-likelihood estimation has even been used to estimate parameters of Rasch models; however, its use has been limited to unidimensional models for pairs of binary items (Arnold and Strauss 1991; Zwinderman 1995) and more recently for a set of dichotomous items (Smit and Kelderman 2000). Our approach goes beyond these applications of PLE because we can accommodate models for multiple latent variables and polytomous items. The expanded use of PLE for estimation of a general class of Rasch models is possible because of the derivation of these models as specific LLLA models for observed responses. Furthermore, we use results from Geys (Geys 1999; Geys *et al.* 1999; Aerts, Geys, Molenberghs, and Ryan 2002) to show that the resulting parameter estimates are asymptotically normal and consistent, provide robust standard errors, and permit statistical tests for the model parameters.

Although the underlying PLE theory is complex, its concrete implementation is rather simple: it involves fitting a single conditional logistic regression model to an appropriately formatted data set. The method can be implemented in any statistical program that fits Poisson (log-linear) models; however, we present an easy to use R package, **plRasch**, that sets up the data, estimates the parameters, and computes the covariance matrix and standard errors for the estimated parameters. The user can choose between MLE (for small problems) and PLE (for small or large problems). In the case of PLE, robust standard errors are computed for the model parameters. The input consists of three matrices: a standard individual-by-item data matrix, a matrix indicating which items load on which latent variable(s), and a matrix indicating which latent variables are correlated. The R code and examples are available at `http://www.ed.uiuc.edu/faculty/cja/plRasch/`[1] and in the near future at the R project Web site at `http://www.R-project.org/`.

The remainder of the paper is structured as follows. In Section 2.2, we start with a review of the family of Rasch models for multicategory responses and multiple latent variables and subsequently show the connection between Rasch and LLLA models. In Section 3, we discuss pseudo-likelihood estimation and how it applies to the estimation of LLLA models. In Section 4, we describe the **plRasch** package and illustrate how to fit models for both MLE and PLE of LLLA (Rasch) models. In Section 5, results from our simulation study are re-

---

[1]Examples using SAS are also available from this Web site.

ported. In Section 6, we conclude with a discussion of possible modifications for fitting more restrictive and more general models, including those with covariates.

2. Rasch models as log-linear-by-linear models

In Section 2.1, we introduce a Rasch family of models for a single latent trait and subsequently show their extension to the multidimensional case. In Section 2.2, we derive the corresponding log-linear-by-linear association models; that is, models for observed response patterns.

2.1. Rasch family of models

Our main reference to the Rasch family of models for polytomous items is Andersen (1995). Out of simplicity but not necessity, we assume for now that all items have the same number of categories[2]. Let X_{vi} represent the response of individual v on item i where response options are represented by $h = 0, 1, \ldots, m$. Deviating slightly from the notation used by Andersen (1995), we let $P(X_{vi} = h|\theta_v)$ equal the probability that individual v selects response option h on item i given θ_v where θ_v equals individual v's value on latent trait Θ. The general uni-dimensional Rasch model specified for $P(X_{vi} = h|\theta_v)$ is a multinomial logistic regression where the predictor variable is unobserved. The unidimensional Rasch model for a polytomous item is

$$P(X_{vi} = h|\theta_v) = \frac{\exp[w_h\theta_v + \beta_{ih}]}{\sum_{l=0}^{m} \exp[w_l\theta_v + \beta_{il}]}, \tag{1}$$

where w_h is a known or assumed category weight or "score"[3] for response h, and β_{ih} is a marginal or "location" effect parameter for response h on item i. The sum in the denominator ensures that for individual v the sum of probabilities $P(X_{vi} = h|\theta_v)$ over all response options to item i equals 1. A property common to all Rasch models is that the slope or discrimination parameters (i.e., w_h's) are the same over all items. The marginal effect parameters β_{ih} depend on both the response option and the particular item.

For dichotomous items, the category weights or scores are usually set to $w_0 = 0$ and $w_1 = 1$, and for identification, $\beta_{i0} = 0$. Using these values yields the standard Rasch or one parameter logistic model for binary data,

$$P(X_{vi} = 1|\theta_v) = \frac{\exp[\theta_v + \beta_{i1}]}{1 + \exp[\theta_v + \beta_{i1}]}, \tag{2}$$

where β_{i1} is an "easiness" parameter. As values of θ_v increase, the probability of selecting response option 1 monotonically increases. For a more complete description of this model see Fischer and Molenaar (1995) or Embretson and Reise (2000).

For unidimensional models, the R package presented in this paper estimates the parameters of equation (1); however, users can modify the R function that sets up the design matrix to fit restricted or extended variants of (1). For example, Master's partial credit model is a simple reparameterization of (1) where $\beta_{ih} = \sum_{l=0}^{h} \alpha_{il}$ (Andersen 1995). Models for the β_{ih}'s can be specified, which lead to more restrictive models, such as Andrich's rating scale model

---

[2]Later in the paper, we discuss how items can be dealt with that have different numbers of response options.
[3]In the literature on log-linear-by-linear association models, which is used later in this paper, the w's are referred to as category scores.

([Andersen 1995](#)). [Mair and Hatzinger](#) ([2007](#)) in this issue discuss a number of restricted and extended forms of uni-dimensional Rasch models, including those that incorporate covariates.

Responses to items on a test may depend on multiple latent variables; therefore, we extend model (1) to the multidimensional case. Let θ_{vp} represent the value on latent variable p for individual v, and r equal the number of latent variables. To represent the multidimensional model, subscripts for items and latent variables are added to the category scores such that w_{ihp} is the known or assumed weight for response option h to item i on latent trait p. If item i is not directly related to latent variable p, then $w_{ihp} = 0$ for all h. If item i is directly related to latent variable p, then $w_{ihp} \neq 0$. The multidimensional model is

$$P(X_{vi} = h|\theta_{v1}, \ldots, \theta_{vr}) = \frac{\exp\left[\sum_{p=1}^{r} w_{ihp}\theta_{vp} + \beta_{ih}\right]}{\sum_{l=0}^{m} \exp\left[\sum_{p=1}^{r} w_{ilp}\theta_{vp} + \beta_{il}\right]}. \tag{3}$$

Two graphs of possible structures for two latent variables and four items are given in Figure 1. The squares represent the items and the circles represent the latent traits. The absence of a line between two variables indicates that the variables are independent given other variables in the graph, and the presence of a line indicates that two variables may be dependent given other variables in the graph. The structure represented by graph (a) is often referred to as "simple structure" (i.e., each item is an indicator of only one latent variable). The items X_1 and X_2 are directly related to latent variable Θ_1, the items X_3 and X_4 are directly related to Θ_2, all of the items are independent given the two latent variables, and the two latent variables are dependent[4]. Since there is no line connecting X_1 or X_2 to Θ_2, the category weights $w_{1h2} = w_{2h2} = 0$, and since there is no line connecting X_3 or X_4 to Θ_1, $w_{3h1} = w_{4h1} = 0$. Given the assumed weights w_{ihp}'s, the parameter ϕ_{12} reflects the strength of the relationship between the two latent variables, and the parameters ϕ_{11} and ϕ_{22} are related to the scale of Θ_1 and Θ_2, respectively. Graph (b) in Figure 1 represents a structure that is slightly more complex where item X_3 is directly related to both Θ_1 and Θ_2. Both types of multidimensional structures can be fit using the R function described in this paper.

From the graph in Figure 1 (a), we expect the (observed) response to items X_1 and X_2 to be associated because they are both directly related to Θ_1, and X_3 and X_4 should be associated because they are both directly related to Θ_2. If Θ_1 and Θ_2 are dependent, then the responses to X_1 and X_2 should also be associated with the responses to X_3 and X_4. A similar pattern of associations is found for graph Figure 1 (b). The expected pattern of observed associations indicates that models for the distribution of response patterns should include terms that represent not only the relationship between items that are indicators of the latent trait to which they are directly associated but also terms that represent associations due to the relationship between latent traits. The corresponding model for the observed response patterns is the subject of the next section.

2.2. Rasch models as LLLA models

Our goal here is to derive a model for the probability of manifest response patterns on the I items assuming an underlying Rasch model; that is, we want a model for $P(\boldsymbol{x})$ where $\boldsymbol{x}$

---

[4]Strictly speaking, the presence of a line indicates that the variables may be dependent given all other variables in the graph. Whether a dependency exists between two variables depends on the value of the parameter relating the two variables.

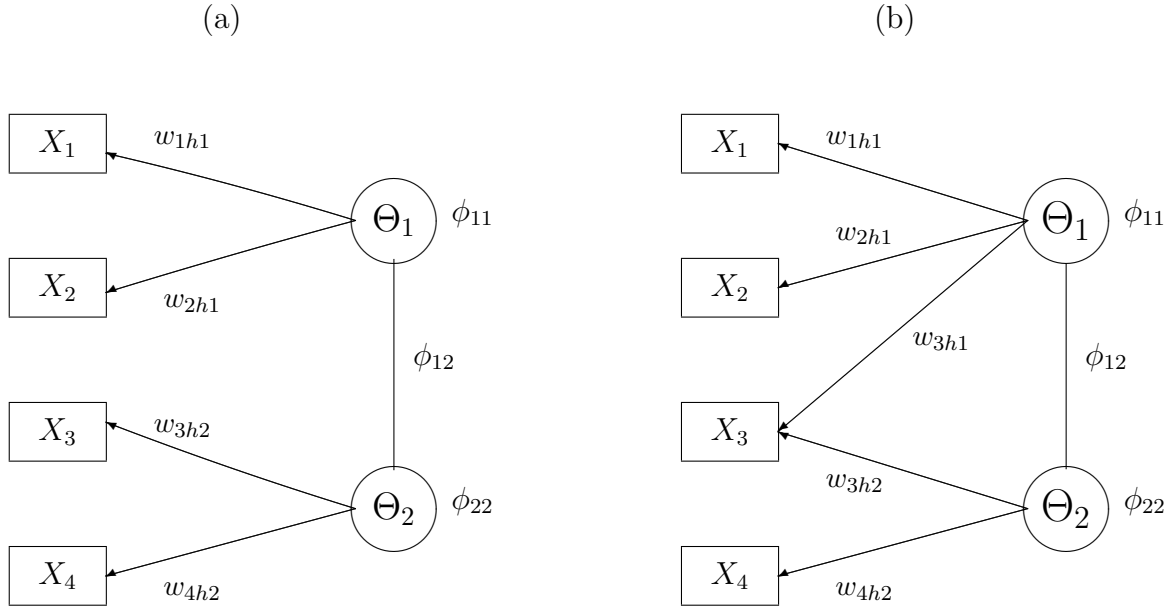(a)   (b)

Figure 1: Graphs that represent two general types of multidimensional structures for four items and two correlated latent traits.

represents a response pattern. Our approach to this problem is to specify a model for each item conditional on the remaining ones (i.e., $P(X_{vi} = h|x_{vk}, k \neq i)$) that is implied by an underlying Rasch model. The joint distribution that is compatible or consistent with such a set of conditionally specified models gives us the model for $P(\boldsymbol{x})$. We introduce this method using the model represented by the graph in Figure 1 (b) and then generalize it to any number of items, latent variables and latent variable structures.

Consider item X_1 in Figure 1 (b). We use the fact that the sum of the category weights w_{ihp} over items on latent trait p is sufficient for θ_p (Andersen 1995; Heinen 1993, 1996; Bartholomew and Knott 1999). When specifying a model for a response function such as $P(X_{v1} = h|x_{v2}, x_{v3}, x_{v4})$, we are conditioning on data rather than the latent variables as in equation (3). As an estimate for θ_{vp}, we will use the sum of the w_{ihp}'s excluding the item for which the model is specified. These sums are rest-scores and equal the sum over all items except one.

Justification and precedence for using rest-scores when modeling response functions for items come from the literature on both classical test theory and item response theory. With respect to classical test theory, the correlation between the item and the test total is artificially inflated when the item is included in the test total (particularly for short tests); therefore, some have argued that correlations in an item analysis should be computed between an item and the rest-score. In the item response literature, justification for the use of rest-scores for binary items comes from nonparametric uni-dimensional monotone item response theory (Junker 1993; Junker and Sijtsma 2000). Junker (1993) showed that item response functions will be monotone when rest-scores are used; however, item response functions can be artificially non-monotonic when test totals are used. Further precedence in item response theory for using

rest-scores comes from studies on the shape of item response functions, model fit, differential item functioning, and identification of incorrectly coded responses (Junker and Sijtsma 2000).

In the case of correlated latent variables, our estimates of the latent variables can be improved by using rest-scores and test totals from items that measure the correlated latent variables. Using collateral information from items that are indicators of other latent variables can lead to an increase in efficiency when estimating θ_{vp} (Wang *et al.* 2004; de la Torre and Patz 2005). For example, when specifying a model for X_1 represented by the graph in Figure 1 (b), a rest-score for θ_{v1} would use only the responses to X_2 and X_3; however, the responses to the items X_3 and X_4 provide information about θ_2, which is correlated to θ_1. When modeling the responses to item X_1, we take as our estimate of θ_1 for individual v

$$\tilde{\theta}_{v1,-1} = \phi_{11}(w_{2j_v1} + w_{3j_v1}) + \phi_{12}(w_{3j_v2} + w_{4j_v2}), \tag{4}$$

where $\tilde{\theta}_{v1,-1}$ is the estimate of θ_1 for individual v that excludes item X_1, $(w_{2j_v1} + w_{3j_v1})$ is the rest-score for individual v of the items that are direct indicators of θ_1, $(w_{3j_v2} + w_{4j_v2})$ is individual v's test-total of the items that are direct indicators of θ_2, and ϕ_{11} and ϕ_{12} are weights for the relative contributions of the rest-score and test-total toward the estimate of θ_1. Note that we have put the sub-script v on the j's to indicate that individual v selected response option j.

Although we will not make any explicit assumptions regarding the distribution of the latent variables, implicitly we assume that their distribution is such that there are only two-way interactions between items. For at least two reasonable distributions, the ϕ_{pq}'s are symmetric. For example, in the graphical latent variable model (Anderson and Vermunt 2000; Anderson and Böckenholt 2000; Anderson 2002) and in Holland's use of the Dutch Identity (Holland 1990), the distribution of the latent variables is assumed to be posterior multivariate normal (i.e., within a response pattern, the latent variables are multivariate normal), which implies that the ϕ_{pq} parameters are variances and covariances of the latent variables. Alternatively, if the underlying distribution of the responses is multivariate normal, then the ϕ_{pq}'s are proportional to the off diagonal elements of the inverse of the correlation matrix (Goodman 1981; Becker 1989), and hence symmetric. In the remainder of this paper, we assume that $\phi_{pq} = \phi_{qp}$, a symmetric measure of association.

Returning to our example, using the estimator $\tilde{\theta}_{v1,-1}$ in equation (4), the conditionally specified response function for X_1 is

$$P(X_{v1} = h|x_{v2}, x_{v3}, x_{v4}) = \kappa_{v1} \exp[w_{1h1}(\phi_{11}(w_{2j_v1} + w_{3j_v1}) + \phi_{12}(w_{3j_v2} + w_{4j_v2})) + \beta_{1h}], \tag{5}$$

where

$$\kappa_{v1} = \left(\sum_{l=0}^{m} \exp[w_{1l1}(\phi_{11}(w_{2j_v1} + w_{3j_v1}) + \phi_{12}(w_{3j_v2} + w_{4j_v2})) + \beta_{1l}]\right)^{-1}. \tag{6}$$

Model (5) can be viewed as a multinomial logistic regression model (i.e., a proportional odds model) with restrictions on the regression coefficients. The coefficient for the rest-score $(w_{2j_v1} + w_{3j_v1})$ is $\phi_{11}w_{1h1}$ and that for the test total $(w_{3j_v2} + w_{4j_v2})$ is $\phi_{12}w_{1h1}$. Alternatively, since the w_{1h1}'s are known (or assumed), the model above can also be viewed as a conditional multinomial logistic regression model with regression coefficients ϕ_{11} for the weighted rest-score $w_{1h1}(w_{2j_v1} + w_{3j_v1})$ and ϕ_{12} for the weighted test total $w_{1h1}(w_{3j_v2} + w_{4j_v2})$.

Specifying models such as that in (5) for each of the items in the Rasch model represented by the graph in Figure 1 (b) gives us a set of conditional models, specifically, equation (5) and

$$P(X_{v2} = h|x_{v1}, x_{v3}, x_{v4}) = \kappa_{v2} \exp[w_{2h1}(\phi_{11}(w_{1j_v1} + w_{3j_v1}) + \phi_{12}(w_{3j_v2} + w_{4j_v2})) + \beta_{2h}] \tag{7}$$

$$\begin{aligned} P(X_{v3} = h|x_{v1}, x_{v2}, x_{v4}) = \ & \kappa_{v3} \exp[w_{3h1}(\phi_{11}(w_{1j_v1} + w_{2j_v1}) + \phi_{12}(w_{4j_v2})) \\ & + w_{3h2}(\phi_{22}(w_{4j_v2}) + \phi_{12}(w_{1j_v1} + w_{2j_v1})) + \beta_{3h}] \end{aligned} \tag{8}$$

$$P(X_{v4} = h|x_{v1}, x_{v2}, x_{v3}) = \kappa_{v4} \exp[w_{4h2}(\phi_{22}(w_{3j_v2}) + \phi_{12}(w_{1j_v12} + w_{2j_v1} + w_{3j_v1})) + \beta_{4h}], \tag{9}$$

where the κ_{vi}'s are defined analogous to equation (6), which ensure that the sums of probabilities over the response options of an item equal one.

The model for $P(X_{v3} = h|x_{v1}, x_{v2}, x_{v4})$ in equation (8) is more complex than the others because X_3 is directly related to both Θ_1 and Θ_2. In this case, the estimates of individual v's values on the latent variables are

$$\begin{aligned} \tilde{\theta}_{v1,-3} &= \phi_{11}(w_{1j_v1} + w_{2j_v1}) + \phi_{12}(w_{4j_v2}) \\ \tilde{\theta}_{v2,-3} &= \phi_{22}(w_{4j_v2}) + \phi_{12}(w_{1j_v1} + w_{2j_v1}). \end{aligned}$$

For any number of items, latent variables or latent variable structures, our estimate of individual v's value on latent variable p in the model for $P(X_{vi} = h|x_{kh}, k \neq i)$ is

$$\begin{aligned} \tilde{\theta}_{vp,-i} &= \phi_{pp} \sum_{k \neq i} w_{kj_vp} + \sum_{q \neq p} \phi_{pq} \sum_{k \neq i} w_{kj_vq} \\ &= \sum_q \phi_{pq} \sum_{k \neq i} w_{kj_vp}, \end{aligned} \tag{10}$$

where the "$-i$" subscript indicates that item i has not been included in the estimate of θ_{vp}. For $i = 1, \ldots, I$, the set of conditionally specified response functions is defined as

$$\begin{aligned} P(X_{vi} = h|x_{vk}, k \neq i) &= \kappa_{vi} \exp\left[\sum_p w_{ihp} \tilde{\theta}_{vp,-i} + \beta_{ih}\right] \\ &= \kappa_{vi} \exp\left[\sum_p w_{ihp}\left(\sum_q \phi_{pq} \sum_{k \neq i} w_{kj_vq}\right) + \beta_{ih}\right], \end{aligned} \tag{11}$$

where $\kappa_{vi} = (\sum_l \exp[\sum_p w_{ilp} \tilde{\theta}_{vp,-i} + \beta_{il}])^{-1}$. Such a set of conditional multinomial logistic regression models is fully conditionally specified (Gelman and Speed 1993); that is, each item is modeled conditional on all of the others.

A set of fully conditionally specified models over-determines the joint distribution of all of the items (i.e., the probabilities of response patterns); therefore, restrictions on parameters in (11) must exist if the set of conditional models are consistent or compatible with a joint distribution. For the set of models given by (11), the compatibility conditions are that $\phi_{pq} = \phi_{qp}$ for all p and q. Since for reasonable distributions of the latent variables $\phi_{pq} = \phi_{qp}$, the compatibility conditions are met. In Appendix A, we generalize Joe and Liu's (1996) proof on the necessary and sufficient conditions for a set of binary logistic regression models to be consistent with some joint distribution to our case of a set of conditional multinomial

logistic regression models. In Appendix A, we also derive the model for the joint distribution that is compatible for the set of models defined by (11).

The joint distribution that is compatible with (11) is the log-linear-by-linear association model

$$\log(P(\boldsymbol{x}_v)) = \beta_0 + \sum_i \beta_{ih_v} + \sum_i \sum_{k>i} \sum_p \sum_{q \geq p} \phi_{pq} w_{ih_v p} w_{kh_v q}, \tag{12}$$

where $\boldsymbol{x}_v$ is individual v's response pattern on the I items, and β_0 ensures that the sum over all possible response patterns equals one. The parameter β_0 does not depend on individual v's response pattern; it depends on the total sample size and the value of the other parameters. Not only is (12) the implied model for the joint distribution, but the set of conditionals *uniquely* implies (12) so long as all the conditional probabilities are greater than zero (Gelman and Speed 1993; Arnold, Castillo, and Sarabia 1999).

Consider the cross-classification of responses to the I items. Model (12) is a Poisson (log-linear) regression model with only bivariate interactions. Since the products of the known category scores $w_{ih_v p} w_{kh_v q}$ appear in the bivariate terms, the model is referred to as a linear-by-linear association model (Agresti 2002).

Returning to our example illustrated in Figure 1 (b), the LLLA model corresponding to the set of conditionally specified models in (5), (7), (8) and (9) is

$$
\begin{aligned}
P(x_{v1h}, x_{v2h}, x_{v3h}, x_{v4h}) \;=\; & \exp\left[\beta_0 + \beta_{1h_v} + \beta_{2h_v} + \beta_{3h_v} + \beta_{4h_v}\right. \\
& + \phi_{11}(w_{1h_v1}w_{2h_v1} + w_{1h_v1}w_{3h_v1} + w_{2h_v1}w_{3h_v1}) \\
& + \phi_{22}(w_{3h_v2}w_{4h_v2}) + \phi_{12}(w_{1h_v1}w_{3h_v2} + w_{1h_v1}w_{4h_v2} \\
& \left. + w_{2h_v1}w_{3h_v2} + w_{2h_v1}w_{4h_v2} + w_{3h_v1}w_{4h_v2})\right]
\end{aligned} \tag{13}
$$

Model (12) includes interactions between items loading on the same latent variable (i.e., $\phi_{pp}w_{ih_v p}w_{kh_v p}$) and between items loading on different but correlated latent variables (i.e., $\phi_{pq}w_{ih_v p}w_{kh_v q}$). In this Poisson regression model, the values of the explanatory variables are dummy or effect codes for the β_{ih_v}'s and the sums of products of the $w_{ih_v p}$'s.

The LLLA model in (12) and (13) can be fit to data by maximum likelihood estimation using the Newton-Raphson algorithm. This is can be done using the existing `glm` function in R or the `llla` function in the **plRasch** package presented in Section 4, which calls the `glm` function for MLE. When fitting (12) to data using Newton-Raphson, the probabilities of all possible response patterns must be iteratively computed. For small numbers of items, the number of possible response patterns is reasonable; however, as the number of items increases, the number of possible response patterns increases exponentially to a point that computational demands are prohibitive. For example, the number of possible response patterns for 20 dichotomous items equals $1,048,576$, for 50 dichotomous items approximately $1.13E15$, and for 100 dichotomous items approximately $1.27E30$. For many situations in education and psychology, estimating models in the Rasch family as an LLLA model requires an alternative estimation method. Our proposed solution to this problem is to use pseudo-likelihood estimation, which for our problem can be done using standard statistical software with no special programing.

3. Pseudo-likelihood estimation

The general idea of pseudo-likelihood estimation (PLE) is to take a complex problem and break it down into simpler ones and ignore the dependency between the simpler problems. Our complex problem is estimating the LLLA model for large numbers of items and the simpler problems turn out to be the conditional multinomial logistic regressions defined in (11). Rather than maximizing the likelihood of (12), in PLE the product of the likelihoods of the conditional multinomial regressions defined by (11) is maximized. The pseduo-likelihood function is the product of the likelihoods of the conditional multinomial logistic regressions, which can be maximized by maximizing the likelihood of a single "stacked" conditional logistic regression.

In Section 3.1, we show how the regression models defined by (11) are stacked and the whole set of conditional models is fit as one large conditional multinomial logistic regression model. The maximum of the pseudo-likelihood function equals the maximum of the likelihood of this single regression model. In Section 3.2, we discuss the statistical properties of the pseudo-likelihood estimates.

3.1. Stacking

In the set of logistic regressions defined by (11), the ϕ_{pq}'s are the same over the item-specific sub-models; however, there is a separate β_{ih}'s for each item. To illustrate how to set up the design matrix for the stacked regression, consider the simple case where there is one latent variable and I items, each with $(m+1)$ response options. In this case, the conditional models defined by in (11) reduce to

$$P(X_{v1} = h | x_{vk}, k \neq 1) \;=\; \frac{\exp\left[\phi_{11}(w_{1h1}\sum_{k\neq 1}w_{kh_v1}) + \beta_{1h}\right]}{\sum_{l=0}^{m}\exp\left[\phi_{11}(w_{1l1}\sum_{k\neq 1}w_{kh_v1}) + \beta_{1l}\right]} \tag{14}$$

$$P(X_{v2} = h | x_{vk}, k \neq 2) \;=\; \frac{\exp\left[\phi_{11}(w_{2h1}\sum_{k\neq 2}w_{kh_v1}) + \beta_{2h}\right]}{\sum_{l=0}^{m}\exp\left[\phi_{11}(w_{2hl1}\sum_{k\neq 2}w_{kh_v1}) + \beta_{2l}\right]} \tag{15}$$

$$\vdots \;=\; \vdots$$

$$P(X_{vI} = h | x_{vk}, k \neq I) \;=\; \frac{\exp\left[\phi_{11}(w_{Ih1}\sum_{k\neq I}w_{kh_v1}) + \beta_{Ih}\right]}{\sum_{l=0}^{m}\exp\left[\phi_{11}(w_{Il1}\sum_{k\neq I}w_{kh_v1}) + \beta_{Il}\right]} \tag{16}$$

To fit the set of models in (14) – (16), there will be $I \times (m+1)$ lines in the design matrix for each individual. In Table 1, we illustrate what the design matrix would look like for individual v where we use dummy codes for the β_{ih}'s and set $\beta_{ih} = 0$ for identification. In a column for a β_{ih}, only one entry equals 1 and all other entries in the column equal 0. The rest-scores are weighted by the w_{ih1} for the corresponding item and response option. The 0 entries in the first line for each conditional model results from the fact that $w_{i01} = 0$ and $\beta_{i0} = 0$. Although not present in Table 1, a data file would also include a 0/1 variable indicating which response option was actually selected by individual v for each item. This variable would be the "response" variable in the conditional logistic regression. In the simple case of one latent variable, the model only has $Im + 1$ parameters, which for realistic values of I and m does not pose an estimation problem.

Conditional distribution	Response option	Item 1			Item 2			$\ldots$	Item I			Weighted rest-score
		β_{11}	$\ldots$	β_{1m}	β_{21}	$\ldots$	β_{2m}	$\ldots$	β_{I1}	$\ldots$	β_{Im}	ϕ_{11}
$P(X_{v1}\lvert \boldsymbol{x}_{v,-1})$	0	0	$\ldots$	0	0	$\ldots$	0	$\ldots$	0	$\ldots$	0	0
	1	1	$\ldots$	0	0	$\ldots$	0	$\ldots$	0	$\ldots$	0	$w_{111}\sum_{k\neq 1}w_{kh_v}1$
	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
	m	0	$\ldots$	1	0	$\ldots$	0	$\ldots$	0	$\ldots$	0	$w_{1m1}\sum_{k\neq 1}w_{kh_v}1$
$P(X_{v2}\lvert \boldsymbol{x}_{v,-2})$	0	0	$\ldots$	0	0	$\ldots$	0	$\ldots$	0	$\ldots$	0	0
	1	0	$\ldots$	0	1	$\ldots$	0	$\ldots$	0	$\ldots$	0	$w_{211}\sum_{k\neq 2}w_{kh_v}1$
	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
	m	0	$\ldots$	0	0	$\ldots$	1	$\ldots$	0	$\ldots$	0	$w_{2m1}\sum_{k\neq 2}w_{kh_v}1$
$\ldots$	$\ldots$											$\ldots$
$P(X_{vI}\lvert \boldsymbol{x}_{v,-I})$	0	0	$\ldots$	0	0	$\ldots$	0	$\ldots$	0	$\ldots$	0	0
	1	0	$\ldots$	0	0	$\ldots$	0	$\ldots$	1	$\ldots$	0	$w_{I11}\sum_{k\neq I}w_{kh_v}1$
	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
	m	0	$\ldots$	0	0	$\ldots$	0	$\ldots$	0	$\ldots$	1	$w_{Im1}\sum_{k\neq I}w_{kh_v}$

Table 1: Illustration of format of design matrix for one individual for PLE of Rasch model as LLLA model with one latent variable and four items. PLE's of parameters equal the MLE parameters of a conditional logistic regression.

For our second example, consider the case of two correlated latent variables represented by graphs such as those in Figure 1. The values for the β_{ih}'s are the same as those in Table 1. A two correlated latent variable model has only two more parameters ϕ_{22} and ϕ_{12} relative to the unidimensional model. The values in the design matrix for ϕ_{22} and ϕ_{12} equal $w_{ih2} \sum_{k \neq i} w_{kh_v 2}$ and $w_{ih1} \sum_{k \neq i} w_{kh_v 2} + w_{ih2} \sum_{k \neq i} w_{kh_v 1}$, respectively.

Although the design matrix can become rather long (i.e., for n individuals, the data file would have $nI(m+1)$ lines), increasing the number of latent dimensions does not dramatically increase the number of parameters of the corresponding multinomial logistic regression model. Adding an item will increase the number of parameters by m. The conditional multinomial logit models will have at most $Im + r(r-1)/2 + r$ parameters.

Each response option must be selected at least once to obtain an estimate of β_{ih}; however, for accurate and precise parameter estimates, more than once is preferable. This has practical implications, especially for multicategory items. If a response option was never selected, this problem can be dealt with by removing from the design matrix the column corresponding to the response option never selected. This strategy can also be used to deal with the situation where items have different numbers of response options. In our example, if item i only had two response options, there would only be a single column in the design matrix, β_{i1} in Table 1 instead of m columns.

3.2. Statistical properties of parameter estimates and inference

Arnold and Strauss (1991) presented a theorem for a single parameter model that states that the pseudo-likelihood estimator of the model parameter will be asymptotically normally and consistent. Geys *et al.* (1999) generalized Arnold and Strauss (1991)'s theorem to models for multiple binary variables. In Appendix B, we show that our problem fits into the exact same form used in the proof of Geys (1999) and Geys *et al.* (1999), which implies that in our multivariate polytomous case, the pseudo-likelihood estimates of parameters of LLLA models are asymptotically multivariate normal and consistent.

The standard errors provided by MLE of the stacked conditional multinomial logistic regression model will be too small because the dependency created by including data from a single examinee I times has been ignored. Three possible methods to obtain correct estimates of standard errors exist: jackknife, bootstrap, and robust ("sandwich"). The jackknife estimator is obtained by removing the data for one individual from the data and re-estimating the model. This is done n times, once for each individual. The bootstrap method consists of taking samples of size n from the data with replacement and fitting the model to resampled data sets. This process is repeated a large number of times (e.g., $1,000$). With both methods, the standard errors of the multiply estimated parameters are computed to yield the jackknife and bootstrap estimates. For our problem, these methods perform well but can become computationally time consuming. The alternative is to compute robust standard errors.

Let $\boldsymbol{S}$ equal the model based estimate of the covariance matrix of the parameters from the MLE of the "stacked" regression. Geys (1999), Geys *et al.* (1999) and Aerts *et al.* (2002) provide a formula in the context of PLE for an empirically corrected or robust estimator, which equals

$$\tilde{\boldsymbol{\Sigma}} = \boldsymbol{S} \left(\sum_v \boldsymbol{\Delta}_v \boldsymbol{\Delta}_v' \right) \boldsymbol{S},$$

where $\boldsymbol{\Delta}_v$ is the score vector (i.e., vector of first partial derivatives of the pseudo-likelihood

function) using individual v's data.

Given an estimated covariance matrix for the parameters, if one wants to linearly reparametrize the model (e.g., Master's partial credit model), standard errors can be computed for the transformed parameters. Alternatively, given the estimated (robust) standard errors and the fact that parameters estimates are asymptotically multivariate normal, pseudo-Wald tests can be performed. The null hypothesis $\boldsymbol{L\gamma} = \boldsymbol{0}$, where $\boldsymbol{\gamma}$ is a vector of parameters and $\boldsymbol{L}$ is a matrix whose rows define linear combinations of the parameters, can be tested using the pseudo-Wald statistic

$$W^* = \boldsymbol{L\tilde{\gamma}}(\boldsymbol{L\tilde{\Sigma}L'})^{-1}\boldsymbol{\tilde{\gamma}'L'}. \tag{17}$$

The statistic W^* has an asymptotic chi-square distribution with degrees of freedom equal to the number of rows in $\boldsymbol{L}$.

Pseudo-likelihood and adjusted pseudo-likelihood ratio tests have also been proposed (Geys 1999; Geys *et al.* 1999; Aerts *et al.* 2002). These test statistics require the maximum of the pseudo-likelihood function and $\boldsymbol{S}$. The performance of these tests, as well as the pseudo-Wald statistics is an area that warrants further investigation in the context of LLLA and Rasch models. An additional area that also warrants further investigation is the applicability and development of modifications of tests of Rasch model assumptions under MLE (Glas and Verhelst 1995; Verhelst and Glas 1995; Glas 2007) to the case of PLE.

4. The plRasch package for R

To fit Rasch models as LLLA models to data using either MLE or PLE in the R programing environment (R Development Core Team 2007), the **plRasch** package needs to be loaded which requires the package **survival** (Therneau and Grambsch 2000).

In the **plRasch** package, the category weights or scale values are set to $w_0 = 0$, $w_1 = 1$, ..., $w_m = m$ for all items, which is common in Rasch modeling; however, other choices are possible. To change the scale values, a user would need to modify the function that sets up the design matrix. In the current implementation, the number of response options are assumed to be equal.

4.1. Short overview of the plRasch package

The **plRasch** package consists of five functions and a number of example sessions. The following are currently available for use in R:

`mlData` is the function that sets up the data for MLE.

`vcov.sandwich.glm` is the function that computes robust covariance matrix for the PLE parameter values for dichotomous items.

`llla` is the main function that calls various functions that fit LLLA models to data.

`RaschPLE` is a function that calls `llla` but reparameterizes the parameter estimates so that they correspond to a parametrization similar to that given by other computer programs that estimate parameters of Rasch models.

`plStackData` is the function that stacks the regressions and sets up the data for PLE.

For illustration, a number of example scripts are available that illustrate the use of `llla` and `RaschPLE` for uni-dimensional and multi-dimensional models for dichotomous and polytomous items.

4.2. The `llla` function

The `llla` function is called using the following syntax

```
llla(data, item.mtx, trait.mtx, useMLE = FALSE, uncorrected = FALSE)
```

where the arguments are

  `data` is an individual-by-item data frame or matrix. The elements of the matrix indicate the response option chosen for each item for each individual. The entries of `data` should be $0, 1, \ldots, m$.

  `item.mtx` is an item-by-latent trait adjacency matrix where an element equals 1 if the item is a direct indicator of a latent trait and 0 otherwise.

  `trait.mtx` is a latent trait-by-latent trait adjacency matrix where an element equals 1 if two latent traits are related and 0 otherwise. The diagonal elements should always equal 1.

  `useMLE` is an optional argument that specifies whether MLE (`TRUE`) should be used rather than the default PLE (`FALSE`).

  `uncorrected` is an optional argument that requests the uncorrected standard errors and covariance matrix to be output when PLE is used. The default is `FALSE`.

When `useMLE = TRUE`, MLE is used to obtain parameter estimates for an LLLA model. The `llla` function calls the function `mlData`, which sets up the data for Poisson regression corresponding to the Rasch model specified by the `item.mtx` and `trait.mtx`. The function `glm` is called by `llla` to estimate the model.

When `useMLE = FALSE`, which is the default, PLE is used to obtain parameter estimates. The `llla` function calls `stackdata` to set up the design matrix corresponding to the Rasch model indicated by the adjacency matrices `item.mtx` and `trait.mtx`. For dichotomous items, `llla` calls the `glm` function and fits a binary logistic regression model. The robust covariance matrix and standard errors of the parameters are computed using `sandwich`. For polytomous items, the conditional multinomial models are fit as discrete time proportional hazard models where time is dummy coded. The `llla` function calls the `coxph` function in the **survival** package to fit the conditional multinomial logistic regression model as a proportional hazard model. The robust standard errors and covariance matrix for the parameters are obtained from `coxph`.

The output from the `llla` function includes

  `coefficients`: The parameter estimates of the LLLA model.

  `se`: The model based standard errors obtained from MLE of a Poisson regression when `useMLE = TRUE` or the robust standard errors from the stacked logistic regression when `useMLE = FALSE`.

covb: The model based covariance matrix from the MLE of a Poisson regression model when useMLE = TRUE, and the robust covariance matrix from the stacked logistic regression when useMLE = FALSE..

se.uncorrected: The estimated standard errors from the MLE of the stacked regression (i.e., the dependency has been ignored), which are the uncorrected standard errors from PLE.

covb.uncorrected: The estimated covariance matrix for parameter estimates from the MLE of the stacked regression (i.e., the dependency has been ignored), which is the uncorrected covariance matrix from PLE.

4.3. The RaschPLE function

Estimating Rasch models as LLLA models will recover the item parameters up to an additive constant. The parameter estimates from llla will differ from those estimated by standard IRT software (e.g., **BILOG**, see **?**) by an additive constant. With the Rasch model, both the location (mean) of item parameters and the mean on latent trait levels are unknown. In this respect, the model is unidentifiable. In practice, the parameter estimates are anchored after the estimation. Either the solution is anchored to latent trait levels by setting the mean latent trait levels to 0 or to item parameters by setting the mean item parameter estimates equal to 0 (Embretson and Reise 2000). In the **plRasch** package, the wrapper function RaschPLE fits the Rasch model by calling llla and then anchors the item parameter estimates to the latent trait.

The syntax is similar to that for llla; that is,

```
RaschPLE(data, item.mtx, trait.mtx)
```

The parameter estimates of the β's obtained from RaschPLE will differ from the parameter estimates obtained by llla by an additive constant. Since a constant is added to the β's, the standard errors are unchanged.

4.4. Illustration of pseudo-likelihood estimation

The **plRasch** package can deal with large numbers of items; however, to illustrate how to use the llla function, we simulated responses to 4 dichotomous items for $1,000$ examinees from the two dimensional Rasch model represented by the graph in Figure 1 (a). The choices are represented in the data matrix as 0 or 1. The data set looks like

```
R> head(sim.data)
     [,1] [,2] [,3] [,4]
[1,]    0    0    1    0
[2,]    0    0    0    0
[3,]    1    1    0    1
[4,]    1    1    1    1
[5,]    0    1    1    0
[6,]    0    1    1    0
```

```
R> tail(sim.data)
         [,1] [,2] [,3] [,4]
[995,]      0    1    0    0
[996,]      0    0    1    1
[997,]      0    1    0    0
[998,]      0    0    1    0
[999,]      0    0    1    0
[1000,]     0    1    1    0
```

To describe the structure of the Rasch model, two adjacency matrices `item.mtx` and `trait.mtx` are used. Matrix `trait.mtx` is a square symmetric matrix with the number of rows equal to the number of latent traits. The diagonal entries of the matrix are always 1. An off-diagonal entry equals 1 if the corresponding two latent nodes are connected. In our example represented by the graph in Figure 1 (a), items 1 and 2 are connected to latent trait Θ_1, items 3 and 4 are connected to latent trait Θ_2, and the two latent traits are correlated. In R (R Development Core Team 2007), we can construct the two adjacency matrices as follows:

```
R> item.mtx <- cbind(c(1,1,0,0), c(0,0,1,1))
R> item.mtx
     [,1] [,2]
[1,]    1    0
[2,]    1    0
[3,]    0    1
[4,]    0    1

R> trait.mtx <- cbind(c(1,1), c(1,1))
R> trait.mtx
     [,1] [,2]
[1,]    1    1
[2,]    1    1
```

Using the defaults (i.e., PLE and robust estimates of the parameter covariance matrix and standard errors), the `llla` function is called as follows:

```
R> lllafit <- llla(sim.data, item.mtx, trait.mtx)
```

The output obtained from this call of `llla` is

```
R> lllafit
Log linear by linear Model Method:  Pseudo-likelihood Method

Call:  llla(sim.data, item.mtx = item.mtx, trait.mtx = trait.mtx)

Coefficients:
      Estimate Std.Error
item1  -1.7119    0.1457
```

```
item2  -0.0368     0.1006
item3   0.5596     0.0993
item4  -1.0525     0.1404
phi11   0.8042     0.1558
phi21   0.2083     0.0634
phi22   0.6142     0.1525


R> lllafit$covb
             item1          item2          item3          item4          phi11
item1  0.021216216  0.0076258178  0.0027995708  0.0027611175 -0.0156506810
item2  0.007625818  0.0101236224  0.0030251402  0.0036994717 -0.0046373620
item3  0.002799571  0.0030251402  0.0098514081  0.0073604088  0.0008531406
item4  0.002761117  0.0036994717  0.0073604088  0.0197059169  0.0005250326
phi11 -0.015650681 -0.0046373620  0.0008531406  0.0005250326  0.0242582541
phi21 -0.004302559 -0.0042649332 -0.0030143146 -0.0030226132 -0.0008862557
phi22  0.001624722  0.0003066558 -0.0069323520 -0.0168262797 -0.0003647315
             phi21          phi22
item1 -0.0043025594  0.0016247215
item2 -0.0042649332  0.0003066558
item3 -0.0030143146 -0.0069323520
item4 -0.0030226132 -0.0168262797
phi11 -0.0008862557 -0.0003647315
phi21  0.0040221621 -0.0008128305
phi22 -0.0008128305  0.0232571453
```

The above parameter estimates are for LLLA models; however, when one wants the parameters estimates scaled such that the item parameters are anchored to the latent trait, the `RaschPLE` function should be used. Continuing our example,

```
R> wrapped <- RaschPLE(sim.data, item.mtx, trait.mtx)
```

This produces the output

```
R> wrapped
$coefficients
     item1      item2      item3      item4
-0.9472009  0.7279415  1.3242807 -0.2878061


$se
     item1      item2      item3      item4
0.14565787 0.10061621 0.09925426 0.14037777


$covb
            item1       item2       item3       item4
item1 0.021216216 0.007625818 0.002799571 0.002761117
item2 0.007625818 0.010123622 0.003025140 0.003699472
item3 0.002799571 0.003025140 0.009851408 0.007360409
item4 0.002761117 0.003699472 0.007360409 0.019705917
```

Notice that the estimated parameter estimates obtained from `llla` and `RaschPLE` differ only by a constant and the `RaschPLE` output does not include the ϕ parameter; however, the standard errors and covariance matrix for the β_{i1}'s are the same.

4.5. Illustration of maximum likelihood estimation

The function `llla` can also fit LLLA models by MLE for small data sets. The `llla` function sets up the data matrix for a Poisson regression model corresponding to the Rasch model indicated by `trait.mtx` and `item.mtx`.

Continuing our example, the LLLA model is fit to the data by MLE by entering

```
R> mlefit <- llla(sim.data, item.mtx, trait.mtx, useMLE = TRUE)
```

The output is

```
R> mlefit
Log linear by linear Model
Method:  Maximum Likelihood Method

Call:  llla(data = sim.data, item.mtx = item.mtx, trait.mtx=trait.mtx,
          useMLE = TRUE)

Coefficients:
            Estimate Std.Error
(Intercept)   4.3481    0.0975
item1        -1.7110    0.1448
item2        -0.0362    0.1000
item3         0.5596    0.0991
item4        -1.0523    0.1404
phi11         0.8038    0.1555
phi21         0.2079    0.0633
phi22         0.6141    0.1525

R> mlefit$covb
            (Intercept)          item1          item2          item3
(Intercept)  0.009512918 -0.0062874003 -0.0062874003 -0.0074024742
item1       -0.006287400  0.0209742056  0.0074606937  0.0028969171
item2       -0.006287400  0.0074606937  0.0099923392  0.0028969171
item3       -0.007402474  0.0028969171  0.0028969171  0.0098158658
item4       -0.007402474  0.0028969171  0.0028969171  0.0073528116
phi11        0.002353834 -0.0157206020 -0.0047387355  0.0006609701
phi21        0.003306652 -0.0041913813 -0.0041913813 -0.0029899827
phi22        0.004395684  0.0009130767  0.0009130767 -0.0069357347
                  item4          phi11          phi21          phi22
(Intercept) -0.0074024742  2.353834e-03  0.0033066521  4.395684e-03
item1        0.0028969171 -1.572060e-02 -0.0041913813  9.130767e-04
item2        0.0028969171 -4.738736e-03 -0.0041913813  9.130767e-04
```

```
item3        0.0073528116  6.609701e-04 -0.0029899827 -6.935735e-03
item4        0.0196984861  6.609701e-04 -0.0029899827 -1.681836e-02
phi11        0.0006609701  2.417130e-02 -0.0008149761  8.649576e-05
phi21       -0.0029899827 -8.149761e-04  0.0040102666 -8.436000e-04
phi22       -0.0168183550  8.649576e-05 -0.0008436000  2.324326e-02
```

Recall that MLE is not feasible for a test with large number of items. For example, if we try to fit responses to 100 dichotomous items using MLE, the procedure will fail because there is not enough memory to hold 2^{100} possible response patterns. If the problem is too large, the following error message is issued

```
Error in 0:(ncat^nitem - 1) : result would be too long a vector
```

In our example with two correlated latent variables, four dichotomous items and 1,000 examinees, the parameter estimates obtained from PLE and MLE are nearly equivalent. The largest difference between parameter estimates equals 0.0009. Also note that the relative efficiencies range between .9935 to .9998. A number of different cases are studied in the simulations presented in the next section.

5. Simulation studies

Simulation studies are reported here to demonstrate the performance of the estimation method implemented in **plRasch** and the ability of **plRasch** to estimate parameters in a Rasch family of models. Unless otherwise noted, in all of the studies reported here, the data were simulated according to the Rasch model in equation (3). For uni-dimensional models, $\theta \sim \mathcal{N}(0,1)$ and for multidimensional models $\boldsymbol{\theta}$ is drawn from a multivariate normal distribution. The simulations in Section 5.1, which model small numbers of items, are designed to demonstrate that parameters of the LLLA model estimated by MLE and PLE are nearly identical. The simulations in Section 5.2, which model large numbers of items, are designed to illustrate that PLE works on large data sets, performs as well as standard IRT estimation methods, and to identify possible limitations of the **plRasch** package.

5.1. Comparisons with MLE

We simulated responses to 4, 5 and 6 items for 50, 100, 500 and 1000 examinees. Since the numbers of items in these simulations are small, MLE of the LLLA models is feasible and permits us to examine the similarity between parameters estimates obtained from MLE and PLE, as well as examine the robust standard errors obtained from PLE relative to those from the MLE of a LLLA model.

Unidimensional models for dichotomous items

In our first set of simulations, we simulated responses to dichotomous items where $\beta_{11} = -1$, $\beta_{21} = 0$, $\beta_{31} = 0.5$, $\beta_{41} = 1$, $\beta_{51} = -0.5$ and $\beta_{61} = 0.2$. The β_{i0}'s were set to zero.

Log-linear-by-linear association models were fit to the data sets by PLE and MLE. The results for the 4 and 5 items data sets and for different sample sizes are basically the same as those for 6 items; therefore, we only report the results for the 6 item data sets. The estimated β_{i1}'s
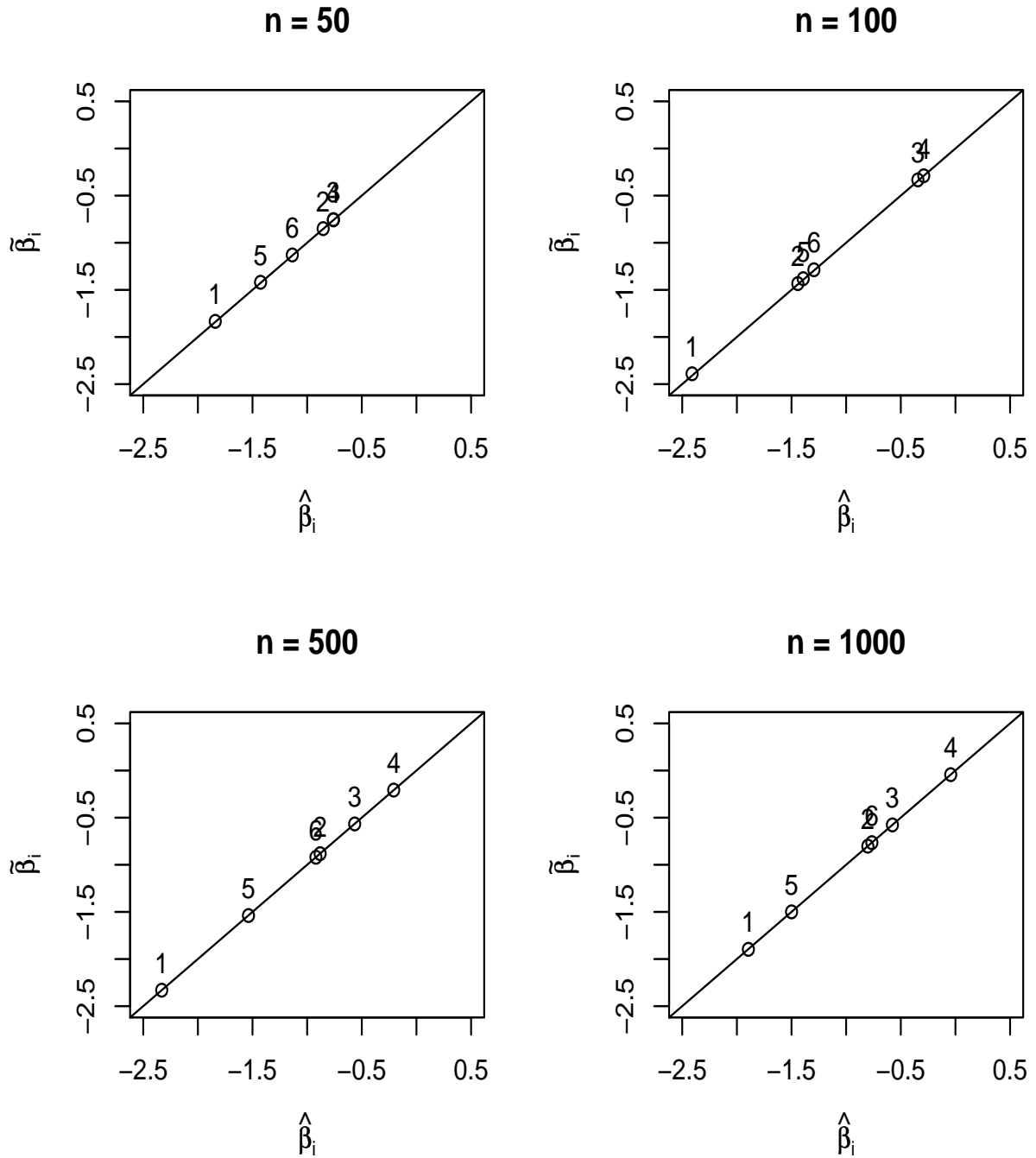
Figure 2: Pseudo-likelihood estimated parameters $\tilde{\beta}_{i1}$ of log-linear-by-linear association models plotted versus maximum likelihood estimates $\hat{\beta}_{i1}$ for 6 items and sample sizes of $n = 50$, 100, 500 and 1,000. The numbers in the plot are the item numbers and the lines are identity lines.
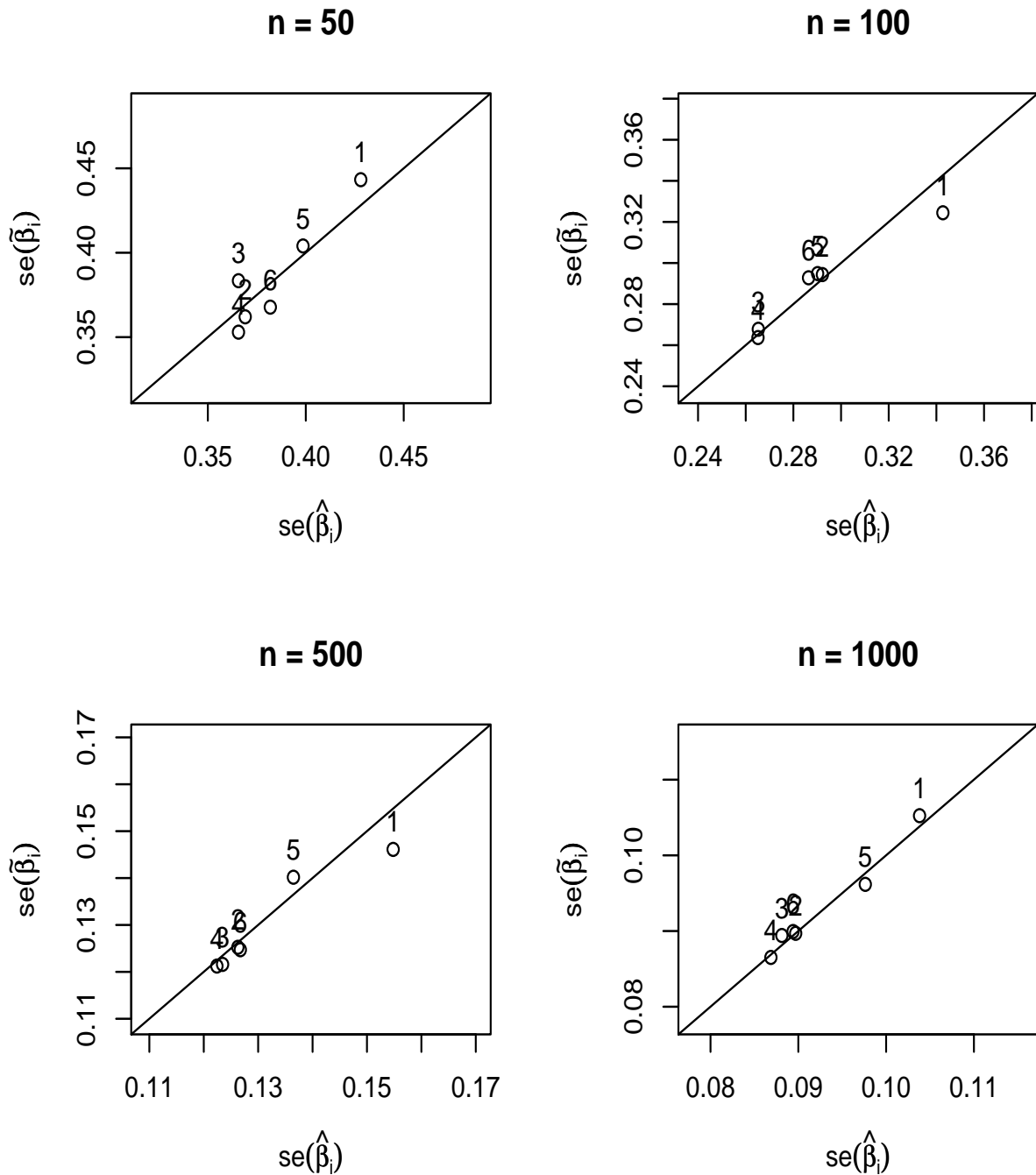
Figure 3: Robust standard errors from PLE $\text{se}(\tilde{\beta}_{i1})$ of log-linear-by-linear-association model plotted versus MLE standard errors $\text{se}(\hat{\beta}_{i1})$ for 6 items and sample sizes of $n = 50, 100, 500$ and $1,000$. The numbers in the plot are the item numbers and the lines are the identity lines.

obtained from PLE and MLE for different sample sizes are plotted in Figure 2. The numbers in the plots are the item numbers and the lines are identity lines. All of the points in these graphs fall on the identity line. The correlations between the $\tilde{\beta}_{i1}$'s and $\hat{\beta}_{i1}$'s for the different sample sizes are all greater than .9999. The ϕ parameters for a given sample size are also extremely close. The absolute values of the differences between the estimates of the ϕ equal 0.0020, 0.0026, .0005 and .0002 for $n = 50, 100, 500$ and $1,000$, respectively.

The robust standard errors from PLE are plotted against the MLE standard errors in Figure 3 for the six item data set. The numbers in the plots are the item numbers and the lines are identity lines. The PLE robust standard errors are very similar to the MLE standard errors. The results for the 4 and 5 item data sets were very similar to those for 6 items and no discernable pattern of differences between them were observed. The relative efficiencies over all parameters, numbers of items and sample sizes (i.e., 72 cases) are all larger than .97 with the exception of three cases where the relative efficiencies were .94, .94 and .96.

Although not shown here, we computed standard errors using the jackknife and bootstrap procedures. The robust standard errors are nearly identical to the jackknife, and the bootstrap standard errors for 1000 samples are similar in size to the robust and jackknife estimates, but are not as strongly correlated. We also computed parameter estimates and standard errors using the IRT program **BILOG** (**?**), which uses marginal maximum likelihood estimation. The MLE standard errors from **BILOG** are slightly smaller than those from `llla`.

Unidimensional models for polytomous items

In this section, we compare PLE and MLE parameter estimates and standard errors of the LLLA model and PLE parameter estimates of the Rasch model to parameters used to simulate the data from a unidimensional Rasch model for multicategory items. Data were simulated for items with either 3 or 5 response options.

Since the results are similar over different numbers of items, examinees and response options, we show those for 6 items with 5 response options for 500 examinees. Figure 4 contains plots of parameter estimates where different symbols (and colors) correspond to different response categories and the lines are identity lines. To distinguish between the LLLA scaled and Rasch scaled parameter estimates, we use λ_{ih} for those of a LLLA model and β_{ih} for those of a Rasch model.

At the top of Figure 4 is a plot of the PLE $\tilde{\lambda}_{ih}$ versus the MLE $\hat{\lambda}_{ih}$ of a LLLA model. The PLE estimates are virtually identical to the MLE ones for all conditions. In the middle of Figure 4 is a plot of $\tilde{\beta}_{ih}$ versus β_{ih} used to simulate the data. The PLE estimates are very similar to those used to simulate the data. As the sample size increases, the PLE estimates $\tilde{\beta}_{ih}$ get closer to those used to simulate the data. For a given sample size, the PLE estimates of $\tilde{\beta}_{ih}$ are closer to those used to simulate the data for items with 3 response options than they are for those with 5 response options.

At the bottom of Figure 4 the robust standard errors are plotted against the MLE standard errors. The Robust standard errors tend to be slightly larger than those from MLE. For a given number of items and number of response options, both robust and MLE estimates of standard errors decrease as sample size increases.
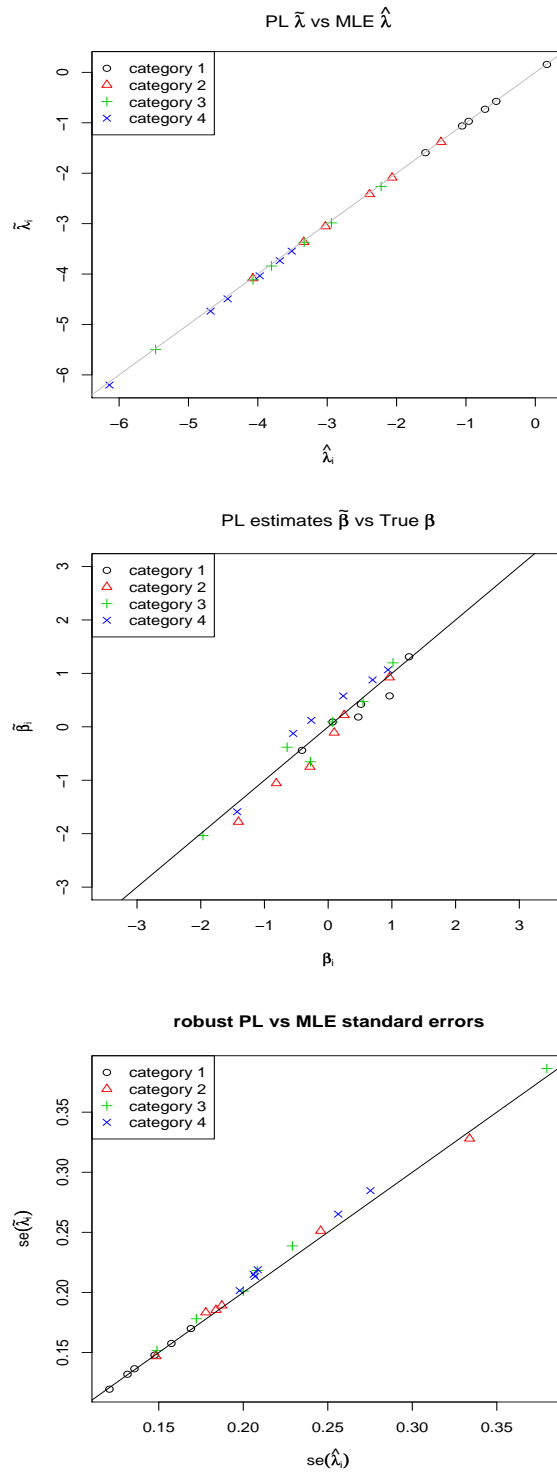
Figure 4: Results from fitting LLLA and Rasch model to data simulated from a unidimensional Rasch model for six items with five response options for 500 examinees. Different response categories are represented by different colors and symbols.

5.2. Large numbers of items

Since PLE was proposed as a way to overcome the problem of fitting LLLA models to large cross-classifications of items, in these simulations we show that PLE solves the problem and recovers parameters of a variety of Rasch models fit to simulated data.

Unidimensional models for dichotomous items

Responses to 20, 50 and 100 dichotomous items were simulated from a unidimensional Rasch model for 1,000 examinees where $\beta_{i1} \sim \mathcal{N}(0,1)$. Rasch models were fit to the data using both **plRasch** and **BILOG** (**?**). The parameters and estimates are plotted in Figure 5. The correlations between $\tilde{\beta}_{i1}$ from `RaschPLE` and $\hat{\beta}_{i1}$ from **BILOG** are all greater than .9999, regardless of the number of items or sample size. The $\tilde{\beta}_{i1}$ are also highly correlated with the β_{i1}'s used to simulate the data (i.e., .9973, .9953, and .9959 for the 20, 50 and 100 item data sets, respectively).

The robust standard errors obtained from `RaschPLE` are plotted against those obtained from **BILOG** in Figure 6. Surprisingly, the standard errors from `RaschPLE` are slightly smaller than those from **BILOG**; however, the differences tend to be .01 or smaller. We also used the jackknife and bootstrap procdures to compute standard errors for the pseudo-likelihood estimates. Those from jackknife are nearly identical to the robust standard errors and those from bootstrap are very similar in size to the robust and jackknife standard errors.

Unidimensional models for polytomous items

In this section, data were simulated from a unidimensional model for items with five response options for 20 and 50 item tests for 1,000 examinees. The $\tilde{\beta}_{ih}$ are plotted against the parameters used to simulate the data for the 20 item test in Figure 7. There is a very close correspondence between $\tilde{\beta}_{ih}$ and β_{ih}.

For 50 items with 5 categories per item and a sample size of 1,000, the length of the stacked data file was 250,000 lines, which was too large for `coxph` in R (R Development Core Team 2007) to handle. We were able to fit this model in SAS (SAS Institute Inc. 2003) using the `PROC MDC` procedure[5]; however, problems arose because not all response options were selected in the simulated data. When encountering such problems, two courses of action are possible: drop the column of the data file for response options not chosen or increase the sample size. For our simulation, we increased the sample size to 10,000, which lead to a data file with 2,500,000 lines, but there were still a couple of items with response options not chosen. With actual data, increasing the sample size may not be possible; however, deleting appropriate columns of the design matrix works well.

Multidimensional models for dichotomous items

Data from a two dimensional model exhibiting simple structure (i.e., Figure 1 (a)) were simulated for 20 and 50 dichotomous items and 1000 examinees where half of the items were direct indicators of one latent variable and the other half of the items were indicators of the other latent variable. The correlation between the latent variables was .50. In Figure 8, the PLE parameter estimates for each case are plotted against the β_{i1}'s used to simulate the data. As with our other simulations, the $\tilde{\beta}_{i1}$ are very close to the β_{i1} used to simulate the data.

---

[5]Examples using `PROC MDC` are available from `http://faculty.ed.uiuc.edu/cja/plRasch/`.
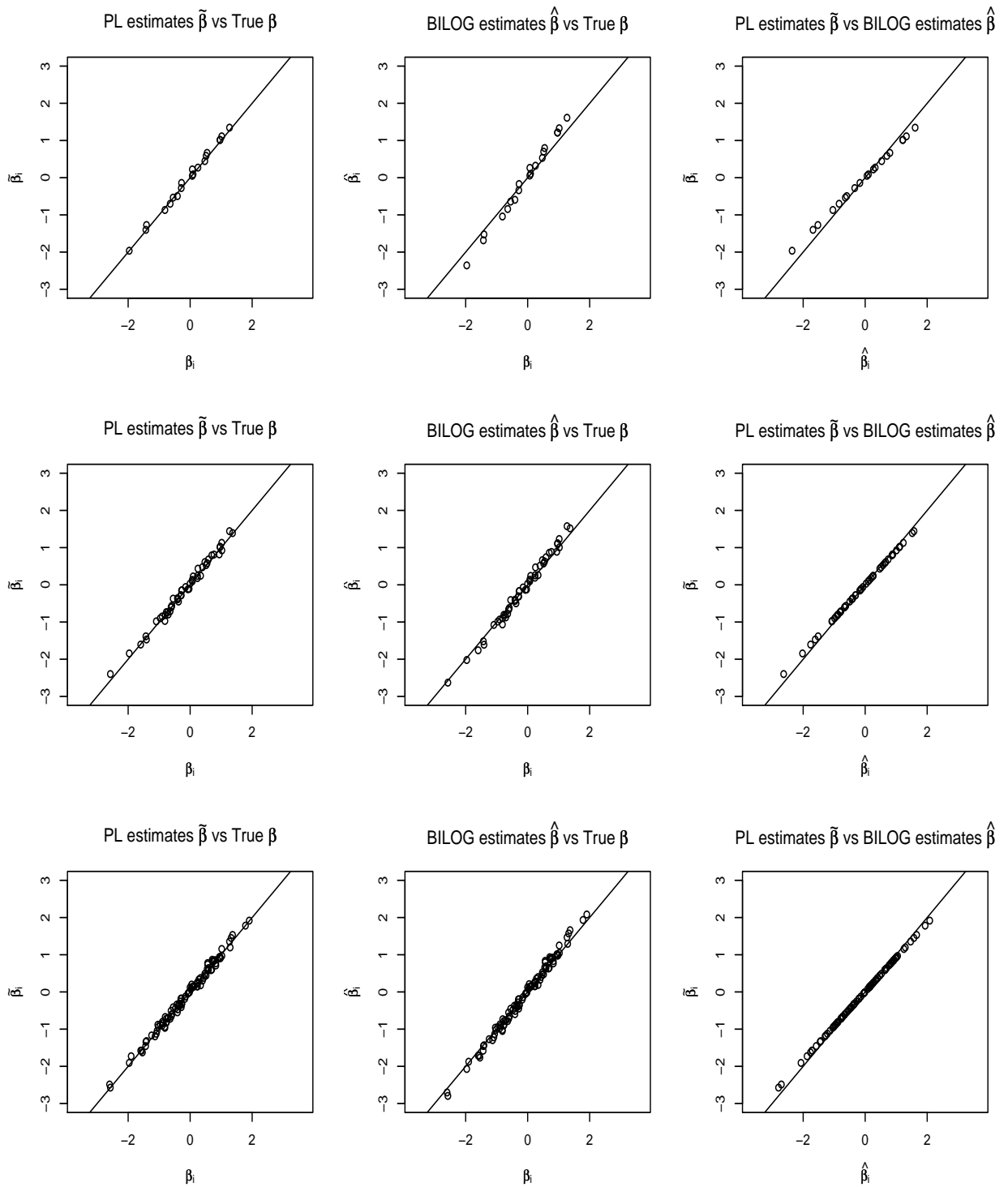
Figure 5: Comparison of $\tilde{\beta}_{i1}$ obtained from PLE, $\hat{\beta}_{i1}$ obtained from **BILOG** and true β_{i1} for $1,000$ simulated responses for 20 items (top row), 50 items (middle row), and 100 items (bottom row) from a unidimensional Rasch model.
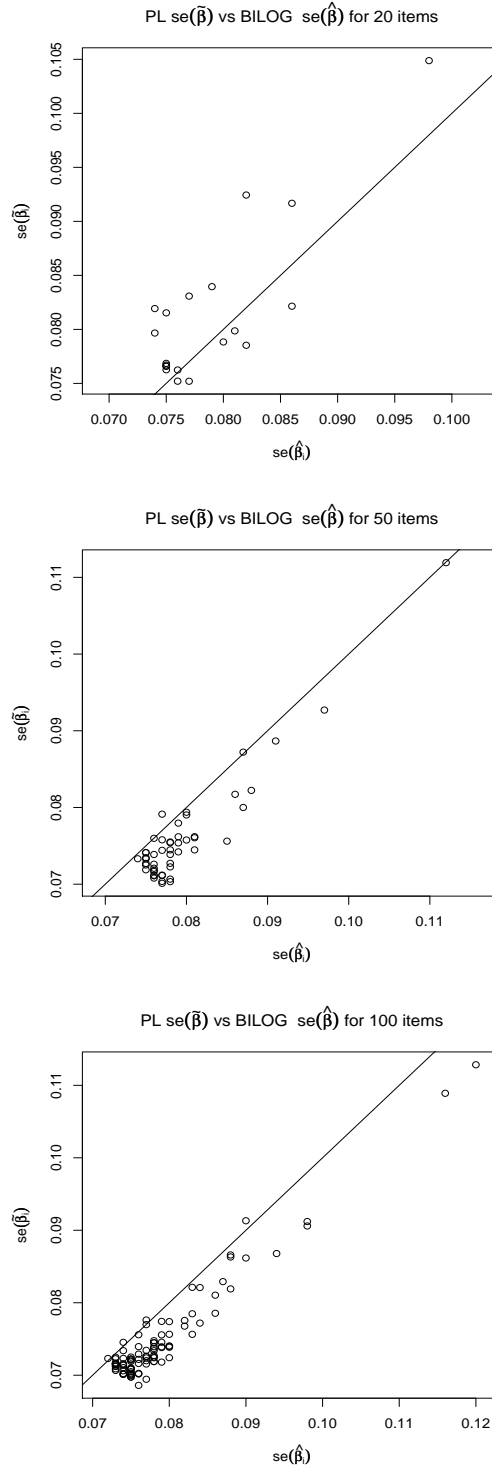
Figure 6: Comparison of standard errors obtained from `RaschPLE` (vertical axis) and **BILOG** (horizontal axis) for $1,000$ simulated responses for 20 items (top row), 50 items (middle row), and 100 items (bottom row) from a unidimensional Rasch model.
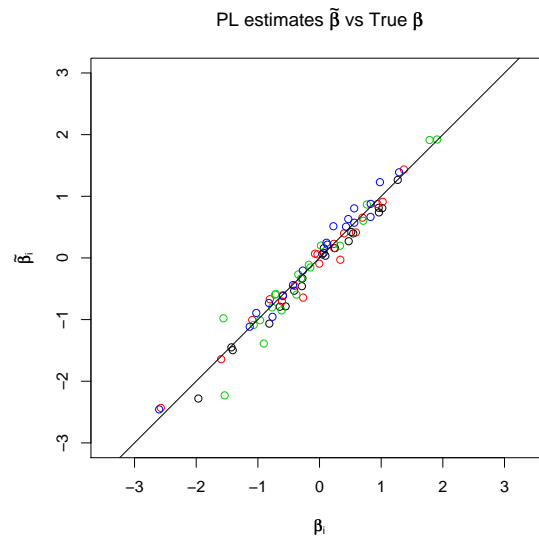
Figure 7: Comparison of $\tilde{\beta}_{ih}$ obtained from PLE and β_{ih} parameters used to simulate responses of $1,000$ examinees to 20 items each with 5 response options.

6. Discussion

The **plRasch** package presented here uses pseudo-likelihood estimation of log-linear-by-linear association models as a way to estimate parameters of a family of multidimensional Rasch models for polytomous items. With respect to parameter recovery, the performance of PLE as implemented in the **plRasch** package is excellent. In the case of dichotomous items and unidimensional models, the resulting parameter estimates are nearly identical to those obtained by **BILOG** (**?**). In all cases examined, the PLE estimates were very highly correlated with parameters used to simulate the data. Furthermore, when MLE estimates of parameters were obtained, the PLE estimates were virtually identical to the MLE parameter estimates.

The current implementation of PLE in the **plRasch** package is limited by the length of the design matrix in the stacked regression. This is mostly a function of sample size and for large numbers of items the number of categories per item. Since PLE of Rasch models requires fitting a single conditional multinomial logistic regression model to data appropriately formatted, the procedure described here can be implemented in any program that estimates conditional multinomial logistic regression models. For example, one could write an SPSS (SPSS Inc. 2006) or SAS (SAS Institute Inc. 2003) macro to prepare the data file, or one could estimate an adjacent-category ordinal logit model (rather than a conditional logit model) using LatentGOLD Version 4.0 (Vermunt and Magidson 2005). The latter computes robust, jackknife and naïve standard errors. We performed some tests using PROC MDS in SAS, which included multidimensional models for polytomous responses for large sample sizes, and did not run into any size (memory) or run-time limitations. For example, we fit a 5 dimensional Rasch model to simulated responses to 100 items with 3 response options per item and a sample size of $2,000$ examinees and obtained excellent results in terms of parameter recovery.

The PLE estimation method is very flexible. The R functions in the **plRasch** package can be
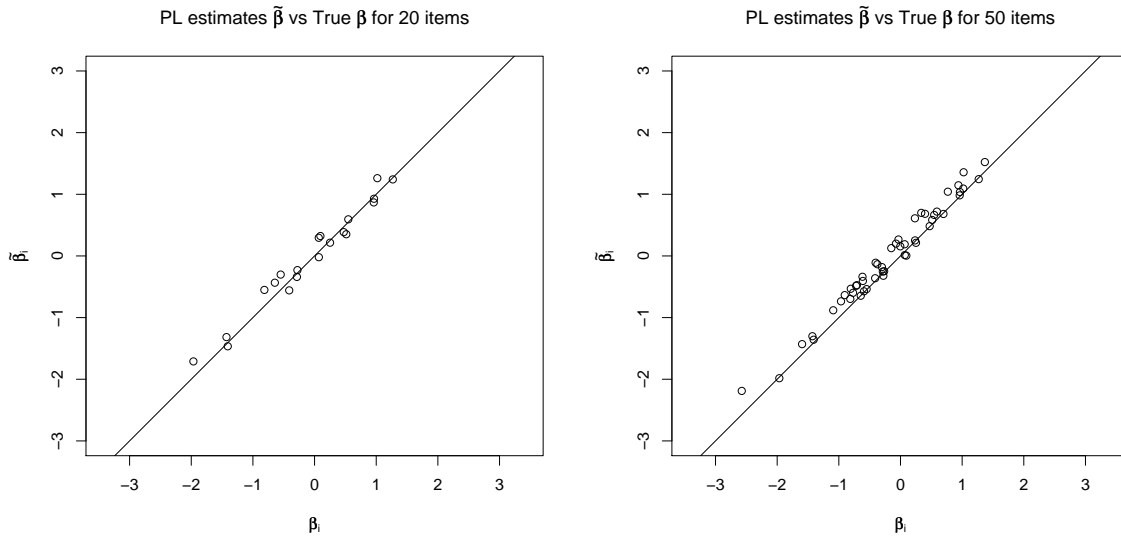
Figure 8: Comparison of $\tilde{\beta}_{i1}$ obtained from PLE and β_{i1} parameters used to simulate responses of $1,000$ examinees to 20 (left) and 50 (right) dichotomous items from a two dimensional model.

modified to fit other models in the Rasch family, including the rating scale model. Modifications are also possible such that covariates can be included in a number of different ways (Zwinderman 1996). Although we set the category weights w_{ihp} to consecutive integers, this is not the only possibility. The category weights could be set to other values that reflect specific hypotheses about response options. Such modifications require changes to the function `stackdata` that sets up the design matrix. This outlook for further developments and modifications should not distract from the fact that PLE performs extremely well, provides numerically stable results, estimates parameters of a rather large family of Rasch models, and is easy to implement, especially given the **plRasch** package in R.

Acknowledgments

References

Aerts M, Geys H, Molenberghs G, Ryan LM (2002). *Topics in Modelling of Clustered Data.* Chapman & Hall/CRC, New York. ISBN 1-58488-185-2.

Agresti A (2002). *Categorical Data Analysis.* Wiley, New York, 2nd edition. ISBN 0-471-36093-7.

Andersen EB (1995). *Rasch Models*, chapter Polytomous Rasch models and Their Estimation, pp. 271–291. Springer-Verlag, New York. ISBN 0-387-94499-0.

Anderson CJ (2002). "The Analysis of Multivariate Frequency Data by Graphical Models and Generalizations of the $RC(M)$ Association Model." *Psychological Methods*, pp. 446–467.

Anderson CJ, Böckenholt (2000). "Graphical Regression Models for Polytomous Variables." *Psychometrika*, **65**, 497–509.

Anderson CJ, Vermunt JK (2000). "Log-Multiplicative Association Models as Latent Variable Models for Nominal and/or Ordinal Data." *Sociological Methodology*, **30**, 81–121.

Anderson CJ, Yu HT (2007). "Log-Multiplicative Association Models as Item Response Theory Models." *Psychometrika*. doi:10.1007/s11336-005-1419-2.

Arnold BC, Castillo E, Sarabia JM (1999). *Conditional Specification of Statistical Models*. Springer-Verlag, New York. ISBN 0-387-98761-4.

Arnold BC, Strauss D (1991). "Pseudolikelihood Estimation: Some Examples." *Sankhya*, **53**, 233–243.

Bartholomew DJ, Knott M (1999). *Latent Variable Models and Factor Analysis*. Arnold, New York, 2nd edition. ISBN 0-340-69243-X.

Becker MP (1989). "On the Bivariate Normal Distribution and Association Models for Ordinal Categorical Data." *Statistics and Probability Letters*, **8**, 435–440.

Besag JE (1974). "Spatial Interaction and the Statistical Analysis of Lattice Systems." *Journal of the Royal Statistical Society B*, **36**, 192–236.

de Boeck P, Wilson M (2004). *Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach*. Springer-Verlag, New York. ISBN 0-387-40275-6.

de la Torre J, Patz R (2005). "Making the Most of What We Have: A Practical Application of Multidimensional Item Response Theory in Test Scoring." *Journal of Educational and Behavioral Statistics*, **30**, 295–311.

Embretson SE, Reise SP (2000). *Item Response Theory for Psychologists*. Lawrence Earlbaum, Mahwah, NJ. ISBN 0-8058-2818-4.

Fischer GH, Molenaar IW (1995). *Rasch Models*. Springer-Verlag, New York. ISBN 0-387-94499-0.

Gelman A, Speed TP (1993). "Characterizing a Joint Probability Distribution by Conditionals." *Journal of the Royal Statistical Society B*, **55**, 185–188.

Geys H (1999). *Pseudo-Likelihood Methods and Generalized Estimating Equations: Efficient Estimation Techinques for the Analysis of Correlated Multivariate Data*. Unpublished PhD Thesis, Limburgs Universitair, Diepenbeek, Belgium.

Geys H, Molenberghs G, Ryan L (1999). "Pseudolikelihood Modeling of Multivariate Outsomes in Developmental Toxicology." *Journal of the American Statistical Association*, **94**, 734–745.

Glas CA (2005). "Review of de Boeck, P. and Wilson, M. *Explanatory Item Reponse Theory Models: A Generalized Linear and Nonlinear Approach.*" *Journal of Educational Measurement*, **42**, 303–307.

Glas CA (2007). *Multivariate and Mixture Distribution Rasch Models*, chapter Testing Generalized Rasch Models, pp. 37–55. Springer-Verlag, New York. ISBN 13-978-0387-32916-1.

Glas CA, Verhelst ND (1995). *Testing the Rasch Model*, pp. 69–95. Springer-Verlag, New York. ISBN 0-387-94499-0.

Goodman LA (1979). "Simple Models for the Analysis of Association in Cross-classifications Having Ordered Categories." *Journal of the American Statistical Association*, **74**, 537–552.

Goodman LA (1981). "Association Models and the Bivariate Normal for Contingency Tables with Ordered Categories." *Journal of the American Statistical Association*, **76**, 320–334.

Goodman LA (1985). "The Analysis of Cross-classified Data Having Ordered and/or Unordered Categories: Association Models, Correlation Models, and Asymmetry Models for Contingency Tables Without Missing Entries." *The Annals of Statistics*, **13**, 10–69.

Heinen T (1993). *Discrete Latent Variable Models*. Tilburg University Press, Tilburg, The Netherlands. ISBN 90-361-9753-8.

Heinen T (1996). *Latent Class and Discrete Latent Trait Models: Similarities and Differences*. Sage, Thousand Oaks, CA. ISBN 0-8039-7433-7.

Holland PW (1990). "The Dutch Identity: A New Tool for the Study of Item Response Models." *Psychometrika*, **55**, 5–18.

Huwang L, Huwang JT (2002). "Prediction and Confidence Intervals for Nonlinear Measurement Error Models." *Statistics & Probability Letters*, **58**, 355–362.

Joe H, Liu Y (1996). "A Model for Multivariate Binary Response With Covariates Based on Conditionally Specified Logistic Regressions." *Statistics & Probability Letters*, **31**, 113–120.

Johnson M, Riezler S (2002). "Statistical Models of Syntax Learning and Use." *Cognitive Science*, **26**, 239–253.

Junker BW (1993). "Conditional Association, Essential Independence and Monotone Unidimensional item response models." *The Annals of Statistics*, **21**, 1359–1378.

Junker BW, Sijtsma K (2000). "Latent and Manifest Monotonicity in Item Respon Models." *Applied Psychological Measurement*, **24**, 65–81.

Liang G, Yu B (2003). "Maximum Pseudo Likelihood Estimation in Network Tomography." *IEEE Transactions on Signal Processing*, **51**, 2043–2053.

Mair P, Hatzinger R (2007). "Extended Rasch Modeling: The **eRm** Package for the Application of IRT Models in R." *Journal of Statistical Software*, **20**(9). URL http://www.jstatsoft.org/v20/i09/.

Molenberghs G, Verbeke G (2005). *Models for Discrete Longitudinal Data*. Springer-Verlag, New York. ISBN 978-0387025144-8.

R Development Core Team (2007). *R: A Language and Environment for Statistical Computing.* Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Rijmen, Tuerlinckx, de Boeck P, Kuppens (2003). "A Nonlinear Mixed Model Framework for Item Response Theory." *Psychological Methods*, **8**, 185–205.

SAS Institute Inc (2003). *SAS/STAT Software, Version 9.1.* Cary, NC. URL http://www.sas.com/.

Shieu CF, Chen CT, Su YH, Wang WC (2005). "Using SAS PROC NLMIXED to Fit Item Response Theory Models." *Behavior Research Methods*, **37**, 202–218.

Smit A, Kelderman H (2000). "Pseudolikelihood Estimation of the Rasch Model." *Journal of Outcome Measurement*, **4**, 513–523.

SPSS Inc (2006). *SPSS for Windows, Release 15.* SPSS Inc., Chicago, IL. URL http://www.spss.com/.

Strauss D, Ikeda M (1990). "Pseudolikelihood Estimation for Social Networks." *Journal of the American Statistical Association*, **85**, 204–212.

Therneau TM, Grambsch PM (2000). *Modeling Survival Data.* Springer-Verlag.

Tjur T (1982). "A Connection Between Rasch's Item Analysis Model and a Mulitiplicative Poisson Model." *Scandanvian Journal of Statistics*, **9**, 23–30.

Verhelst ND, Glas CA (1995). *Rasch Models: Foundations, Recent Developments and Applications*, chapter The One Parameter Logistic Model, pp. 215–237. Springer-Verlag, New York. ISBN 0-387-94499-0.

Vermunt JK, Magidson J (2005). **Latent Gold** *4.0 User's Guide.* Statistical Innovations, Inc., Belmont MA. URL http://www.statisticalinnovations.com/products/latentgold_v4.html.

Wang WC, Chen PH, Cheng YY (2004). "Improving Measurement Precision of Test Batteries Using Multidimensional Item Response Models." *Psychological Methods*, **9**, 116–136.

Wasserman S, Pattison P (1990). "Logit Models and Logistic Regressions for Social Networks: I. An Introduction to Markov Graphs and p^*." *Psychometric*, **61**, 401–425.

Zwinderman AH (1995). "Pairwise Estimation of Rasch Models." *Applied Psychological Measurement*, **19**, 369–375.

Zwinderman AH (1996). *Handbook of Modern Item Response Theory*, chapter Response Models With Manifest Predictors, pp. 245–256. Springer-Verlag, New York. ISBN 0-387-94661-6.

A. Compatibility conditions and the joint distribution

The joint distribution implied by a set of fully conditionally specified multinomial logistic regressions is shown here, as well as a derivation of the restrictions on the parameters of the conditional distributions that are both necessary and sufficient to guarantee compatibility of the conditionals distributions with a joint distribution. We show that the joint distribution is a LLLA model.

The joint distribution of variables $X_1, \ldots, X_I$ equivalent to equation (12) can be written as

$$P(\boldsymbol{x}) = \exp\left[\beta_0 + \sum_i \boldsymbol{y}_i'\left(\boldsymbol{\beta}_i + \sum_{k>i} \boldsymbol{\Psi}_{i|k}\boldsymbol{y}_k\right)\right], \tag{18}$$

where $\boldsymbol{x}$ is an $(I \times 1)$ vector that is a realization on items $X_1, \ldots, X_I$, β_0 is a normalization constant, $\boldsymbol{y}_i' = (0, \ldots, 1, 0, \ldots)$ is an $((m+1) \times 1)$ vector that indicates a specific category (response option) of variable X_i selected, $\boldsymbol{\beta}_i' = (\beta_{i0}, \ldots, \beta_{i(m)})$ is a $((m+1) \times 1)$ vector of marginal effect terms for variable i, and the $((m+1) \times (m+1))$ matrix $\boldsymbol{\Psi}_{i|k} = \boldsymbol{W}_i\boldsymbol{\Phi}\boldsymbol{W}_k'$ where $\boldsymbol{W}_i = (\boldsymbol{w}_{i1}, \ldots, \boldsymbol{w}_{iq})$ and $\boldsymbol{W}_k = (\boldsymbol{w}_{k1}, \ldots, \boldsymbol{w}_{kq})$ are $((m+1) \times r)$ matrices. where $\boldsymbol{w}_{ip} = (w_{i0p}, w_{i1p}, \ldots, w_{i(m)p})'$ is an $((m+1) \times 1)$ vector of category weights. The elements of the $(r \times r)$ matrix $\boldsymbol{\Phi}$ equal the ϕ_{pq}'s. Thus, the elements of $\boldsymbol{\Psi}_{i|k}$ (i.e., $\psi_{ih|kh}$) equal the terms that relate variable X_k to X_i in the conditional distribution of X_i given the remaining X's; that is, $\psi_{ih|kh} = \sum_p \sum_q \phi_{pq} w_{ihp} w_{khq}$.

A.1. Sufficiency

As stated in the text, the compatibility conditions are that all $\boldsymbol{\Psi}_{i|k} = \boldsymbol{\Psi}_{k|i}'$ for all i and k. We first show that given equation (18) leads to the conditional distributions of on item conditional the remaining $(I-1)$ variables defined by equation (11). The conditional distribution of (12) or equivalently (18) is

$$
\begin{aligned}
P(X_i = h | \boldsymbol{x}_{-i}) &= \frac{P(\boldsymbol{x})}{\sum_{ih=0}^m P(\boldsymbol{x})} \\
&= \frac{\exp\left[\beta_0 + \sum_{i^* \neq i} \boldsymbol{y}_{i^*}'\left(\boldsymbol{\beta}_{i^*} + \sum_{k>i^*} \boldsymbol{\Psi}_{i^*|k}\boldsymbol{y}_k\right)\right]\exp\left[\boldsymbol{y}_{ih}'\left(\boldsymbol{\beta}_i + \sum_{k \neq i} \boldsymbol{\Psi}_{i|k}\boldsymbol{y}_k\right)\right]}{\exp\left[\beta_0 + \sum_{i^* \neq i} \boldsymbol{y}_{i^*}'\left(\boldsymbol{\beta}_{i^*} + \sum_{k>i^*} \boldsymbol{\Psi}_{i^*|k}\boldsymbol{y}_k\right)\right]\sum_{ih}\exp\left[\boldsymbol{y}_{ih}'\left(\boldsymbol{\beta}_i + \sum_{k \neq i} \boldsymbol{\Psi}_{i|k}\boldsymbol{y}_k\right)\right]} \\
&= \kappa_{-i}\exp\left[\boldsymbol{y}_{ih}'\left(\boldsymbol{\beta}_i + \sum_{k \neq i} \boldsymbol{\Psi}_{i|k}\boldsymbol{y}_k\right)\right]
\end{aligned}
\tag{19}
$$

where $\boldsymbol{x}_{-i}$ is the response for all items except item i, $\boldsymbol{y}_{ih}$ is the $((m+1) \times 1)$ vector with 1 in the $(h+1)^{\text{th}}$ position and zeros elsewhere, and the normalizing constant equals

$$\kappa_{-i} = \left(\sum_{ih=0}^m \exp\left[\boldsymbol{y}_{ih}'\left(\boldsymbol{\beta}_i + \sum_{k \neq i} \boldsymbol{\Psi}_{i|k}\boldsymbol{y}_k\right)\right].\right)^{-1}$$

Equation (19) is a conditional multinomial logistic regression model that is equivalent to equation (11).

For identification, we set $\beta_{i0} = 0$ for all i, and for mathematical convenience (and consistency with the text), we set $w_{i0)} = 0$ for each item. We use these constraints and specifications

here and throughout Appendix A. Note that when taking the odds of X_{ih} and X_{i0} using conditional model (19), the κ_{-i}'s cancel out; that is,

$$\frac{P(x_{ih}|\boldsymbol{x}_{-i})}{P(x_{i0}|\boldsymbol{x}_{-i})} = \exp\left[\boldsymbol{y}_{ih}'\left(\boldsymbol{\beta}_i + \sum_{k\neq i}\boldsymbol{\Psi}_{i|k}\boldsymbol{y}_{kh}\right)\right]. \tag{20}$$

A.2. Joint distribution

We now generalize the proof in Joe and Liu (1996) to multi-category variables and derive the joint distribution that is compatible with the conditionals.

The joint distribution function for I variables is proportional to the ratio of the conditional distributions; that is,

$$\begin{aligned}
f(\boldsymbol{x}) &= f(x_1, x_2, \ldots, x_I) \\
&\propto \frac{\prod_{i=1}^I f_{i|-i}(x_i|x_1^*, \ldots, x_{i-1}^*, x_{i+1}, \ldots, x_I)}{\prod_{i=1}^I f_{i|-i}(x_i^*|x_1^*, \ldots, x_{i-1}^*, x_{i+1}, \ldots, x_I)},
\end{aligned} \tag{21}$$

where $f_{i|-i}$ is the conditional mass function for variable x_i given $\boldsymbol{X}_{-i}$, and $\boldsymbol{x}^* = (x_1^*, \ldots, x_I^*)$ is fixed and arbitrary. Using (20) in (21) with $\boldsymbol{x}$ and fixed and arbitrary $\boldsymbol{x}^*$ yields

$$\begin{aligned}
f(\boldsymbol{x}) &\propto \frac{\prod_{i=1}^I \exp\left[\boldsymbol{y}_i'\left(\boldsymbol{\beta}_i + \sum_{k<i}\boldsymbol{\Psi}_{i|k}\boldsymbol{y}_k^* + \sum_{k>i}\boldsymbol{\Psi}_{i|k}\boldsymbol{y}_k\right)\right]}{\prod_{i=1}^I \exp\left[\boldsymbol{y}_i^{*'}\left(\boldsymbol{\beta}_i + \sum_{k<i}\boldsymbol{\Psi}_{i|k}\boldsymbol{y}_k^* + \sum_{k>i}\boldsymbol{\Psi}_{i|k}\boldsymbol{y}_k\right)\right]} \\
&\propto \exp\left[\sum_i\left(\boldsymbol{y}_i'\boldsymbol{\beta}_i - \boldsymbol{y}_i^{*'}\boldsymbol{\beta}_i + \boldsymbol{y}_i'\sum_{k<i}\boldsymbol{\Psi}_{i|k}\boldsymbol{y}_k^*\right.\right. \\
&\qquad\left.\left. + \boldsymbol{y}_i'\sum_{k>i}\boldsymbol{\Psi}_{i|k}\boldsymbol{y}_k - \boldsymbol{y}_i^{*'}\sum_{k<i}\boldsymbol{\Psi}_{i|k}\boldsymbol{y}_k^* - \boldsymbol{y}_i^{*'}\sum_{k>i}\boldsymbol{\Psi}_{i|k}\boldsymbol{y}_k\right)\right].
\end{aligned} \tag{22}$$

Note that when substituting (20) in (21) we simplified the notation for $\boldsymbol{y}_{ih}$ to $\boldsymbol{y}_i$. The $\boldsymbol{y}_i$ equals the $((m+1)\times 1)$ vector of all zeros and 1 in the $(1+h)^{\text{th}}$ position where $x_i = h$ in response pattern $\boldsymbol{x}$. Since $(x_1^*, \ldots, x_I^*)$ is fixed and arbitrary, the terms $(\boldsymbol{y}_i^{*'}\boldsymbol{\beta}_i)$ and $(\boldsymbol{y}_i^{*'}\sum_{k<i}\boldsymbol{\Psi}_{i|k}\boldsymbol{y}_k^*)$ in (22) are constants, so

$$f(\boldsymbol{x}) \propto \exp\left[\sum_i\left(\boldsymbol{y}_i'\boldsymbol{\beta}_i + \boldsymbol{y}_i'\sum_{k<i}\boldsymbol{\Psi}_{i|k}\boldsymbol{y}_k^* + \boldsymbol{y}_i'\sum_{k>i}\boldsymbol{\Psi}_{i|k}\boldsymbol{y}_k - \boldsymbol{y}_i^{*'}\sum_{k>i}\boldsymbol{\Psi}_{i|k}\boldsymbol{y}_k\right)\right]. \tag{23}$$

Since $\boldsymbol{x}^*$ is arbitrary (and fixed), for convenience we will set it to be the response pattern where $x_1^* = 0, \ldots, x_I^* = 0$, which means that $\boldsymbol{\Psi}_{i|k}\boldsymbol{y}_k^* = \boldsymbol{W}_i\boldsymbol{\Psi}\boldsymbol{W}_k'\boldsymbol{y}_k^* = 0$ and $\boldsymbol{y}_i^*\boldsymbol{\Psi}_{i|k} = \boldsymbol{y}_i^*\boldsymbol{W}_i\boldsymbol{\Psi}\boldsymbol{W}_k' = 0$ (recall that the category weight for the first category equal zero, $w_{i0} = 0$), yields

$$f(\boldsymbol{x}) \propto \exp\left[\sum_i \boldsymbol{y}_i'\left(\boldsymbol{\beta}_i + \sum_{k>i}\boldsymbol{\Psi}_{i|k}\boldsymbol{y}_k\right)\right], \tag{24}$$

which is exactly the form the joint distribution in (18). For the joint distribution, we need a normalizing constant to ensure that the probabilities sum to 1 over all possible response

patterns $\boldsymbol{x}$. The joint distribution is

$$f(\boldsymbol{x}) = \exp\left[\beta_0 + \sum_i \boldsymbol{y}_i'\left(\boldsymbol{\beta}_i + \sum_{k>i} \boldsymbol{\Psi}_{i|k}\boldsymbol{y}_k\right)\right], \tag{25}$$

with

$$\beta_0 = -\log\left(\sum_{\boldsymbol{x}} \exp\left[\sum_i \boldsymbol{y}_i'\left(\boldsymbol{\beta}_i + \sum_{k>i} \boldsymbol{\Psi}_{i|k}\boldsymbol{y}_k\right)\right]\right), \tag{26}$$

where the first summation is over all possible response patterns $\boldsymbol{x}$.

A.3. Necessity

For the compatibility conditions, the ratio of (21) with fixed and arbitrary $(x_1^*, \ldots, x_I^*)$ relative to (21) with fixed and arbitrary $(x_1^\dagger, \ldots, x_I^\dagger)$ must not depend on $(x_1, \ldots, x_I)$. This ratio is

$$\begin{aligned}
R(\boldsymbol{x}; \boldsymbol{x}^*; \boldsymbol{x}^\dagger) &= \frac{\prod_{i=1}^I f_{i|-i}(x_i|x_1^*, \ldots, x_{i-1}^*, x_{i+1}, \ldots, x_I)}{\prod_{i=1}^I f_{i|-i}(x_i^*|x_1^*, \ldots, x_{i-1}^*, x_{i+1}, \ldots, x_I)} \\
&\quad \times \frac{\prod_{i=1}^I f_{i|-i}(x_i^\dagger|x_1^\dagger, \ldots, x_{i-1}^\dagger, x_{i+1}, \ldots, x_I)}{\prod_{i=1}^I f_{i|-i}(x_i|x_1^\dagger, \ldots, x_{i-1}^\dagger, x_{i+1}, \ldots, x_I)}.
\end{aligned} \tag{27}$$

Substituting the ride side of (23), which equals the product of the conditionals, in (27) with fixed and arbitrary $(x_1^*, \ldots, x_I^*)$ and $(x_1^\dagger, \ldots, x_I^\dagger)$ yields

$$\begin{aligned}
R(\boldsymbol{x}; \boldsymbol{x}^*; \boldsymbol{x}^\dagger) &= \frac{\exp\left[\sum_i \left(\boldsymbol{y}_i'\left(\sum_{k<i} \boldsymbol{\Psi}_{i|k}\boldsymbol{y}_k^*\right) - \boldsymbol{y}_i^{*'}\left(\sum_{k>i} \boldsymbol{\Psi}_{i|k}\boldsymbol{y}_k\right)\right)\right]}{\exp\left[\sum_i \left(\boldsymbol{y}_i'\left(\sum_{k<i} \boldsymbol{\Psi}_{i|k}\boldsymbol{y}_k^\dagger\right) - \boldsymbol{y}_i^{\dagger'}\left(\sum_{k>i} \boldsymbol{\Psi}_{i|k}\boldsymbol{y}_k\right)\right)\right]} \tag{28} \\[2mm]
&= \frac{\exp\left[\sum_i \left(\sum_{k>i} \boldsymbol{y}_i^{*'}\boldsymbol{\Psi}_{k|i}'\boldsymbol{y}_k - \sum_{k>i} \boldsymbol{y}_i^{*'}\boldsymbol{\Psi}_{i|k}\boldsymbol{y}_k\right)\right]}{\exp\left[\sum_i \left(\sum_{k>i} \boldsymbol{y}_{i'}^\dagger\boldsymbol{\Psi}_{k|i}'\boldsymbol{y}_k - \sum_{k>i} \boldsymbol{y}_{i'}^\dagger\boldsymbol{\Psi}_{i|k}\boldsymbol{y}_k\right)\right]} \tag{29} \\[2mm]
&= \exp\left[\sum_i \left((\boldsymbol{y}_i^* - \boldsymbol{y}_i^\dagger)' \sum_{k>i}(\boldsymbol{\Psi}_{k|i}' - \boldsymbol{\Psi}_{i|k})\boldsymbol{y}_k\right)\right]. \tag{30}
\end{aligned}$$

To go from (28) to (29), we switched the order of the summation from $k < i$ to $k > i$ by switching the labels i and k such that, for example, $\sum_{k<i} \boldsymbol{y}_i'\boldsymbol{\Psi}_{i|k}\boldsymbol{y}_k^* = \sum_{k>i} \boldsymbol{y}_i^{*'}\boldsymbol{\Psi}_{k|i}'\boldsymbol{y}_k$. For any choices of $\boldsymbol{x}^*$ and $\boldsymbol{x}^\dagger$, (30) does not depend on $\boldsymbol{x}$ (i.e., $\boldsymbol{y}_k$) if $\boldsymbol{\Psi}_{i|k} = \boldsymbol{\Psi}_{k|i}'$; that is, the compatibility conditions on the parameters of the conditional distributions are that $\boldsymbol{W}_i\boldsymbol{\Phi}\boldsymbol{W}_k' = (\boldsymbol{W}_k\boldsymbol{\Phi}\boldsymbol{W}_i')'$ for all i and k. In terms of scalars, $\psi_{ih|kl} = \sum_p \sum_q \phi_{pq}w_{ihp}w_{klq}$ and $\psi_{kl|ih} = \sum_p \sum_q \phi_{qp}w_{klq}w_{ihp}$. For $\psi_{ih|kl}$ to equal ψ_{klih} requires that $\phi_{pq} = \phi_{qp}$.

B. Pseudo-likelihood function

We show here that our problem can be put into the exact form as used by Geys *et al.* (1999), which in turn generalizes their proof regarding the consistency and asymptotic normality of pseudo-likelihood parameters for multicategory variables.

The pseudo-likelihood function is the product of a set of conditional likelihood functions or equivalently sum the log of the conditional likelihood functions, pl, which is easier to work with,

$$pl = \sum_{v=1}^{n} \sum_{s \in \mathcal{S}} \delta_s \ln[f_s(\boldsymbol{y}_v^{(s)}|\boldsymbol{\Gamma}_v)] \tag{31}$$

where the terms in (31) are defined below. In pseudo-likelihood estimation, we seek the parameters that maximize (31). If we can express our problem as in (31), then the theorem proved by Geys *et al.* (1999) applies to our problem.

The key to representing our problem in the form that Geys *et al.* (1999) use is to represent response patterns $\boldsymbol{x}$ as an indicator vector. We use the indicator vector $\boldsymbol{y}$ for this purpose. Let $y_{vih} = 1$ if individual v selected response option h on item i, and 0 otherwise. We can then represent the value on variable i by the vector of indicators $\boldsymbol{y}_{vi} = (y_{vi0}, y_{vi1}, y_{vi2}, \ldots, y_{vim})'$, which will have all 0's and one 1. For the full set of I items, we concatenate the indicator vectors $\boldsymbol{y}_{vi}$ to obtain the $((I(m+1)) \times 1)$ vector

$$\boldsymbol{y}_v = \begin{pmatrix} \boldsymbol{y}_{v1} \\ \boldsymbol{y}_{v2} \\ \vdots \\ \boldsymbol{y}_{vI} \end{pmatrix}.$$

The set of $(I(m+1) \times 1)$ vectors, $\mathcal{S}$, contains all possible vectors $\boldsymbol{s}$ with elements of 0's and 1's but not all elements equal to 0. Note that the set $\mathcal{S}$ includes vectors that do not represent possible response patterns (e.g., vectors whose elements consist of all 0's except one, or vectors that consist of all 1's). Lastly, we define the vector $\boldsymbol{y}_v^{(s)}$ as a sub-vector of $\boldsymbol{y}_v$ corresponding to a set of component vectors $\boldsymbol{y}_{vi}$ that do not have all elements equal to 0. Since our problem is fully conditionally specified, the pertinent vectors in $\mathcal{S}$ are

$$\{\boldsymbol{s}_{y_0}, \boldsymbol{s}_{y_1}, \boldsymbol{s}_{y_2}, \ldots, \boldsymbol{s}_{y_I}\} = \left\{ \begin{pmatrix} \boldsymbol{y}_{v1} \\ \hline \boldsymbol{y}_{v2} \\ \hline \boldsymbol{y}_{v3} \\ \hline \vdots \\ \hline \boldsymbol{y}_{vI} \end{pmatrix}, \begin{pmatrix} \boldsymbol{0}_{m+1} \\ \hline \boldsymbol{y}_{v2} \\ \hline \boldsymbol{y}_{v3} \\ \hline \vdots \\ \hline \boldsymbol{y}_{Iv} \end{pmatrix}, \begin{pmatrix} \boldsymbol{y}_{v1} \\ \hline \boldsymbol{0}_{m+1} \\ \hline \boldsymbol{0}_{m+1} \\ \hline \boldsymbol{y}_{v3} \\ \hline \vdots \\ \hline \boldsymbol{y}_{vI} \end{pmatrix}, \ldots, \begin{pmatrix} \boldsymbol{y}_{v1} \\ \hline \boldsymbol{y}_{v2} \\ \hline \vdots \\ \hline \boldsymbol{y}_{v,I-1} \\ \hline \boldsymbol{0}_{m+1} \end{pmatrix} \right\},$$

where $\boldsymbol{0}_{m+1}$ is a $((m+1) \times 1)$ vector with all elements equal to 0. The sub-scripts on the $\boldsymbol{s}$'s indicate the position of the sub-vector $\boldsymbol{0}_{m+1}$ (e.g., $\boldsymbol{s}_{y_0}$ has no sub-vector equal to $\boldsymbol{0}_{m+1}$, and $\boldsymbol{s}_{y_1}$ has the sub-vector $\boldsymbol{0}_{m+1}$ in position for $\boldsymbol{y}_1$ in $\boldsymbol{y}_v$.

Let δ be the set of real numbers with at least one non-zero component, in particular, $\delta = \{\delta_s | \boldsymbol{s} \in \mathcal{S}\}$. The δ values corresponding to the $\boldsymbol{s}$ vectors given above are

$$\{\delta_{s_0}, \delta_{s_{y_1}}, \delta_{s_{y_2}}, \ldots, \delta_{s_{y_I}}\} = \{I, -1, -1, \ldots, -1\},$$

and the δ values for the remaining vectors in $\mathcal{S}$ equal 0.

The joint density of $\boldsymbol{y}_v^{(s)}$ for individual v where $\boldsymbol{\Gamma}_v$ is a set of parameters for individual v is represented as $f_s(\boldsymbol{y}_v^{(s)}|\boldsymbol{\Gamma}_v)$. In our problem, $\boldsymbol{\Gamma}_v = \boldsymbol{\Gamma} = \{\beta_{11}, \ldots, \beta_{im}, \beta_{i1}, \ldots, \phi_{11}, \ldots, \phi_{pq}, \ldots\}$. Using these definitions, we obtain

$$
\begin{aligned}
pl &= \sum_{v=1}^{n} \sum_{s \in \mathcal{S}} \delta_s \ln[f_s(\boldsymbol{y}_v^{(s)}|\boldsymbol{\Gamma})] & (32)\\
&= \sum_v \left\{ I \log[f_{s_{y_0}}(\boldsymbol{y}_v^{(s_{y_0})}|\boldsymbol{\Gamma})] - \sum_i \log[f_{s_{y_i}}(\boldsymbol{y}_v^{(s_{y_i})}|\boldsymbol{\Gamma})] \right\}\\
&= \sum_v \sum_i \left\{ \log \left[\frac{(f_{(s_{y_0})}(\boldsymbol{y}_v^{(s_{y_0})}|\boldsymbol{\Gamma}))}{(f_{(s_{y_i})}(\boldsymbol{y}_v^{(s_{y_i})}|\boldsymbol{\Gamma}))} \right] \right\},\\
&= \sum_v \sum_i \log[f(\boldsymbol{y}_i|\boldsymbol{\Gamma}, \boldsymbol{y}_k, k \neq i)] & (33)
\end{aligned}
$$

where $f_{s_{y_0}}(\boldsymbol{y}_v^{(s_{y_i})}|\boldsymbol{\Gamma})$ is the joint distribution for $\boldsymbol{x}$, $f_{s_{y_i}}(\boldsymbol{y}_v^{(s_{y_i})}|\boldsymbol{\Gamma})$ is the marginal distribution of $\boldsymbol{x}$ excluding item i, and $f(\boldsymbol{y}_i|\boldsymbol{\Gamma}, \boldsymbol{y}_k, k \neq i)$ is the conditional distribution of $\boldsymbol{y}_i$ (i.e., x_i) given the remaining items (i.e., x_k for $k \neq i$).

Affiliation:

Carolyn J. Anderson
University of Illinois
Department of Educational Psychology
1310 South Sixth St, MC-708
Champaign, IL 61820, United States of America
E-mail: cja@uiuc.edu
URL: http://www.psych.uiuc.edu/~canderso/