Research Report

Efficient Full Information Maximum Likelihood Estimation for Multidimensional IRT Models

Frank Rijmen

February 2009

ETS RR-09-03

Listening. Learning. Leading.®

Efficient Full Information Maximum Likelihood Estimation for Multidimensional IRT Models

Frank Rijmen
ETS, Princeton, New Jersey

February 2009

As part of its nonprofit mission, ETS conducts and disseminates the results of research to advance quality and equity in education and assessment for the benefit of ETS's constituents and the field.

To obtain a PDF or a print copy of a report, please visit:

http://www.ets.org/research/contact.html

Copyright © 2009 by Educational Testing Service. All rights reserved.

ETS, the ETS logo, and LISTENING. LEARNING.
LEADING. are registered trademarks of Educational Testing
Service (ETS).

Abstract

Maximum marginal likelihood estimation of multidimensional item response theory (IRT) models has been hampered by the calculation of the multidimensional integral over the ability distribution. However, the researcher often has a specific hypothesis about the conditional (in)dependence relations among the latent variables. Exploiting these relations may result in more efficient estimation algorithms. A well-known example is the bi-factor model, in which each item measures a general dimension and one of *K* other dimensions, for which Gibbons and Hedeker (1992) showed that full information maximum likelihood estimation only requires the integration over two-dimensional integrals. In this paper, it is shown how the approach of Gibbons and Hedeker (1992) can be placed into a graphical model framework. The advantage of the graphical model framework is that efficient estimation schemes can be derived in a fully automatic way by applying algorithms to the graphical representation of a statistical model. This renders the approach fairly generally applicable, and tedious derivations by hand are no longer involved. The generality of the approach is demonstrated by applying it to a multidimensional IRT model with a second order dimension. It turns out that full information maximum likelihood estimation for such a model also requires the evaluation of two-dimensional integrals only.

Key words: Item response theory, MML estimation, graphical model framework, estimation schemes

Acknowledgments

The author would like to thank Alina von Davier for her support to pursue the research presented in this report.

Table of Contents

Page

The Bi-Factor Model 3

Efficient Computations Based on Graphical Modeling 6

Statement of the Problem 6

Directed Acyclic Graphs 9

Transforming the Directed Acyclic Graph (DAG) to a Junction Tree 11

Junction Tree Algorithm 14

A Multidimensional Model With a Second-Order Dimension 18

Discussion 19

References 21

Appendixes

A . Identification Restrictions for the Bi-factor and the Testlet Model 24

B . Computations for the Bi-Factor Model 26

List of Figures

Page

Figure 1. Directed acyclic graph of the bi-factor model with four subsets of items and three items within each subset..... 10

Figure 2. Moral graph for the bi-factor model with four subsets of items..... 12

Figure 3. Junction tree for the bi-factor model with four subsets of items..... 13

Figure 4. Scheduling of flows along the junction tree for the bi-factor model with four subsets of items. 16

Figure 5. Directed acyclic graph for a four-dimensional model with a second-order dimension. 18

Figure 6. Junction tree for a four-dimensional model with a second-order dimension. 19

Maximum marginal likelihood (MML) estimation of multidimensional item response theory (IRT) models has been hampered by the fact that the computations involve a numerical integration over the latent ability distribution. With increasing dimensionality, brute force integration becomes computationally demanding to a degree that the applicability of higher dimensional IRT models is jeopardized in practical settings. More specifically, when the integral over the ability distribution is evaluated using Gaussian quadrature (e.g., Bock & Aitkin, 1981), the number of calculations involved increases exponentially with the number of dimensions. Even though the number of quadrature points per dimension can be reduced when using adaptive Gaussian quadrature (Pinheiro & Bates, 1995), the total number of points again increases exponentially with the number of dimensions. Furthermore, adaptive Gaussian quadrature involves the computation of the posterior mode and variance of the latent distribution for each response pattern, with a level of complexity that increases with the number of dimensions as well.

As an alternative, so-called limited information techniques can be used to estimate the parameters of multidimensional IRT models (Jöreskog, 1994; Muthén, 1984). Limited information techniques have been developed in the field of structural equation modeling to deal with ordered categorical observed (indicator) variables. Because many IRT models can be formulated as confirmatory factor analysis models for ordinal data (Skrondal & Rabe-Hesketh, 2004; Takane & de Leeuw, 1987), limited information estimation methods can be used to estimate the parameters of these IRT models as well. Unlike MML estimation methods, the limited information techniques do not take into account the complete joint contingency table of all items, but only marginal tables up to the fourth order (Mislevy, 1985). In this way, parameter estimation can be carried out using weighted least squares estimation and is reasonably fast, even for high-dimensional models. However, the number of elements in the optimal weight matrix, which has to be invertible, grows with the fourth power of the number of items (Mislevy, 1985), so that the sample size needed to estimate an IRT model with many items, which is the typical situation in educational measurement, becomes prohibitive in many practical applications. Alternatively, Muthén, du Toit, and Spisic (1997) proposed a robust weighted least squares approach where the optimal weight matrix is replaced by a diagonal matrix, having as elements the diagonal elements of the optimal weight matrix.

IRT models that do not incorporate item discrimination parameters, such as the Rasch (1960) or the partial credit model (Masters, 1982), can be formulated as generalized linear mixed

models (Rijmen, Tuerlinckx, De Boeck, & Kuppens, 2003). For those models, quasi-likelihood methods (Breslow & Clayton, 1993) have been developed as relatively fast alternatives to MML estimation methods. Research showing that the early versions of these methods resulted in inconsistent estimates has led to attempts to improve upon these methods (Goldstein & Rasbash, 1996; Rodríguez & Goldman, 1995).

Notwithstanding the widespread use of limited information and quasi-likelihood estimation techniques and ongoing efforts for further improvements in these methods, one can safely assume that many researchers would prefer or at least consider using full-information MML techniques if they were available. In this paper, it is shown that full information MML estimation is feasible in many cases by exploiting the conditional independence relations that are implied by the model. Instead of using brute force integration over the joint distribution of all latent variables, the marginal likelihood and other quantities used during estimation can be computed by carrying out a sequence of integrations over subsets of latent variables. The particular collection of subsets depends on the conditional independence relations implied by the model. The level of computational complexity of a full information MML procedure that exploits the conditional independence relation implied by the model scales with the number of latent variables within a subset, which may be substantially lower than the total number of latent variables.

Graphical model theory offers a general framework for deriving these subsets of variables and also provides the sequence in which to carry out the integrations. A main advantage of adopting the graphical model framework is that these subsets of variables can be derived in a fully automatic way by applying algorithms to the graphical representation of the model. This renders the approach fairly generally applicable, and complicated or tedious derivations done by hand become obsolete.

In the next section, the bi-factor model is presented. The bi-factor model is an interesting model for many practical testing situations in that it incorporates both a general dimension that is common to all items and specific dimensions that pertain to subsets of items only. A second reason to start with the bi-factor model is that Gibbons and Hedeker (1992) showed how full information MML estimation in this case only requires the evaluation of integrals of dimension not higher than two.

Using the bi-factor model as a leading example, the next section of the paper will show how efficient estimation algorithms can be constructed by relying on graphical models. For the bi-factor model, the procedure involves the same set of two-dimensional integrations as specified by Gibbons and Hedeker (1992), but the result is obtained under much more general conditions. In the following section, the generality of the graphical model framework will be illustrated with a multidimensional model that incorporates a second-order dimension to account for the correlations between dimensions.

The Bi-Factor Model

In the bi-factor model, each item is an indicator of a general dimension and one of K other dimensions. Typically, the general dimension stands for the latent variable of central interest (e.g., reading ability), whereas the K other dimensions are incorporated to take into account additional dependencies between items belonging to the same subset (e.g., items of a reading test referring to the same reading passage). That is, conditional on general ability, items are assumed to be independent between but not within subsets.

For binary data, the bi-factor model can be defined as follows. Let $y_{j(k)}$ denote the binary scored response on the jth item, $j = 1,\ldots, J$, embedded within item subset k, $k = 1,\ldots, K$. There are J_k items embedded within each subset k, hence $\sum_{k=1}^{K} J_k = J$. The response vector pertaining to item subset k is denoted by $\mathbf{y}_k$, and the vector of all responses is denoted by $\mathbf{y}$. Conditional on K subset specific latent variables θ_k and a general latent variable θ_g that is common to all items, the responses are assumed to be statistically independent,

$$P\left(\mathbf{y}\middle|\boldsymbol{\theta}\right) = \prod_{j=1}^{J} P\left(y_{j(k)}\middle|\theta_g, \theta_k\right), \tag{1}$$

where $\boldsymbol{\theta} = \left(\theta_g, \theta_1, \ldots, \theta_k, \ldots, \theta_K\right)$. Furthermore, $\pi_j = P\left(y_{j(k)} = 1\middle|\theta_g, \theta_k\right)$ is related to a linear function of the latent variables through a link function $g\left(\cdot\right)$,

$$g\left(\pi_j\right) = \alpha_{jg}\theta_g + \alpha_{jk}\theta_k + \beta_j, \tag{2}$$

where $g(\cdot)$ is typically the probit or logit link function. The parameter β_j is the intercept parameter for item *j*, and α_{jg} and α_{jk} are the slopes or loadings of item *j* on the general and specific latent variables. Note that several distinct but formally equivalent parameterizations are being used in the IRT and factor analysis literature for the model presented in Equation 2.

The slope parameters $\alpha_{jg,k}$ are assumed to be known in the so-called one parameter IRT models. When an item guessing parameter is also incorporated into the expressions for the π_j's, a so-called three parameter IRT model is obtained. Furthermore, for polytomous responses, the model can be extended in a straightforward way by choosing a link function $g(\cdot)$ for polytomous data (Fahrmeir & Tutz, 2001).

In order to identify the model, the location and scale of all dimensions have to be fixed. Typically, the mean and variance of each dimension are set to zero and one, respectively. In addition, *K* restrictions are needed stemming from the invariance of the model. This can be achieved by setting the correlations between each of the *K* specific dimensions and the general dimension to zero. Further details on identification are provided in Appendix A.

The testlet model (Bradlow, Wainer, & Wang, 1999) is a special case of the bi-factor model. It is obtained by constraining the loadings on the specific dimension to be proportional to the loadings on the general dimension within each testlet (Li, Bolt, & Fu, 2006; see also Appendix A).

In a MML estimation framework, the latent variables are assumed to be random variables with joint distribution $r(\boldsymbol{\theta})$. Then, the marginal probability of a response pattern is obtained by integrating out the latent variables,

$$P(\mathbf{y}) = \int_{\boldsymbol{\theta}} P(\mathbf{y}|\boldsymbol{\theta}) r(\boldsymbol{\theta})\, d\boldsymbol{\theta}\,. \tag{3}$$

When the latent variables are discrete rather than continuous, the integral in Equation 3 is replaced by a finite sum.

MML estimation, whether through a direct maximization of the likelihood (i.e., Newton-Raphson or quasi-Newton) or an indirect maximization (i.e., the EM algorithm), requires the evaluation of Equation 3 for each response pattern. Since there is no known closed-form

solution, one relies on numerical integration techniques, approximating $P(\mathbf{y})$ as a finite sum. The computational complexity of brute force numerical integration becomes very high with an increasing number of dimensions. With a fixed number M of points for each latent variable for example, the number of terms in the finite sum amounts to M^D, where D is the number of dimensions ($D = K+1$ for the bi-factor model).

Fortunately, the computation of the marginal probabilities $P(\mathbf{y})$ and other quantities that are used during estimation do not require numerical integration over a D-dimensional grid of points, if one is willing to make some assumptions with regard to the distribution of the latent variables. Exploiting these assumptions and the conditional independence relations implied by the bi-factor structure, $P(\mathbf{y})$ can be obtained through a sequence of integrations in two dimensions. More specifically, the integrations involve the sets (θ_g, θ_k), $k = 1, ..., K$.

Gibbons and Hedeker (1992) derived this result for the bi-factor model under a specific set of conditions:

$$g(\cdot) \text{ is the probit link}$$

$$\boldsymbol{\theta} \sim N(\mathbf{0}, \mathbf{I}), \text{ where } \mathbf{I} \text{ is an identity matrix.}$$

These conditions arise from the fact that Gibbons and Hedeker (1992) base their derivation on a property of the multinormal integral (Stuart, 1958). For polytomous responses, see Gibbons et al. (2007) for a derivation along the same lines.

In the next section, it is shown that $P(\mathbf{y})$ can be obtained efficiently through a sequence of integrations over the sets (θ_g, θ_k) under much more general conditions. More specifically, the derivation holds for any link function rather than being limited to the probit link; the distribution of $\boldsymbol{\theta}$ is not assumed to be multivariate normal; and finally, the assumption of independent latent variables is relaxed to the assumption that all specific latent variables are conditionally independent given the general latent ability.

The relaxation of the last assumption deserves some further discussion. As is discussed in Appendix A, some constraints have to be imposed in order to identify the bi-factor model. When all latent variables are assumed to be independent (as in Gibbons & Hedeker, 1992), the model

can be identified by constraining the means and variances of the latent variables. When relaxing the assumption that all latent variables of the bi-factor model are independent into the assumption that the specific latent variables are conditionally independent, given the general latent ability, K additional identification restrictions are needed. One way to identify the model is to constrain the correlations between the specific and general ability to zero. In the multivariate normal case, a correlation of zero implies statistical independence. Hence, it may seem there is no point in relaxing the assumption that all latent variables of the bi-factor model are independent into the assumption that the specific latent variables are conditionally independent, given the general latent ability. However, the argument that is made in this paper is that, regardless of what constraints are needed for model identification, the derivation of Gibbons and Hedeker (1992) requires independent latent variables, whereas the derivation in the next section only requires conditional independence. Furthermore, a correlation of zero implies statistical independence for the normal distribution, but not for an arbitrary distribution $r(\boldsymbol{\theta})$. Finally, the researcher may have reasons to identify the bi-factor model by restrictions other than imposing uncorrelated dimensions.

Efficient Computations Based on Graphical Modeling

Statement of the Problem

The efficient MML estimation method that will be presented for the bi-factor model in this section and for a second-order factor model in the subsequent section are essentially specific instantiations of a general expectation maximization (EM) algorithm that has been developed in the field of graphical modeling (Lauritzen, 1995). The algorithm is efficient in that the E-step is carried out using local computations on subsets of variables that are conditionally independent. These sets are derived by working from the graphical representation of a statistical model. A thorough account of the general procedure involves a substantial amount of graph theory and is outside the scope of this paper. The main results will be stated without proof. The interested reader is referred to Cowell, Dawid, Lauritzen, and Spiegelhalter (1999) for a more in-depth account. Instead, a more intuitively based account is presented using the bi-factor model as a leading example.

After computing the relevant quantities in the E-step in an efficient way, parameters can be updated in the M-step in the same way as in a traditional EM algorithm that relies on brute

force integration in the E-step (see Bock, Gibbons, & Muraki, 1988, for an EM algorithm for the multidimensional IRT model; or Fahrmeir & Tutz, 2001, for a description of an EM algorithm for generalized linear mixed models).

As a starting point, consider how an EM algorithm proceeds for the bi-factor model. The E-step consists of the computation of the expected complete data log likelihood, where the expectation is taken over the posterior distribution $q\left(\boldsymbol{\theta}\middle|\mathbf{y};\hat{\boldsymbol{\alpha}}^{old}\right)$ of the latent variables given the data and the vector of provisional parameter estimates $\hat{\boldsymbol{\alpha}}^{old}$:

$$Q\left(\boldsymbol{\alpha};\hat{\boldsymbol{\alpha}}^{old}\right)=E\left\{\log L_c\left(\mathbf{y},\boldsymbol{\theta};\boldsymbol{\alpha}\right)\middle|\mathbf{y};\hat{\boldsymbol{\alpha}}^{old}\right\},$$

$$=\sum_i\int_{\boldsymbol{\theta}}\log P\left(\mathbf{y}\middle|\boldsymbol{\theta}\right)q\left(\boldsymbol{\theta}\middle|\mathbf{y};\hat{\boldsymbol{\alpha}}^{old}\right)d\boldsymbol{\theta}+\sum_i\int_{\boldsymbol{\theta}}\log r\left(\boldsymbol{\theta}\right)q\left(\boldsymbol{\theta}\middle|\mathbf{y};\hat{\boldsymbol{\alpha}}^{old}\right)d\boldsymbol{\theta} \tag{4}$$

The first term, using the bi-factor structure can be rewritten as

$$\begin{aligned}
&\sum_i\int_{\boldsymbol{\theta}}\log P\left(\mathbf{y}\middle|\boldsymbol{\theta}\right)q\left(\boldsymbol{\theta}\middle|\mathbf{y};\hat{\boldsymbol{\alpha}}^{old}\right)d\boldsymbol{\theta}\\
&\quad=\sum_i\int_{\boldsymbol{\theta}}\sum_k\log P\left(\mathbf{y}_k\middle|\theta_g\theta_k\right)q\left(\boldsymbol{\theta}\middle|\mathbf{y};\hat{\boldsymbol{\alpha}}^{old}\right)d\boldsymbol{\theta}\\
&\quad=\sum_i\sum_k\int_{\boldsymbol{\theta}}\log P\left(\mathbf{y}_k\middle|\theta_g\theta_k\right)q\left(\boldsymbol{\theta}\middle|\mathbf{y};\hat{\boldsymbol{\alpha}}^{old}\right)d\boldsymbol{\theta}\\
&\quad=\sum_i\sum_k\int_{\theta_g}\int_{\theta_k}\log P\left(\mathbf{y}_k\middle|\theta_g\theta_k\right)\int_{\boldsymbol{\theta}_{(k)}}q\left(\boldsymbol{\theta}\middle|\mathbf{y};\hat{\boldsymbol{\alpha}}^{old}\right)d\boldsymbol{\theta}_{(k)}d\theta_k d\theta_g\\
&\quad=\sum_i\sum_k\int_{\theta_g}\int_{\theta_k}\log P\left(\mathbf{y}_k\middle|\theta_g\theta_k\right)q\left(\theta_g,\theta_k\middle|\mathbf{y};\hat{\boldsymbol{\alpha}}^{old}\right)d\theta_k d\theta_g
\end{aligned} \tag{5}$$

where $\boldsymbol{\theta}_{(k)}$ is the vector of subset specific latent variables less θ_k. Hence, the first term of $Q\left(\boldsymbol{\alpha};\hat{\boldsymbol{\alpha}}^{old}\right)$, which is the quantity to be maximized with respect to the parameters during the subsequent M-step, involves a sequence of bidimensional integrations over the sets $\left(\theta_g,\theta_k\right)$. Such a reduction does not hold in general for the second term. It is easily verified, however, that integrations over the same sets $\left(\theta_g,\theta_k\right)$ are obtained when incorporating the assumption that the

specific latent variables are conditionally independent of each other given the general latent variable,

$$r(\boldsymbol{\theta}) = r(\theta_g)\prod_k r(\theta_k|\theta_g). \tag{6}$$

Through algebraic manipulations, the very important result is obtained that $Q(\boldsymbol{\alpha};\hat{\boldsymbol{\alpha}}^{old})$ only contains two dimensional integrals over the sets (θ_g,θ_k). Note that these are the same sets as the ones obtained by Gibbons and Hedeker (1992). One is still short of proving their result under more general conditions, however. The reason is the computation of the posterior density $q(\theta_g,\theta_k|\mathbf{y};\hat{\boldsymbol{\alpha}}^{old})$. The default but cumbersome way to compute $q(\theta_g,\theta_k|\mathbf{y};\hat{\boldsymbol{\alpha}}^{old})$ would be through an application of Bayes' theorem,

$$q(\theta_g,\theta_k|\mathbf{y};\hat{\boldsymbol{\alpha}}^{old}) = \frac{r(\theta_k,\theta_g)P(\mathbf{y}|\theta_k,\theta_g)}{\int_{\theta_g}\int_{\theta_k} r(\theta_k,\theta_g)P(\mathbf{y}|\theta_k,\theta_g)d\theta_k d\theta_g}, \tag{7}$$

where $r(\theta_k,\theta_g)P(\mathbf{y}|\theta_k,\theta_g) = s(\mathbf{y},\theta_k,\theta_g)$ can be obtained by marginalizing over the K-1 other latent variables,

$$s(\mathbf{y},\theta_k,\theta_g) = \int_{\boldsymbol{\theta}_{(k)}} P(\mathbf{y}|\boldsymbol{\theta})r(\boldsymbol{\theta})d\boldsymbol{\theta}_{(k)}. \tag{8}$$

Consequently, a brute force computation of the denominator in Equation 7 still requires the (numerical) evaluation of a K + 1-dimensional integral.

To conclude, in order to be able to exploit the conditional independence relations implied by the bi-factor model, a procedure is needed to compute the posterior densities $q(\theta_g,\theta_k|\mathbf{y};\hat{\boldsymbol{\alpha}}^{old})$ in a way that is more efficient than first carrying out numerical evaluations on a K+1 dimensional grid, and subsequently collapsing the resulting table over K-1 dimensions. This is where graphical modeling comes into the picture.

Directed Acyclic Graphs

Even though many useful results of graphical modeling apply to models that incorporate both discrete and continuous (normally distributed) variables (Lauritzen, 1996), a condition for models represented by a directed acyclic graph (DAG) is that continuous parents should not have discrete children. This condition is obviously not met in the bi-factor model. Therefore, we will approximate each latent variable θ by a discrete latent variable *z*, with, as equivalents of Equations 1 and 6,

$$P\left(\mathbf{y}\middle|\mathbf{z}\right)=\prod_{j=1}^{J}P\left(y_{j(k)}\middle|z_g,z_k\right) \tag{9}$$

$$P\left(\mathbf{z}\right)=P\left(z_g\right)\prod_{k}P\left(z_k\middle|z_g\right). \tag{10}$$

As a matter of fact, replacing the vector of continuous latent variables $\boldsymbol{\theta}$ with a vector of discrete latent variables **z** is tantamount to what is done when evaluating the integral over $\boldsymbol{\theta}$ using numerical integration in traditional MML estimation procedures. That is, from a computational viewpoint, there is no difference at all between having $\boldsymbol{\theta}$ in the model formulation and approximating the integrals over $\boldsymbol{\theta}$ through numerical integration over a discrete grid **z** on the one hand, and approximating the model through the estimation of its discrete counterpart incorporating **z** on the other hand. By estimating $P\left(\mathbf{z}\right)$ (and **z**) from the data, any distribution $r\left(\boldsymbol{\theta}\right)$ can be approximated (Aitkin, 1999; Laird, 1978). When $\boldsymbol{\theta}$ is assumed to be multivariate normally distributed, both the values **z** and their probabilities $P\left(\mathbf{z}\right)$ are pre-specified using the formulas for Gaussian quadrature (Abramowitz & Stegun, 1974), and then centered and scaled separately for each score pattern based on the maximum and variance of the posterior distribution $q\left(\theta_g,\theta_k\middle|\mathbf{y};\hat{\boldsymbol{\alpha}}^{old}\right)$ in the case of adaptive Gaussian quadrature (Pinheiro & Bates, 1995).

The starting point is to represent the bi-factor model (with discrete latent variables *z*) as a DAG. DAGs have been used for a long time to represent statistical models, especially in the structural equation and factor analysis literature. Figure 1 (top panel) represents the DAG of the bi-factor model characterized by Equations 9 and 10 for a model with four subsets of items and three items within each subset. In the graph, each node corresponds to a random variable, and the

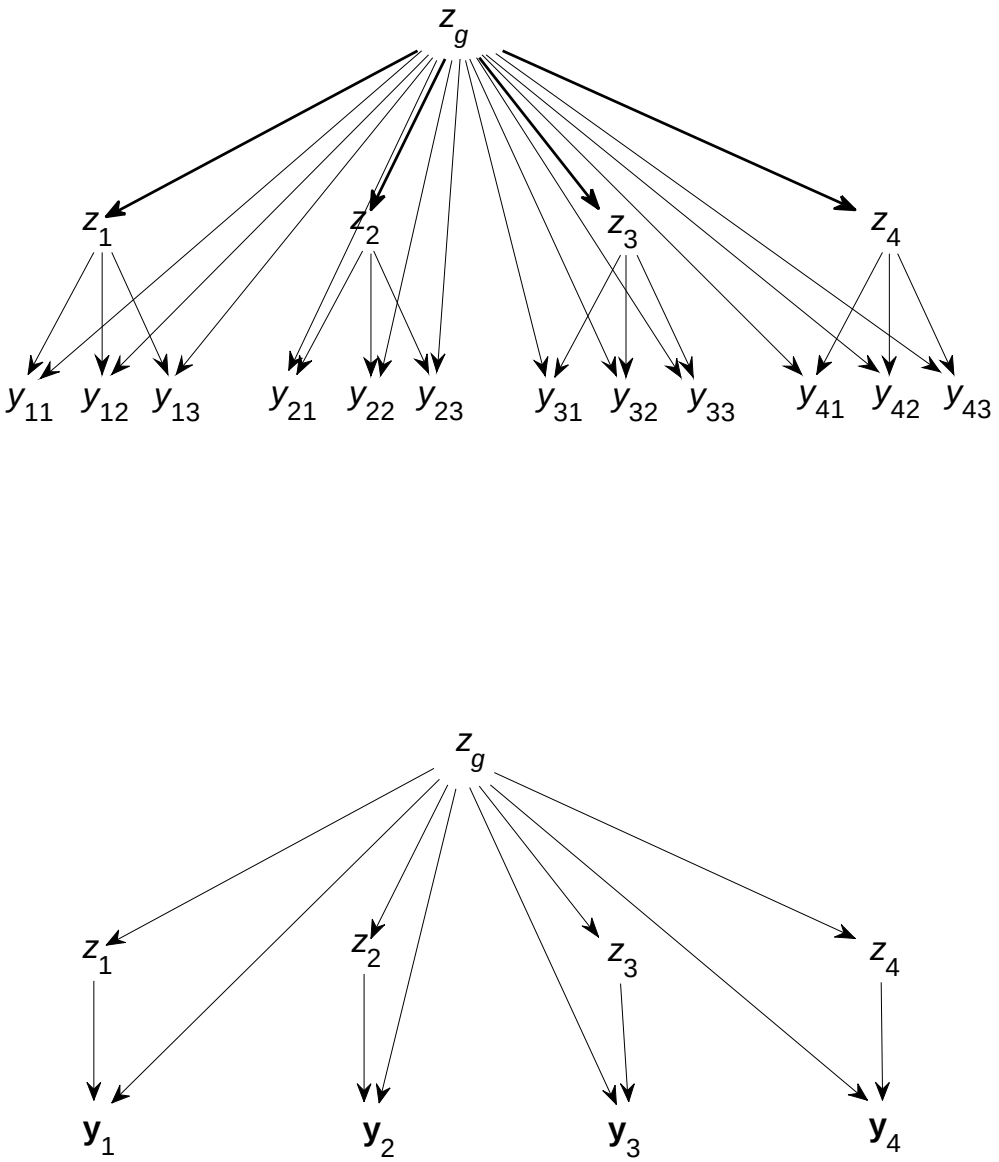

Figure 1. **Directed acyclic graph of the bi-factor model with four subsets of items and three items within each subset. Top: Individual items represented as nodes. Bottom: Item subsets represented by a single node.**

(absence of) directed edges between nodes represent conditional (in)dependence relationships. For example, the directed edge between z_g and z_1 (z_g is called a parent of z_1) represents the conditional dependence of z_1 on z_g, whereas the conditional independence between z_1 and z_4 is represented by the fact that both are children of z_g, and that they are in no other way connected. The items have a special status in that they appear as terminal nodes in the DAG. As explained in Rijmen, Vansteelandt, and De Boeck (2008), the DAG (and its transformation into other types of graphs) can be simplified by collecting the set of observed variables that share the same parents and appear as terminal nodes into a single node (see Figure 1, bottom panel).

A DAG associated with a probabilistic model for a set of discrete random variables $x_1, \ldots, x_M$ admits a recursive factorization of the joint probability function,

$$P(\mathbf{x}) = \prod_{m=1}^{M} P\left(x_m \middle| pa\left(x_m\right)\right), \tag{11}$$

where $pa(x_m)$ denotes the set of variables that are parents of x_m. Applying Equation 11 to the DAG in Figure 1, one indeed obtains the factorization of the probability of the complete data vector $P(\mathbf{y}, \mathbf{z}) = P(\mathbf{y}|\mathbf{z})P(\mathbf{z})$ according to the bi-factor model characterized by Equations 9 and 10.

Note that adding edges in the DAG of Figure 1 still results in a valid factorization of the joint probability function. However, some conditional independence relations will no longer be represented in the DAG. In general, adding unnecessary edges results in loss of efficiency of computational schemes such as the one described below.

Transforming the Directed Acyclic Graph (DAG) to a Junction Tree

The core of the construction of efficient computational schemes relies on the transformation of a DAG into a junction tree. For a detailed account of the algorithms for transforming a DAG into a junction tree, see Cowell et al. (1999). I suffice by presenting the transformations with minimal details.

A first step is transforming the DAG into an undirected graph. The undirected graph is called the moral graph. It is obtained by adding an undirected edge between all nodes with a common child that are not yet joined, and dropping directions from all edges. Figure 2 displays the moral graph of the directed acyclic graph of Figure 1. No edges have to be added to the DAG

of the bi-factor model, because there was already an edge between Z_g and Z_1, Z_2, Z_3, and Z_4, respectively, with which Z_g has respectively Y_1, Y_2, Y_3, and Y_4, each as a common child.

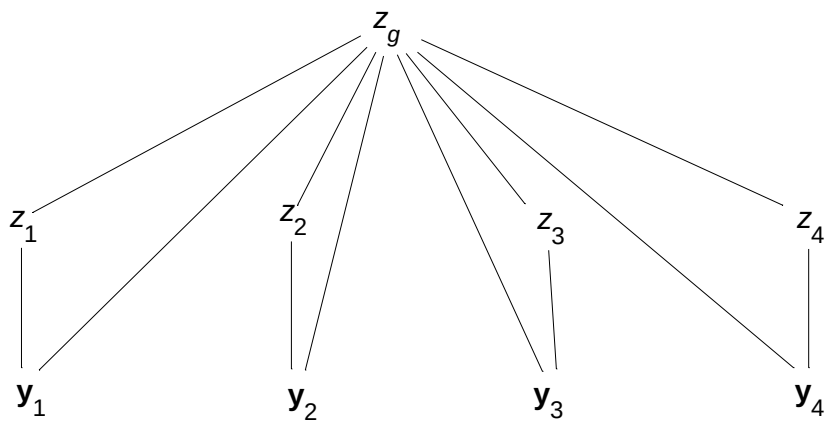

***Figure 2*. Moral graph for the bi-factor model with four subsets of items.**

Note that, were independence between all latent variables assumed, as in Gibbons and Hedeker (1992), rather than the more lenient assumption that the specific dimensions are conditionally independent given the general dimension (Equation 10), the same moral graph would have been obtained. Specifically, in that case, there would be no edges in the DAG from Z_g to Z_1, Z_2, Z_3, and Z_4, respectively. However, these edges would have been added when forming the moral graph. The fact that both DAGs result in the same moral graph is precisely why the assumption of independent latent variables can be relaxed into the assumption that the specific latent variables are conditionally independent given the general latent variable without sacrificing computational efficiency.

The moralization step ensures that a probabilistic model that satisfies the conditional independence relations implied by a DAG also satisfies the conditional independence relations implied by the undirected moral graph of the DAG. In the process of moralization, conditional

independence relations that were implied by the DAG might lose their representation in the moral graph by the process of adding edges.

Second, the moral graph is triangulated by adding edges so that chordless cycles contain no more than three nodes. A chordless cycle is a cycle in which there are only edges between consecutive nodes. In general, a triangulated graph can be obtained in many different ways, but one tries to add as few edges as possible to retain the graphical representation of the conditional independence relations that were implied by the DAG. Finding an optimal triangulation is nondeterministic polynomial-time (NP) hard (Yannakakis, 1981; for the reader not familiar with computational complexity theory, NP-hard is *very hard*), but well performing heuristic algorithms are available (Kjærulff, 1992). The moral graph in Figure 2 contains no cycles and thus is already triangulated.

A graph being triangulated is a necessary and sufficient condition for the existence of an associated junction tree. A tree is a graph whose undirected version (obtained by dropping all the directions from the edges) has a path between all pairs of nodes and has no cycles. In a junction tree, the nodes correspond to cliques. Cliques are complete subsets of nodes. A set of nodes is complete if there is an edge between every pair of nodes. The intersection between two neighboring cliques C_k and C_l is called a separator, $S_{kl} = C_k \cap C_l$.

A junction tree possesses the running intersection property: the intersection $C_k \cap C_l$ of a pair C_k, C_l of cliques is contained in every node on the unique path in the junction tree between C_k and C_l. Figure 3 shows a junction tree of cliques obtained from the triangulated moral graph of Figure 2. Again, more than one junction tree can be constructed in general.

$$\mathbf{y}_1\, z_1\, z_g \text{———} \mathbf{y}_2\, z_2\, z_g \text{———} \mathbf{y}_3\, z_3\, z_g \text{———} \mathbf{y}_4\, z_4\, z_g$$

***Figure 3*. Junction tree for the bi-factor model with four subsets of items.**

Junction Tree Algorithm

A crucial result is that a junction tree offers an alternative factorization of the joint probability function. In particular, the joint probability function of all variables can be factorized as the product of all marginal clique probabilities over the product of all marginal separator probabilities:

$$P(\mathbf{x}) = \frac{\prod_C P(\mathbf{x}_C)}{\prod_S P(\mathbf{x}_S)}, \tag{12}$$

where $\mathbf{x}_c$ and $\mathbf{x}_s$ denote the random variables that constitute clique c and separator s, respectively. For the bi-factor model, it is verified in Appendix B that the probability of the complete data vector $P(\mathbf{y}, \mathbf{z})$ can indeed be rewritten as

$$P(\mathbf{y}, \mathbf{z}) = \frac{\prod_k P(\mathbf{y}_k, z_k, z_g)}{\left[P(z_g)\right]^{K-1}}. \tag{13}$$

The cliques correspond to the sets ($\mathbf{y}_k$, z_k, z_g), and are exactly the sets of variables that appeared in the individual terms of $Q(\boldsymbol{\alpha}; \hat{\boldsymbol{\alpha}}^{old})$, the expected complete data log-likelihood for the bi-factor model (see Equation 5). Hence, computing $Q(\boldsymbol{\alpha}; \hat{\boldsymbol{\alpha}}^{old})$ using the factorization of Equation 13 results in the same dimensionality reduction as was obtained in Equation 5 through algebraic manipulations. More importantly, graphical modelling theory offers an efficient way for computing the posterior probabilities $P(z_g, z_k \mid \mathbf{y}; \hat{\boldsymbol{\alpha}}^{old})$, as is explained in the following.

The factorization of Equation 12 serves as the basis for an efficient computational scheme using local computations. A first step is to associate a non-negative *potential* function ψ to each clique and separator of the junction tree. The domain of ψ is the set of all possible realizations of the random variables in the clique or separator.

The potential values or potentials may be assigned as follows. First, all potentials are initialized with value 1. Then, each factor of Equation 11 is multiplied into the potential function of a clique that contains all the nodes corresponding to the random variables in the factor. The

way a junction tree is constructed implies that there is always such a clique; if there is more than one, it does not matter which one is chosen. Then, by definition,

$$P(\mathbf{x}) = \frac{\prod_C \psi(\mathbf{x}_C)}{\prod_S \psi(\mathbf{x}_S)}. \tag{14}$$

Next, a schedule of flows is passed along the edges of the junction tree. Let C_k and C_l be two consecutive nodes of the junction tree, with separator S_{kl} . New potentials are defined as

$$\begin{aligned} &\psi^*\left(\mathbf{x}_{S_{kl}}\right) = \sum_{C_k \setminus S_{kl}} \psi\left(\mathbf{x}_{C_k}\right), \\ &\text{and} \\ &\psi^*\left(\mathbf{x}_{C_l}\right) = \psi\left(\mathbf{x}_{C_l}\right)\lambda\left(\mathbf{x}_{S_{kl}}\right), \\ &\text{where } \lambda\left(\mathbf{x}_{S_{kl}}\right) = \frac{\psi^*\left(\mathbf{x}_{S_{kl}}\right)}{\psi\left(\mathbf{x}_{S_{kl}}\right)}, \lambda\left(\mathbf{x}_{S_{kl}}\right) \equiv 0 \text{ if } \psi\left(\mathbf{x}_{S_{kl}}\right) = 0, \end{aligned} \tag{15}$$

and $\sum_{C_k \setminus S_{kl}}$ denotes marginalization over all random variables whose corresponding nodes are in $C_k \setminus S_k$.

Jensen, Lauritzen, and Olesen (1990) proposed the following two-phase schedule: First, select an arbitrary clique of the junction tree as the root-clique. In the collection phase, flows are passed along the edges towards the root-clique. In the distribution-phase, flows are passed in the reverse direction. See Figure 4 for a scheduling of flows.

After applying the two-phase schedule, equilibrium is reached, and the clique and separator potentials correspond to the marginal probability functions of the cliques and separators, respectively. Numerical underflow can be avoided by normalizing the potentials of each clique after updating the potentials of that clique, for example by dividing each potential by the sum of the clique potentials. After applying the two-phase schedule, the clique and separator potentials then become *proportional* to the marginal probability functions of the cliques and separators, respectively. The proportionality constant equals the product of the clique normalizing constants.

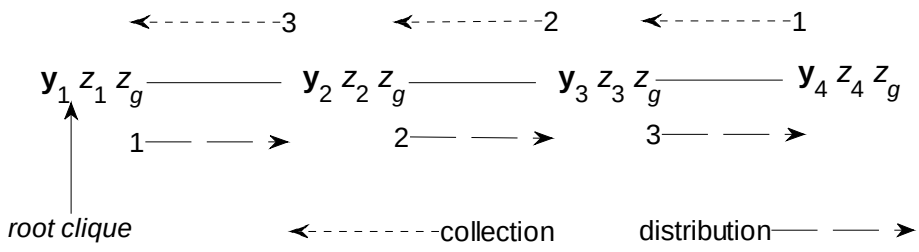

Figure 4. **Scheduling of flows along the junction tree for the bi-factor model with four subsets of items.**

When some of the variables are observed, posterior probability distributions of the unobserved variables can be obtained along similar lines. The only change is that, for each observed variable, some arbitrary clique that contains the variable is selected, and the potentials of all realizations of the clique variables involving a different state for the observed variable than the observed one are set to 0. This step is called "entering the evidence." After applying the two-phase schedule, the clique and separator potentials now are proportional to the posterior marginal probability functions of the cliques and separators, respectively. Normalizing the potentials of a clique or separator so that they sum to one yields its posterior probability distribution. Posterior distributions for individual variables are obtained by marginalizing over all other variables in the clique or separator.

For the bi-factor model, the observed variables are the items, and they all appear as terminal nodes. As mentioned before, the subsets of items sharing the same set of parents can be merged into a single node in this case. The corresponding DAG is much simpler, and consequently the construction of a junction tree and the propagation of evidence along the junction tree are also simplified considerably. The effective state space of such a merged node is only of size one and equals, for each case, the observed response pattern on the corresponding set of observed variables. This is because, when entering evidence, the potentials of all configurations other than the observed one are set to zero. Including these configurations would only result in a needless propagation of zeroes (Huang & Darwiche, 1996). Initialization is done in the same way as described before, except for the fact that now, for each case, the conditional

probabilities of the observed response patterns on the sets of terminal nodes sharing the same parent(s) are used. These conditional probabilities are computed over all data that are not missing (assuming ignorable missingness; Little & Rubin, 1987). If all data on such a set of observed variables are missing, this probability is set to 1, leaving the potentials unaltered.

In summary, by applying transformations to the graph, all conditional independence relations of the probabilistic model could be rendered explicit that did not lose their representation in graphical form during the process of moralization and triangulation. The factorization of the joint probability function along these conditional independence relations will allow for computations to be carried out locally to yield marginal or conditional posterior distributions of interest, given observed data.

This is a very important result, because it shows how the E-step of the EM algorithm can be carried out efficiently (Lauritzen, 1995). First, the junction tree is initialized using provisional parameter estimates $\hat{\boldsymbol{\alpha}}^{old}$. Then, separately for each case, the observed response pattern is entered as evidence into the network. After propagating the evidence along the junction tree, the posterior clique probabilities are obtained. These are the posterior probabilities needed to compute $Q\left(\boldsymbol{\alpha};\hat{\boldsymbol{\alpha}}^{old}\right)$, the expected complete data log-likelihood. In the M-step, updated parameters are obtained by maximizing $Q\left(\boldsymbol{\alpha};\hat{\boldsymbol{\alpha}}^{old}\right)$ with respect to the parameters.

The complexity of the E-step of this efficient EM algorithm scales with the sum of the clique state spaces. Hence, the smaller the clique state spaces, the greater the gain in efficiency obtained by using the modified instead of the standard EM algorithm.

For the bi-factor model, the cliques of the junction tree correspond to the sets ($\mathbf{y}_k$, z_k, z_g), which are exactly the sets of variables that appeared in the individual terms of $Q\left(\boldsymbol{\alpha};\hat{\boldsymbol{\alpha}}^{old}\right)$ as defined in Equation 5. Hence, after propagating the evidence along the junction tree, the (renormalized) posterior marginal clique probabilities correspond to the posterior probabilities $P\left(z_g, z_k \middle| \mathbf{y};\hat{\boldsymbol{\alpha}}^{old}\right)$. It follows that an efficient full information MML estimation method, involving a sequence of two-dimensional integrations over the sets $\left(\theta_g, \theta_k\right)$, can be derived under much more general conditions than the ones stated by Gibbons and Hedeker (1992). Appendix B contains a more detailed description of how the E-step of the EM algorithm is carried out efficiently for the bi-factor model. A main advantage of graphical modeling is that it offers a

very general procedure that can be carried out in an algorithmic way, rendering tedious algebraic manipulations of the likelihood function obsolete.

A Multidimensional Model With a Second-Order Dimension

In this section, the generality of the approach is illustrated with a multidimensional IRT model incorporating a second-order dimension. In the model, all dependencies between the first-order dimensions are explained by the second-order dimension. The DAG for a model with four first-order dimensions is given in Figure 5. The absence of edges between the first-order dimensions represents their conditional independence given the second-order factor. The moral graph is simply obtained by dropping the directions of the edges, and a junction tree is given in Figure 6. From Figure 6, it follows that an efficient E-step of the EM algorithm involves a sequence of two-dimensional integrals over the sets of latent variables (z_k, z_g), $k = 1,\ldots, 4$, as was the case for the bi-factor model. As a matter of fact, comparing Figure 5 to Figure 1 reveals that the second-order model is a bi-factor model with the additional restriction that the conditional item response probabilities do not directly depend on the general dimension.

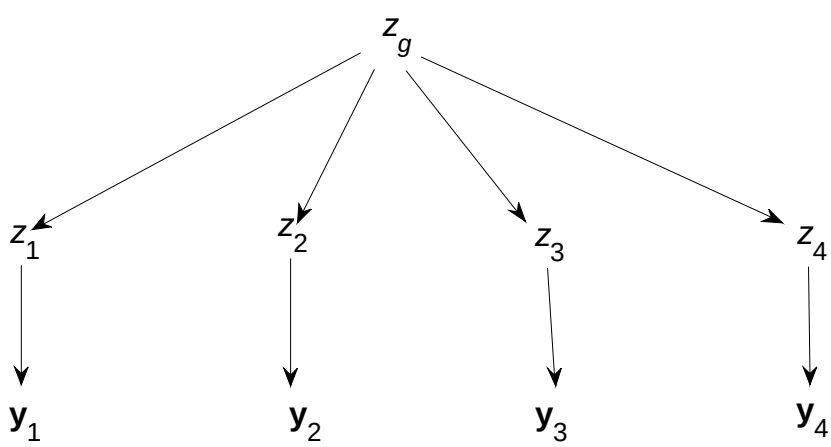

***Figure 5*. Directed acyclic graph for a four-dimensional model with a second-order dimension.**

***Figure 6*. Junction tree for a four-dimensional model with a second-order dimension.**

Discussion

Graphical models have been used widely to represent statistical models in a visually attractive way. In this paper, it is argued that graphical modeling has more to offer. It was explained in some detail how efficient MML estimation procedures can be developed by applying transformations to the directed graph representing the statistical model. Working in a graphical modeling framework has the advantage that, once the statistical model has been represented in a DAG, the transformations can be applied in a fully algorithmic way, and the sets of conditionally independent variables are obtained automatically.

In this paper, attention was limited to multidimensional IRT and item factor analysis methods, because these methods represent an important class of statistical models in the quantitative social and behavioral sciences. Moreover, especially in the IRT community with its focus on full information MML estimation, the applicability of multidimensional models has long been considered rather limited because of the intractable multidimensional integral that appears in the marginal probability of a response pattern. The important result that was established in this paper is that this dismissal of multidimensional models because of computational considerations has not been entirely fair with respect to such models. As long as one is willing to impose some structure on the model, the estimation can be simplified by exploiting the imposed conditional independence relations.

When one is not willing to impose any conditional independence relations, full information MML estimation still becomes infeasible with an increasing number of dimensions. However, such models are mostly considered in an exploratory stage only, after which typically

a simple structure model is constructed for (a subset of) the items. In the exploratory stage, obtaining full information MML results is not of crucial interest. The following sequential procedure could be followed: first, several models could be explored using limited-information estimation techniques. Once a good candidate model is established that incorporates a simplified structure, the parameters of this model can be estimated using the efficient EM algorithm.

References

Abramowitz, M., & Stegun, I. (1974). *Handbook of mathematical functions*. New York: Dover.

Aitkin, M. A. (1999). A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics, 55,* 117–128.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: an application of the EM algorithm. *Psychometrika, 46*, 443-459.

Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full-information item factor analysis. *Applied Psychological Measurement, 12,* 261-280.

Bradlow, E. T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika, 64,* 153-168.

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

Cowell, R. G., Dawid, A. P., Lauritzen, S. L., & Spiegelhalter, D. J. (1999). *Probabilistic networks and expert systems.* New York: Springer.

Fahrmeir, L., & Tutz, G. (2001). *Multivariate statistical modelling based on generalized linear models* (2nd ed.). New York: Springer.

Gibbons, R. D., & Hedeker, D. (1992). Full-information item bi-factor analysis. *Psychometrika, 57*, 423-436.

Gibbons, R. D., Bock, R. D., Hedeker, D., Weiss, D. J., Segawa, E., Bhaumik, et al. (2007). Full-information item bifactor analysis of graded response data. *Applied Psychological Measurement, 31,* 4-19.

Goldstein, H., & Rasbash, J. (1996). Improved approximations for multilevel models with binary responses. *Journal of the Royal Statistical Society, Series A, 159*, 505–513.

Huang, C., & Darwiche, A. (1996). Inference in belief networks: A procedural guide. *International Journal of Approximate Reasoning, 15*, 225-263.

Jensen, F. V., Lauritzen, S. L., & Olesen, K. G. (1990). Bayesian updating in causal probabilistic networks by local computation. *Computational Statistics Quarterly, 4,* 269-282.

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

Kjærulff, U. (1992). Optimal decomposition of probabilistic networks by simulated annealing. *Statistics and Computing, 2*, 7-17.

Laird, N. M. (1978). Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association, 73,* 805–811.

Lauritzen, S. L. (1995). The EM algorithm for graphical association models with missing data. *Computational Statistics and Data Analysis, 19,* 191–201.

Lauritzen, S. L. (1996). *Graphical models.* Oxford, UK: Clarendon Press.

Li, Y. Bolt, D. M., & Fu, J. (2006). A comparison of alternative models for testlets. *Applied Psychological Measurement, 30,* 3-21.

Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data*. New York: Wiley.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47,* 149–174.

Mislevy, R. J. (1985). *Recent developments in the factor analysis of categorical variables* (ETS Research Rep. No. RR-85-24). Princeton, NJ: ETS.

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

Muthén, B., du Toit, S. H. C., & Spisic, D. (1997). *Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes*. Unpublished manuscript.

Pinheiro, P. C., & Bates, D. M. (1995). Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of Computational and Graphical Statistics, 4,* 12–35.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danish Institute for Educational Research.

Rijmen, F., Tuerlinckx, F., De Boeck, P., & Kuppens, P. (2003). A nonlinear mixed model framework for item response theory. *Psychological Methods, 8,* 185–205.

Rijmen, F., Vansteelandt, K., & De Boeck, P. (in press). Latent class models for diary method data: Parameter estimation by local computations. *Psychometrika*.

Rodriguez, G., & Goldman, N. (1995). An assessment of estimation procedures for multilevel models with binary response. *Journal of the Royal Statistical Society, Series A, 158,* 73–89.

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal and structural equation models*. Boca Raton, FL: Chapman & Hall/CRC.

Stuart, A. (1958). Equally correlated variates and the multinormal integral. *Journal of the Royal Statistical Society, Series B, 20,* 373–378.

Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, *52*, 393–408.

Yannakakis, M. (1981). Computing the minimum fill-in is NP-complete. *SIAM Journal on Algebraic and Discrete Methods*, *2*, 77–79.

Appendix A

Identification Restrictions for the Bi-factor and the Testlet Model

The bi-factor model was defined in Equation 2 as

$$g\left(\pi_j\right) = \alpha_{jg}\theta_g + \alpha_{jk}\theta_k + \beta_j \tag{A1}$$

Identification Restrictions

There are three types of identification restrictions:

1. K + 1 restrictions to fix the origin of the general and K item subset specific dimensions. In a MML framework, this can achieved through the restrictions $E\left(\boldsymbol{\theta}\right) = \mathbf{0}$.

2. K + 1 restrictions to fix the scale of the general and K item subset specific dimensions. In a MML framework, this can achieved through the restrictions $\sigma_g = 1, \sigma_k = 1, k = 1, ..., K$.

3. K restrictions to deal with the rotational invariance. Algebraically, rotational invariance can be shown as follows. Assuming that the location and scale are already fixed through $E\left(\boldsymbol{\theta}\right) = \mathbf{0}$ and $\sigma_g = 1, \sigma_k = 1, k = 1, ..., K$, we rewrite Equation A1 ,

$$\begin{aligned} g\left(\pi_j\right) &= \alpha_{jg}\theta_g + \alpha_{jk}\theta_k + \beta_j \\ &= \alpha_{jg}\theta_g + \alpha_{jk}\theta_k + \beta_j + c_k\alpha_{jk}\theta_g - c_k\alpha_{jk}\theta_g \\ &= \left(\alpha_{jg} - c_k\alpha_{jk}\right)\theta_g + \alpha_{jk}\left(\theta_k + c_k\theta_g\right) + \beta_j \\ &= \alpha^*_{jg}\theta_g + a_k\alpha_{jk}\left(\theta_k + c_k\theta_g\right)/a_k + \beta_j \\ &= \alpha^*_{jg}\theta_g + \alpha^*_{jk}\theta^*_k + \beta_j, \end{aligned} \tag{A2}$$

with $\alpha^*_{jg} = \left(\alpha_{jg} - c_k\alpha_{jk}\right), \alpha^*_{jk} = a_k\alpha_{jk}$, $\theta^*_k = \left(\theta_k + c_k\theta_g\right)/a_k$, and $a_k^2 = 1 + c_k^2 + 2c_k\rho_{gk}$. a_k^2 is the variance of $\left(\theta_k + c_k\theta_g\right)$, and is determined by the subset specific constant c_k and the correlation ρ_{gk} between θ_g and θ_k. Dividing $\left(\theta_k + c_k\theta_g\right)$ by its standard deviation a_k ensures that the variance of θ^*_k is 1. We obtained again the expression for a bi-factor model, but now expressed

in a different basis. The rotational invariance can be fixed by setting the correlations ρ_{gk} to zero for all *k*.

The Testlet Model

Bradlow et al. (1999) formulated the testlet model in a Bayesian framework. Its analogue in a maximum likelihood framework for a model without guessing parameter can be formulated as

$$g\left(\pi_j\right) = \alpha_{jg}\left(\theta_g + \theta_k\right) + \beta_j, \tag{A3}$$

To identify the model, $E\left(\boldsymbol{\theta}\right) = \mathbf{0}$, $\sigma_g = 1$, and $\rho_{gk} = 0$ for all *k* as in the bi-factor model. In contrast, the variances σ_k^2 are free parameters. The restrictions $\rho_{gk} = 0$ fix the rotational invariance of the testlet model. Similarly to the bi-factor model, the rotational invariance of the testlet model can be shown algebraically as follows:

$$\begin{aligned} g\left(\pi_j\right) &= \alpha_{jg}\theta_g + \alpha_{jg}\theta_k + \beta_j \\ &= \alpha_{jg}\theta_g + \alpha_{jg}\theta_k + c_k\alpha_{jg}\theta_g - c_k\alpha_{jg}\theta_g + \beta_j \\ &= \alpha_{jg}\left(1-c_k\right)\theta_g + \alpha_{jg}\left(1-c_k\right)\left(\theta_k + c_k\theta_g\right)/\left(1-c_k\right) + \beta_j \\ &= \alpha_{jg}^{**}\theta_g + \alpha_{jg}^{**}\theta_k^{**} + \beta_j, \end{aligned} \tag{A4}$$

where $\alpha_{jg}^{**} = \alpha_{jg}\left(1-c_k\right)$, and $\theta_k^{**} = \left(\theta_k + c_k\theta_g\right)/\left(1-c_k\right)$.

Expressing Equation A3 as a model in which all latent variables have unit variance, Li et al. (2006) showed that the testlet model is a bi-factor model in which the loadings on the specific dimension are proportional to the loadings on the general dimension within each testlet,

$$\begin{aligned} g\left(\pi_j\right) &= \alpha_{jg}\left(\theta_g + \theta_k\right) + \beta_j \\ &= \alpha_{jg}\theta_g + \alpha_{jg}\sigma_k\theta_k / \sigma_k + \beta_j \\ &= \alpha_{jg}\theta_g + \alpha_{jk}^{+}\theta_k^{+} + \beta_j, \end{aligned} \tag{A5}$$

with $\alpha_{jk}^{+} = \alpha_{jg}\sigma_k$ and $\theta_k^{+} = \theta_k / \sigma_k$ so that $\sigma_{k^*} = 1$.

Appendix B
Computations for the Bi-Factor Model

First, it is shown how the complete data likelihood can be rewritten according to Equation 13. From Equations 9 and 10,

$$P(\mathbf{y},\mathbf{z}) = P(\mathbf{z})P(\mathbf{y}|\mathbf{z}) = P(z_g)\prod_k P(z_k|z_g)P(\mathbf{y}_k|z_g,z_k) \tag{B1}$$

where $P(\mathbf{y}_k|z_g,z_k)$ is the vector of responses to the items of testlet *k*. Then,

$$\begin{aligned} P(\mathbf{y},\mathbf{z}) &= \frac{\left[P(z_g)\right]^{K-1} P(z_g)\prod_k P(z_k|z_g)P(\mathbf{y}_k|z_g,z_k)}{\left[P(z_g)\right]^{K-1}} \\ &= \frac{\prod_k \left[P(z_g)\right]^{K} P(z_k|z_g)P(\mathbf{y}_k|z_g,z_k)}{\left[P(z_g)\right]^{K-1}} \\ &= \frac{\prod_k P(z_k,z_g)P(\mathbf{y}_k|z_g,z_k)}{\left[P(z_g)\right]^{K-1}} \\ &= \frac{\prod_k P(\mathbf{y}_k,z_g,z_k)}{\left[P(z_g)\right]^{K-1}}, \end{aligned}$$

resulting in Equation 13.

It is now shown in detail how the junction tree can be used to obtain the marginal posterior probabilities $(\theta_g,\theta_k|\mathbf{y})$ for all *k*. A first step is to initialize the clique and separator potentials.

For the bi-factor model, there are *K* cliques corresponding to the sets ($\mathbf{y}_k$, z_k, z_g), and *K*-1 separators that all consist of the single variable z_g. All separator potentials are initialized with a value of 1 for all possible realizations of z_g. The clique potentials are initialized as follows. For each of the factors in Equation B1, a clique is selected that contains all its variables. Specifically, the factors $P(\mathbf{y}_k|z_g,z_k)$ and $P(z_k|z_g)$ are associated with clique ($\mathbf{y}_k$, z_k, z_g), for k =1,…, *K*.

$P\left(z_g\right)$ can be associated with any clique since z_g occurs in all cliques. Here, it is associated to clique ($\mathbf{y}_K$, z_K, z_g). Next, the factors of Equation B1 are multiplied into the cliques they are associated with. Hence, for the cliques ($\mathbf{y}_k$, z_k, z_g), k =1,…, K -1, the potential for every possible realization of z_k and z_g is defined as

$$\Psi_{\mathbf{y}_k, z_k, z_g} = P\left(z_k \middle| z_g\right) P\left(\mathbf{y}_k \middle| z_g, z_k\right) \tag{B2}$$

Both factors are computed straightforwardly given provisional estimates for the parameters that characterize $P\left(\mathbf{y}_k \middle| z_g, z_k\right)$ and $P\left(z_k \middle| z_g\right)$, and the observed response pattern $\mathbf{y}_k$. Obviously, the thus defined clique potentials will be specific for each observed response pattern $\mathbf{y}_k$.

For clique ($\mathbf{y}_K$, z_K, z_g), the potential for every possible realization of z_K and z_G is defined as

$$\Psi_{\mathbf{y}_K, z_K, z_g} = P\left(z_g\right) P\left(z_K \middle| z_g\right) P\left(\mathbf{y}_K \middle| z_g, z_K\right) \tag{B3}$$

It is easily verified that the thus defined potentials obey they property stated in Equation 14 that the joint probability function can be expressed as the product of all clique marginals over the product of all separator marginals:

$$\begin{aligned} P(\mathbf{y}, \mathbf{z}) = \mathrm{P}(\mathbf{x}) &= \frac{\prod_C \psi(\mathbf{x}_C)}{\prod_S \psi(\mathbf{x}_S)} \\ &= \prod_C \psi(\mathbf{x}_C) / 1 \\ &= \prod_{k=1}^{K} \psi\left(y_k, z_k, z_g\right) \\ &= P\left(z_g\right) \prod_{k=1}^{K} P\left(z_k \middle| z_g\right) P\left(\mathbf{y}_k \middle| z_g, z_k\right) \end{aligned} \tag{B4}$$

Finally it is shown how the marginal posterior clique probabilities are obtained by applying the two-phase schedule of flows. During this process, the clique and state potentials are altered but the property that the joint probability function can be expressed as the product of all clique marginals over the product of all separator marginals is preserved. Without loss of generality, the schedule is illustrated for K = 4.

Using the junction tree of Figure 4, the first flow is passed from clique ($\mathbf{y}_4$, z_4, z_g) to clique ($\mathbf{y}_3$, z_3, z_g) as described in Equation 15. First, the potential of the separator in between the two cliques is updated. Denoting the two cliques by C_4 and C_3 respectively, and the separator by S_{34},

$$\begin{aligned}\psi^*\left(\mathbf{x}_{S_{34}}\right) &= \sum_{C_4 \setminus S_{34}} \psi\left(\mathbf{x}_{C_4}\right) \\ &= \sum_{z_4, \mathbf{y}_4} P\left(z_g\right) P\left(z_4 \middle| z_g\right) P\left(\mathbf{y}_4 \middle| z_g, z_4\right) \\ &= P\left(\mathbf{y}_4, z_g\right),\end{aligned}$$

$$\lambda\left(\mathbf{x}_{S_{34}}\right) = \frac{\psi^*\left(\mathbf{x}_{S_{34}}\right)}{\psi\left(\mathbf{x}_{S_{34}}\right)} = \frac{P\left(\mathbf{y}_4, z_g\right)}{1}, \tag{B5}$$

and

$$\psi^*\left(\mathbf{x}_{C_3}\right) = \psi\left(\mathbf{x}_{C_3}\right)\lambda\left(\mathbf{x}_{S_{34}}\right) = P\left(z_3 \middle| z_g\right) P\left(\mathbf{y}_3 \middle| z_g, z_3\right) P\left(\mathbf{y}_4, z_g\right).$$

Note that $\mathbf{y}_4$, as explained before, is treated for any given person as a merged node with only one possible realization, which is the observed response pattern of the person. It is easily verified that Equation 13 still holds,

$$\begin{aligned}&\frac{\prod_C \psi\left(\mathbf{x}_C\right)}{\prod_S \psi\left(\mathbf{x}_S\right)} \\ &= P\left(z_1 \middle| z_g\right) P\left(\mathbf{y}_1 \middle| z_g, z_1\right) P\left(z_2 \middle| z_g\right) P\left(\mathbf{y}_2 \middle| z_g, z_2\right) P\left(z_3 \middle| z_g\right) P\left(\mathbf{y}_3 \middle| z_g, z_3\right) \times \\ &\quad \frac{P\left(\mathbf{y}_4, z_g\right) P\left(z_4 \middle| z_g\right) P\left(\mathbf{y}_4 \middle| z_g, z_4\right) P\left(z_g\right)}{1 \times 1 \times P\left(\mathbf{y}_4, z_g\right)} \\ &= P\left(z_1 \middle| z_g\right) P\left(\mathbf{y}_1 \middle| z_g, z_1\right) P\left(z_2 \middle| z_g\right) P\left(\mathbf{y}_2 \middle| z_g, z_2\right) P\left(z_3 \middle| z_g\right) P\left(\mathbf{y}_3 \middle| z_g, z_3\right) P\left(z_4 \middle| z_g\right) P\left(\mathbf{y}_4 \middle| z_g, z_4\right) P\left(z_g\right) \\ &= P\left(\mathbf{y}, \mathbf{z}\right).\end{aligned}$$

The second flow is passed from clique C_3 to clique C_2, over separator S_{23}, where the potentials for C_3 have been updated according to Equation B5,

$$
\begin{aligned}
\psi^*\left(\mathbf{x}_{S_{23}}\right) &= \sum_{C_3 \setminus S_{23}} \psi\left(\mathbf{x}_{C_3}\right) \\
&= \sum_{z_3, \mathbf{y}_3} P\left(z_3 \middle| z_g\right) P\left(\mathbf{y}_3 \middle| z_g, z_3\right) P\left(\mathbf{y}_4, z_g\right) \\
&= P\left(\mathbf{y}_4, z_g\right) P\left(\mathbf{y}_3 \middle| z_g\right) \\
&= P\left(\mathbf{y}_3 \middle| z_g\right) P\left(\mathbf{y}_4 \middle| z_g\right) P\left(z_g\right) \\
&= P\left(\mathbf{y}_3, \mathbf{y}_4, z_g\right)
\end{aligned}
$$

using the conditional independence of $\mathbf{y}_3$ and $\mathbf{y}_4$, given z_g for the last equality;

$$
\lambda\left(\mathbf{x}_{S_{23}}\right) = \frac{\psi^*\left(\mathbf{x}_{S_{23}}\right)}{\psi\left(\mathbf{x}_{S_{23}}\right)} = \frac{P\left(\mathbf{y}_3, \mathbf{y}_4, z_g\right)}{1},
$$

and

$$
\psi^*\left(\mathbf{x}_{C_2}\right) = \psi\left(\mathbf{x}_{C_2}\right) \lambda\left(\mathbf{x}_{S_{23}}\right) = P\left(z_2 \middle| z_g\right) P\left(\mathbf{y}_2 \middle| z_g, z_2\right) P\left(\mathbf{y}_3, \mathbf{y}_4, z_g\right).
$$

The next flow concludes the collection phase, and results in the following potentials

$$
\begin{aligned}
\psi^*\left(\mathbf{x}_{S_{12}}\right) &= \sum_{C_2 \setminus S_{12}} \psi\left(\mathbf{x}_{C_2}\right) \\
&= \sum_{z_2, \mathbf{y}_2} P\left(z_2 \middle| z_g\right) P\left(\mathbf{y}_2 \middle| z_g, z_2\right) P\left(\mathbf{y}_3, \mathbf{y}_4, z_g\right) \\
&= P\left(\mathbf{y}_3, \mathbf{y}_4, z_g\right) P\left(\mathbf{y}_2 \middle| z_g\right) \\
&= P\left(\mathbf{y}_2 \middle| z_g\right) P\left(\mathbf{y}_3, \mathbf{y}_4 \middle| z_g\right) P\left(z_g\right) \\
&= P\left(\mathbf{y}_2, \mathbf{y}_3, \mathbf{y}_4, z_g\right)
\end{aligned}
$$

using the conditional independence of $\mathbf{y}_2$, $\mathbf{y}_3$ and $\mathbf{y}_4$, given z_g for the last equality;

$$
\lambda\left(\mathbf{x}_{S_{12}}\right) = \frac{\psi^*\left(\mathbf{x}_{S_{12}}\right)}{\psi\left(\mathbf{x}_{S_{12}}\right)} = \frac{P\left(\mathbf{y}_2, \mathbf{y}_3, \mathbf{y}_4, z_g\right)}{1},
$$

and

$$
\psi^*\left(\mathbf{x}_{C_1}\right) = \psi\left(\mathbf{x}_{C_1}\right) \lambda\left(\mathbf{x}_{S_{12}}\right) = P\left(z_1 \middle| z_g\right) P\left(\mathbf{y}_1 \middle| z_g, z_1\right) P\left(\mathbf{y}_2, \mathbf{y}_3, \mathbf{y}_4, z_g\right).
$$

It is left to the reader to verify that Equation 13 remains valid after the second and third propagation of flows.

From the conditional independence given z_g between $\mathbf{y}_2$, $\mathbf{y}_3$, and $\mathbf{y}_4$ on the one hand and $\mathbf{y}_1$ and z_1 on the other hand,

$$\begin{aligned}\psi^*\left(\mathbf{x}_{C_1}\right) &= P\left(z_1 \middle| z_g\right) P\left(\mathbf{y}_1 \middle| z_g, z_1\right) P\left(\mathbf{y}_2, \mathbf{y}_3, \mathbf{y}_4 \middle| z_g\right) P\left(z_g\right) \\ &= P\left(\mathbf{y}, z_g, z_1\right)\end{aligned} \tag{B6}$$

Hence, the clique potentials for C_1 are proportional to the marginal posterior probabilities of its latent variables z_1 and z_g. The exact posterior probabilities for the latent variables z_1 and z_g are obtained through a renormalization of the clique potentials,

$$P\left(z_g, z_1 \middle| \mathbf{y}\right) = \frac{P\left(\mathbf{y}, z_g, z_1\right)}{\sum_{z_g, z_1} P\left(\mathbf{y}, z_g, z_1\right)} \tag{B7}$$

To obtain the marginal posterior clique probabilities for the remaining cliques, the second phase of flows has to be propagated. In this distribution phase, cliques are updated in reversed order. So, first a flow is passed from clique C_1 to clique C_2, over separator S_{12},

$$\begin{aligned}\psi^*\left(\mathbf{x}_{S_{12}}\right) &= \sum_{C_1 \setminus S_{12}} \psi\left(\mathbf{x}_{C_1}\right) \\ &= \sum_{z_1, \mathbf{y}_1} P\left(\mathbf{y}, z_g, z_1\right) \\ &= P\left(\mathbf{y}, z_g\right),\end{aligned}$$

$$\lambda\left(\mathbf{x}_{S_{12}}\right) = \frac{\psi^*\left(\mathbf{x}_{S_{12}}\right)}{\psi\left(\mathbf{x}_{S_{12}}\right)} = \frac{P\left(\mathbf{y}, z_g\right)}{P\left(\mathbf{y}_2, \mathbf{y}_3, \mathbf{y}_4, z_g\right)} = P\left(\mathbf{y}_1 \middle| z_g\right),$$

and

$$\psi^*\left(\mathbf{x}_{C_2}\right) = \psi\left(\mathbf{x}_{C_2}\right)\lambda\left(\mathbf{x}_{S_{12}}\right) = P\left(z_2 \middle| z_g\right) P\left(\mathbf{y}_2 \middle| z_g, z_2\right) P\left(\mathbf{y}_3, \mathbf{y}_4, z_g\right) P\left(\mathbf{y}_1 \middle| z_g\right) = P\left(\mathbf{y}, z_2, z_g\right).$$

Now, the clique potentials for C_2 are proportional to the marginal posterior probabilities of its latent variables z_2 and z_g.

For the last two flows, the clique and separator potentials are updated as, respectively,

$$\psi^*\left(\mathbf{x}_{S_{23}}\right) = \sum_{C_2 \setminus S_{23}} \psi\left(\mathbf{x}_{C_2}\right)$$
$$= \sum_{z_2, \mathbf{y}_2} P\left(\mathbf{y}, z_g, z_2\right)$$
$$= P\left(\mathbf{y}, z_g\right),$$
$$\lambda\left(\mathbf{x}_{S_{23}}\right) = \frac{\psi^*\left(\mathbf{x}_{S_{23}}\right)}{\psi\left(\mathbf{x}_{S_{23}}\right)} = \frac{P\left(\mathbf{y}, z_g\right)}{P\left(\mathbf{y}_3, \mathbf{y}_4, z_g\right)} = P\left(\mathbf{y}_1, \mathbf{y}_2 \middle| z_g\right),$$

and

$$\psi^*\left(\mathbf{x}_{C_3}\right) = \psi\left(\mathbf{x}_{C_3}\right)\lambda\left(\mathbf{x}_{S_{23}}\right) = P\left(z_3 \middle| z_g\right) P\left(\mathbf{y}_3 \middle| z_g, z_3\right) P\left(\mathbf{y}_4, z_g\right) P\left(\mathbf{y}_1, \mathbf{y}_2 \middle| z_g\right) = P\left(\mathbf{y}, z_3, z_g\right),$$

and

$$\psi^*\left(\mathbf{x}_{S_{34}}\right) = \sum_{C_3 \setminus S_{34}} \psi\left(\mathbf{x}_{C_3}\right)$$
$$= \sum_{z_3, \mathbf{y}_3} P\left(\mathbf{y}, z_3, z_g\right)$$
$$= P\left(\mathbf{y}, z_g\right),$$
$$\lambda\left(\mathbf{x}_{S_{34}}\right) = \frac{\psi^*\left(\mathbf{x}_{S_{34}}\right)}{\psi\left(\mathbf{x}_{S_{34}}\right)} = \frac{P\left(\mathbf{y}, z_g\right)}{P\left(\mathbf{y}_4, z_g\right)} = P\left(\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3 \middle| z_g\right),$$

and

$$\psi^*\left(\mathbf{x}_{C_4}\right) = \psi\left(\mathbf{x}_{C_4}\right)\lambda\left(\mathbf{x}_{S_{34}}\right) = P\left(z_g\right) P\left(z_4 \middle| z_g\right) P\left(\mathbf{y}_4 \middle| z_g, z_4\right) P\left(\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3 \middle| z_g\right) = P\left(\mathbf{y}, z_4, z_g\right).$$

Posterior clique probabilities are obtained by normalizing the clique potentials analogously to Equation B7.

It is again easily verified that the clique and separator potentials have reached equilibrium: no passage of a flow between any two cliques will alter the potentials of the separator or receiving clique. All the separator potentials are equal to $P\left(\mathbf{y}, z_g\right)$, which is also obtained after marginalizing any of the K clique potentials over z_k. Hence, the update factors $\lambda\left(\mathbf{x}_{S_{kl}}\right) = \frac{\psi^*\left(\mathbf{x}_{S_{kl}}\right)}{\psi\left(\mathbf{x}_{S_{kl}}\right)}$ are all equal to one, and the potentials are unaltered after passing a flow.