Log-Linear and Log-Multiplicative Association Models for Categorical Data

Carolyn J. Anderson, Maria Kateri, & Irini Moustaki

20 January 2021

1 Introduction

Log-linear models are useful for determining whether dependencies exist between categorical variables; however, when there are interactions, the nature of the association needs to be described. Unfortunately, the descriptions can be challenging especially when the categorical variables have a large number of categories and/or the table is high-dimensional. To fully capture the dependency structure would require computing all possible conditional odds ratios (ORs), which in the case of large tables is often not very enlightening. Association models (AMs) provide a solution to this problem by imposing special structures on the interactions between categorical variables thus leading to more parsimonious models that facilitate insightful interpretation of interactions. A central characteristic of all AMs is that interactions are represented by multiplicative terms.

Basically, AMs have a special multiplicative structure imposed on some or all interaction terms of a standard log-linear model. The parameters of the multiplicative terms have high interpretative value and reduce the number of parameters needed to describe the nature and strength of interactions. In some AMs, the model remains log-linear, while others are log-multiplicative, i.e. non-linear in their parameters. ORs, which play a predominant role in log-linear model (and AMs) analysis and interpretation, are functions of the parameters introduced in an AM, and plots of these parameters give pictures of the features and structure of the associations.

In addition to providing visual plots representing associations between variables, the models themselves have graphical representations. The graphics greatly aid in communication because they represent scientific content and in some cases underlying processes. To differentiate between models and for clarity, we advocate that models should be presented both graphically and algebraically. Many of the AMs that we discuss have the same basic graphical representation. The algebraic representations without also using their graphical representations tend to cloud the relationships between models, but the algebraic form provides details that may be lacking in the graphical representation.

AMs developed for the analysis of categorical variables have been derived from numerous frameworks. They provide useful structured representations of interactions among variables allowing a special treatment for ordinal variables. AMs have been developed either directly for specific modeling purposes (e.g. contingency table analysis) or have been arisen through a theorized underlying processes (e.g. item response theory (IRT)). They have been proposed over different fields and sub-fields, often independently, which has lead to a fractured literature on the subject. It is evident that AMs offer a powerful and flexible platform for diverse areas of applications. The class of models that we generically refer to as AMs consists of many models with different names but of the same general form. These include, among others, linear by linear models (LL), row models (R), column models (C), uniform models (U), and M-dimensional row-column AMs ($RC(M)$) ([28, 30, 31]), generalized additive effects and multiplicative interaction model used to study plant genetics ([21]), graphical latent variable models for categorical data ([3]), IRT models ([4, 5, 36, 49]), Ising model ([45]), generalized Newton's law of gravity ([17]), network psychometrics ([49]), fused graphical models ([14]), formative response models, distance based models ([59, 19, 17, 18]), conditional multinomial models ([2, 4, 35]), and discretized multivariate normal distributions ([29, 7, 58, 65, 66]). Worth mentioning are efforts to build bridges between different fields and further explore their utility, such as connecting IRT to log-linear models ([43, 42]), and to log-multiplicative interactions ([4, 5, 49], and others).

AMs are closely linked to log-linear models. For this, we start in Section 2 with a brief presentation of log-linear models upon which we build the family of AMs for two-way tables (i.e., LL, R, C, and $RC(M)$ models). Many of the basic features of these AMs for two-way tables carry over to models for more variables and more complex situations. We subsequently review statistical graphical representations of log-linear and AMs and use this as a step toward high-dimensional generalizations of the $RC(M)$ model. Subsequently, we present high-dimensional models in detail, including estimation and the equivalence with IRT models. Some discussion on testing and model selection under the pseudo-likelihood framework is given. To illustrate the use and benefits afforded by AMs, two examples are given: (i) the analysis of a (16×6) table by models for two-way tables, and (ii) responses to 42 four category items from three correlated scales by models for high-dimensional tables. Lastly, we follow with a discussion that reflects on the material presented in the chapter and provides future research directions.

2 Preliminaries

Throughout this chapter, we assume that we have I items (or variables), $\boldsymbol{Y} = (Y_1, \ldots, Y_I)'$, measured on n subjects. Let $\boldsymbol{Y} = (Y_1, \ldots, Y_I)'$ be a random response vector where $Y_i \in \mathcal{C}_i = \{1, \ldots, J_i\}$, and let $\boldsymbol{y}_s = \{y_{1s}, \ldots, y_{Is}\}$ be observed responses for subject $s \in \{1, \ldots, n\}$, i.e. $y_{is} = j_i \in \mathcal{C}_i$, for $i = 1, \ldots, I$. Furthermore, assume that there exists a set of M latent variables, $\boldsymbol{\Theta} = \{\Theta_1, \ldots, \Theta_M\}$ and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_M)'$ is a realization of them. We restrict to models with $M \leq I$ while more complex models with $M > I$ are possible.

In a contingency table representation, data form an I-dimensional table, produced by cross-classifying the subjects' responses on all items, having cell entries $n_{j_1,\ldots,j_I}$, the frequencies of subjects with responses $\boldsymbol{y} = (j_1, \ldots, j_I)'$, where $j_i \in \mathcal{C}_i$, $i = 1, \ldots, I$. In this set-up, the subject index s is suppressed, but will be needed later in the chapter. Obviously, $\sum_{j_1,\ldots,j_I} n_{j_1,\ldots,j_I} = n$ and the underlying distribution, depending on the study design, can be a multinomial $\mathcal{M}(n, \boldsymbol{\pi})$ with probability table $\boldsymbol{\pi} = \{\pi_{\boldsymbol{y}}\} = \{\pi_{j_1,\ldots,j_I}\}$, or independent Poisson distributions $\mathcal{P}(m_{\boldsymbol{y}})$ in every cell, where $m_{\boldsymbol{y}}$ is the predicted or expected cell frequency. Given the sample size n, the expected cell frequencies equal $m_{\boldsymbol{y}} = m_{j_1,\ldots,j_I} = n\pi_{j_1,\ldots,j_I}$.

2.1 Hierarchical log-linear models

Contingency tables are traditionally analyzed by hierarchical log-linear models, expressed in terms of cell probabilities or expected cell frequencies[1]. Here

[1] Note that "hiearchical" refers to different models (e.g., linear regression, multi-level models, log-linear models). In this chapter "hierarchical" refers to models where all lower order terms that comprise an interaction are included in the model.

we shall model the cell probabilities. In the case of many items, the corresponding contingency table is high-dimensional and is often extremely sparse, which causes inferential and estimation problems. Usually lower order interactions (even only two factor interactions) are sufficient to model the response patterns and the corresponding marginal tables are not sparse. The two-way marginal tables are sufficient statistics for estimating two-factor interactions. Thus, the response probabilities $P(\boldsymbol{y})$ can be modeled, for example, by a log-linear model with all two-factor interactions,

$$\log P(\boldsymbol{y}) = \log(\pi_{\boldsymbol{y}}) = \lambda + \sum_{i=1}^{I} \lambda_{j_i}^{[i]} + \sum_{\substack{i,k \\ i<k}}^{I} \lambda_{j_i j_k}^{[ik]}, \quad j_i \in \mathcal{C}_i,\ j_k \in \mathcal{C}_k\ , \tag{1}$$

where λ ensures that probabilities sum to 1, $\lambda_{j_i}^{[i]}$ is the marginal (main) effect term for the category j_i of the i-th item, and $\lambda_{j_i j_k}^{[ik]}$ is the interaction term between the levels j_i and j_k of the i-th and k-th items, respectively.

Identification constraints required on parameters in (1) to obtain parameter estimates. Common constraints are setting the first category to zero, i.e.,

$$\lambda_1^{[i]} = \lambda_{11}^{[ik]} = \lambda_{1j_k}^{[ik]} = \lambda_{j_i 1}^{[ik]} = 0, \quad \text{for all possible values of } i, k, j_i, j_k. \tag{2}$$

Alternative constraints set last category to zero or set the sum of over categories equal to zero.

In some applications, we are only interested in the relationship between variables; however, in other modeling applications, we make a distinction between response and explanatory variables. Regardless of the situation, the model for way tables is the same. For example, when modeling response behavior, explanatory variables, such as demographic ones, may be present that may be categorical or on an interval scale. In such cases, log-linear models of type (1) can be employed that incorporate main effects for the explanatory variables and interactions among explanatory variables and that response variable.

The simplest case of having just two items reduces (1) to

$$\log P(\boldsymbol{y}) = \log(\pi_{\boldsymbol{y}}) = \lambda + \lambda_{j_1}^{[1]} + \lambda_{j_2}^{[2]} + \lambda_{j_1 j_2}^{[12]}, \qquad j_1 \in \mathcal{C}_1,\ \ j_2 \in \mathcal{C}_2\ . \tag{3}$$

In the log-linear modeling framework, log-linear models with interactions may have difficultly dealing with sparse tables that include zero cell frequencies. Using (3) as an example, if a cell of the $[ik]$ marginal table has a zero frequency, the corresponding parameter $\lambda^{[12]}_{j_1 j_2}$ for that cell cannot be estimated. Necessary and sufficient conditions for the existence of the maximum likelihood estimates (MLE) of the log-linear model parameters, with a focus on the role of sampling zeros in the observed table, are provided by [24]. Fitted values(MLE) for (3) can be obtained using iterative proportional fitting, but we cannot fully describe the interaction because some odds ratios are not estimable. This is not the case for unsaturated AMs. For example, the $RC(M)$ model described in the next section encounters no problems if there are zeros in a (marginal) table. Only the univariate marginals need to be non-zero. The ability of AMs to deal with sparse tables becomes especially important when we have high-dimensional tables.

3 Association models for two-way tables

Model (3) is saturated (i.e. has 0 degrees of freedom). For a $J_1 \times J_2$ table, in the classical log-linear modeling framework, there are no models in between the saturated model and that of independence, which has $(J_1 - 1)(J_2 - 1)$ degrees of freedom. A class of non-saturated models is derived by imposing a structure or restrictions on the interaction parameters of a log-linear model which requires fewer parameters. Fewer parameters leads to more parsimonious models that fill the gap between the two extreme models (independence and saturated) and at the same time, offer sound interpretation. These models are known as dependency or *association models* (AMs, often called Goodman's AMs) and are based on the concept of assigning scores or estimating scale values for the categories of the classification variables (items).

For a two-dimensional table, association models are of the form

$$\log P(\boldsymbol{y}) = \log(\pi_{\boldsymbol{y}}) = \lambda + \lambda^{[1]}_{j_1} + \lambda^{[2]}_{j_2} + \sigma^2 \nu_{1j_1} \nu_{2j_2} \ , \qquad (4)$$

for $j_1 \in \mathcal{C}_1$, $j_2 \in \mathcal{C}_2$, where $\boldsymbol{\nu}_1 = (\nu_{11}, \ldots, \nu_{1J_1})'$ and $\boldsymbol{\nu}_2 = (\nu_{21}, \ldots, \nu_{2J_2})'$ are scores corresponding to the rows and columns of the contingency table, respectively, and σ^2 is an intrinsic association parameter. Notice that in the literature on association models, the association parameter is usually denoted by ϕ and, for row and column scores that are monotone in the same direction, the sign of ϕ indicates the direction of the underlying association. The model is invariant under linear transformation of the row and column scores and the direction of the scores are generally set such that σ^2 is positive. An important point is that σ^2 reflects the strength of the association and the row and column scores reflect the structure.

The row and columns scores, $\boldsymbol{\nu}_1$ and $\boldsymbol{\nu}_2$, respectively, can be fixed (known) or parameters to be estimated. The simplest association model that considers both of them fixed, has just one parameter more than the independence model and is known as the *linear by linear* (LL) model. If additionally the scores are equidistant for successive row and column categories, then under this specific LL model all local odds ratios, which are odds ratios between adjacent rows and columns, are equal. This is called as the *uniform* (U) association model. When the row scores are fixed and column scores are estimated, the model is called the column effect (C) model. The row effect (R) model is defined analogously. Models LL, U, C and R are all log-linear. When both row and column scores are parameters to be estimated, model (4) becomes the *multiplicative row-column* effect (RC) model and no longer has a log-linear structure.

The main effects parameters of model (4) satisfy the corresponding identifiability constraints in (2) while the scores, whenever they are parameters, satisfy

$$\sum_{j_1=1}^{J_1} \nu_{1j_1} = \sum_{j_2=1}^{J_2} \nu_{2j_2} = 0 \quad \text{and} \quad \sum_{j_1=1}^{J_1} \nu_{1j_1}^2 = \sum_{j_2=1}^{J_2} \nu_{2j_2}^2 = 1 \ . \tag{5}$$

Since model (4) is invariant under linear transformations of the scores, for comparability, also in the case of fixed or known scores, scores are transformed to fulfill (5). The intrinsic association parameter in (4) is redundant and can be set $\sigma^2 = 1$, abandoning the second set of constraints in (5), as given by Goodman [28].

An extension of the RC model is the multidimensional row-column or $RC(M)$ association model, which includes multiple sets of scores for each item. It is defined as

$$\log P(\boldsymbol{y}) = \log(\pi_{\boldsymbol{y}}) = \lambda + \lambda_{j_1}^{[1]} + \lambda_{j_2}^{[2]} + \sum_{m=1}^{M} \sigma_m^2 \nu_{1j_1m} \nu_{2j_2m} \ , \tag{6}$$

for $M \in \{1, \ldots, M^*\}$, $M^* = \min(I, J) - 1$, where scores and association parameters are assigned to each dimension m, with $\sigma_1^2 \geq \ldots \geq \sigma_M^2 \geq 0$, reflecting that the strength of association accounted for each dimension m is decreasing in m. Constraints (5) hold for the scores on every dimension, and additionally, scores on different dimensions are orthogonal to each other, i.e.,

$$\sum_{j_1=1}^{J_1} \nu_{1j_1m}\nu_{1j_1m'} = \sum_{j_2=1}^{J_2} \nu_{2j_2m}\nu_{2j_2m'} = 0, \quad \text{for all } m \neq m'. \tag{7}$$

Constraints (5) and (7) are the most commonly used ones, but are not the only possible ones. Model (6) has $(I-M-1)(J-M-1)$ degrees of freedom (df) and note that $RC(1) = RC$ and $RC(M^*)$ is an equivalent expression of the saturated log-linear model given in (3).

The AMs for two-way tables presented in this section can be extended in a straight forward manner to tables of higher dimensions and we will point out how the models for high-way tables are the same and different from the $RC(M)$ association models. Before considering the high-way case, we discuss estimation and present an example for a 2-way table.

3.1 Estimation and Goodness of Fit of AMs

Maximum likelihood estimation of AMs is the most commonly used method to fit the models to data and we focus our attention to ways to do this in **R** ([60]). Models that are log-linear can be fit through packages for generalized linear models (GLM), in particular the `glm` function. Models that are non-linear in their parameters, like the $RC(M)$ model introduced above, require special packages for their implementation, such as the `gnm` package of Turner and Firth [61] or the `VGAM` of Yee [67]. The implementation of association models via `gnm` is extensively illustrated in Section 6.6 of Kateri [40], while functions for fitting specific AMs are provided in the web-appendix of [40]. Here, we fit AMs using maximum likelihood estimation as implemented in the R ([60], version 4.0.0) package `logmulti` ([11]), which is a wrapped for the more general `gnm` package.

Goodness of fit (GoF) of AMs can be tested by the standard GoF tests for contingency table models, i.e., the likelihood ratio statistic (G^2) or the Pearson's X^2. Since the values of the G^2 and X^2 statistics are strongly influenced from the sample size, we consider two additional statistics that give the practical significance and a more intuitive sense of GoF. The value of G^2 from independence can be thought of as a measure of the amount of dependency in the data. The percent of association accounted for by a model equals

$$\frac{(G^2_{ind} - G^2_{model})}{G^2_{ind}} \times 100,$$

where G^2_{ind} and G^2_{model} are the likelihood ratio test statistics for the model of independence and the model of interest. A second index, the dissimilarity index (D), equals the proportion of the data that would have to be moved from one cell to another for the model to fit perfectly. The dissimilarity index can be computed using frequencies or proportions; namely,

$$\frac{\sum_i |n_i - \hat{m}_i|}{2n} = \frac{\sum_i |p_i - \hat{p}_i|}{2},$$

where the sum is over all cells, n_i is observed frequency, $\hat{m}_i$ is the estimated expected frequency, p_i is the proportion of data in cell i, and $\hat{p}_i$ is the estimated probability of being in cell i. The rule of thumb is that a $D \leq .03$ is a good fitting model. We should note that D does not perform well for large tables, because, to achieve perfect fit, observations that would need move to an adjacent cell has the same weight as those that would need to be moved many cells away.

3.2 Example: Who Takes Which MOOCs

The data in Table 1 come from a study examining engagement in massively open online courses (MOOC) with the goal of determining who is being served by taking which course ([10]). The data come from MOOCs covering six different disciplines where all MOOCs except one were offered multiple times. In total, there are 16 course offerings. The topics of the MOOCs were computer science (CS1, CS2), education (Educ1, Educ2), organic chemistry (Chem1, Chem2), business administration on subsistence (Bus1, Bus2, Bus3), environmental science (Env1 – Env6), and animal and veterinary science (Animal). The students' ages were collected on a category scale of six age groups.

The models that are relevant to this data set are (4) and (6) with $J_1 = 16$ and $J_2 = 6$. The models were fit to the data using maximum likelihood estimation as implemented in the `R` package `logmulti`. Goodness-of-fit statistics for six models fit to the data are reported in Table 2. For each model, we report df, G^2, p-value, and the two additional statistics discussed in Section 3.1.

The MOOCs and age groups show a significant relationship ($G^2_{ind} = 1098.3$, $df = 75$, $p \leq .01$). Since G^2 may be significant due to large frequencies or extra heterogeneity between students within each of the combinations of MOOC by age group, we also fit a model of independence using the Negative Binomial distribution. This independence model also showed significant dependency ($G^2_{ind} = 101.46$, $df = 75$, $p = .02$). Furthermore, D is relatively large for both of these two models. The top two plots in Figure 1 are the qqplot's of standardized residuals from the two independence models, and they show considerable departure from normality for smaller and larger frequencies, which gives us further evidence against independence. Examining a table of $(16 \times 6 =)$ 96 residuals does not lead to insight into the relationship between age and MOOCs.

The simplest association model that can be fit to data is the R model, where we can reasonably assign scores to the column variable (i.e., age). The two natural options for known scores would be either equidistant for successive categories or the midpoints of the corresponding age intervals. Neither of these two models provide satisfactory representations of the association in the data. The $RC(2)$ association model fits better than any of the simpler models ($G^2 = 52.17$, $df = 39$, $p = .08$, the percent association=95%, $D = .02$). Furthermore, the bottom right qq-plot plot in Figure 1 shows that the standardized residuals from the $RC(2)$ model are very close of normal. The parameter estimates from the $RC(1)$ and $RC(2)$ association models are plotted in Figure (2) where the category scale values for the MOOCs and age groups are weighted by the square root of the association parameter (i.e., $\hat{\nu}_{ij_i1}\sqrt{\hat{\sigma}^2_1}$ and $\hat{\nu}_{kj_k2}\sqrt{\hat{\sigma}^2_2}$).

Even though $RC(2)$ model is our best model, for the purpose of illustration, the scale value plots for both the $RC(1)$ and $RC(2)$ models are given in Figure 2. For both models, the scale values for the courses contrast STEM and non-STEM courses; that is, at one extreme are the chemistry and computer science courses and at the other extreme the education courses. On the first dimension of both models, the scale values for student's age are ordered from younger to older; however, they are not equally spaced. The age groups 25-29, 30-39 and 40-49 are relatively close in value in the $RC(1)$ model but less so in the $RC(2)$ model. From the $RC(2)$ graph we can say that students aged 18-24 have higher odds of taking STEM courses than the odds for any of the other age groups. Conversely, the ≥ 60 aged students have higher odds of taking the education courses than any of the students in other age groups. Different offerings of the same course tend to have similar scale values, especially Bus1 & Bus2, Educ1 & Educ2. The odds of taking one or the other of these courses (regardless of age groups) is close to 1.

As illustrated in this example, the scale values from the $RC(1)$ and $RC(2)$

Table 1: Who takes which MOOC: a cross-classification of MOOC courses by age groups of students who take the courses.

	Age Groups					
Course	18–24	25–29	30–39	40–49	50–59	≥60
Animal	33	43	64	30	30	17
Bus1	59	101	100	68	49	28
Bus2	45	57	68	38	21	31
Bus3	20	47	62	32	28	35
Chem1	164	149	174	86	69	48
Chem2	71	52	54	27	18	13
CS1	1472	1472	2068	1110	580	254
CS2	198	199	342	199	124	46
Educ1	13	34	114	117	91	46
Educ2	10	20	77	81	65	37
Env1	92	216	313	154	139	117
Env2	126	265	342	197	176	147
Env3	89	155	217	143	149	114
Env4	90	163	216	99	77	63
Env5	111	175	206	134	109	111
Env6	42	78	119	60	62	72

association models need not be the same. Also for a given model, the scale values may be reflected (i.e., multiplied by -1) and this is illustrated in the scale values plots. For $RC(1)$ model on dimensional one, the ages go from low to high, but for the $RC(2)$ model go from high to low. The scale values for courses also are reflected in the $RC(2)$ model compared to the $RC(1)$ model, which leads to the same interpretations for the models.

4 Graphical Models

Log-linear models for categorical data have graphical representations that are visual representations of theory or scientific information, and they can be used to determine whether tables can be collapsed over items without impacting associations ([46, 20]). Graphs also aid us in generalizing the

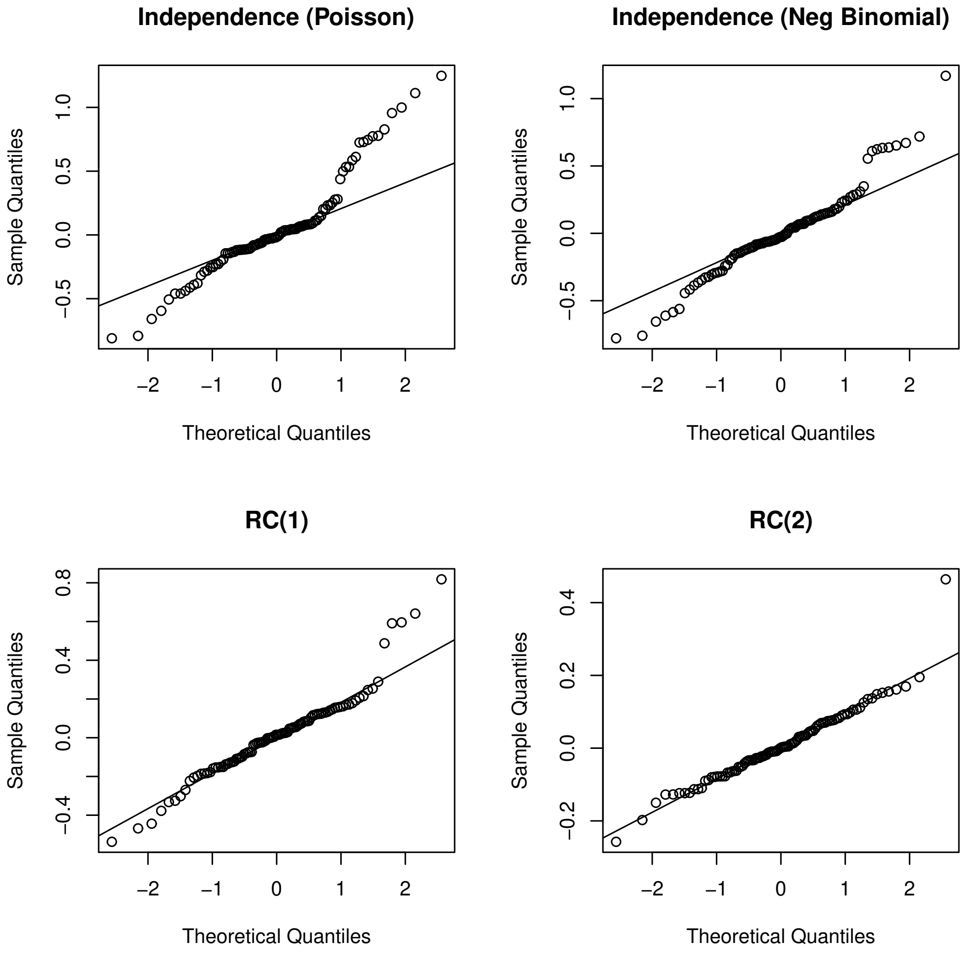


Figure 1: QQplot of standardized residuals from models fit to the MOOC data.

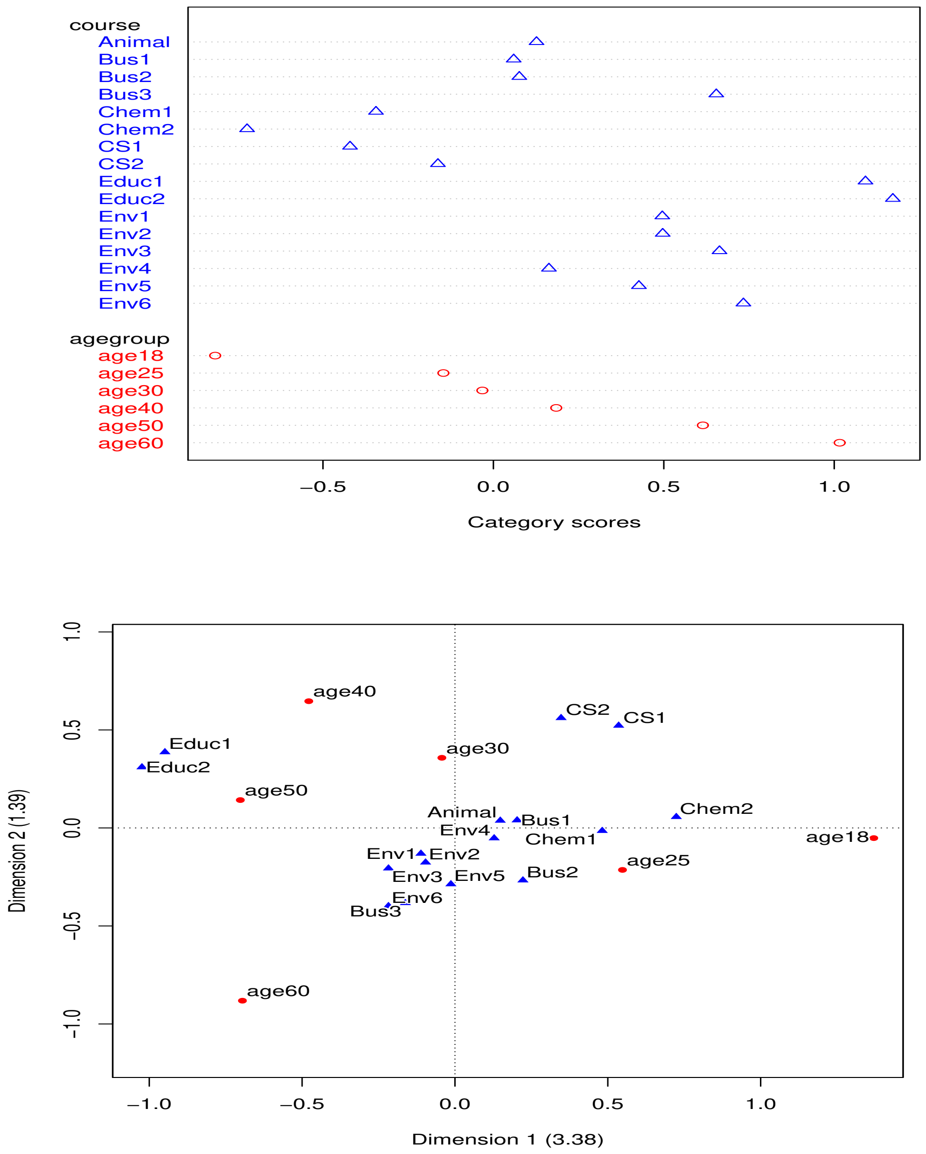


Figure 2: Plot of estimated scales scale values from $RC(1)$ association model (top) and $RC(2)$ association model (bottom) fit to MOOC data.

Table 2: Goodness-of-fit statistics for models fit to the MOOC data in Table 1.

Model	*df*	G^2	p	Percent of association	Dissimilarity index
Independence (Poisson)	74	1098.3	$< .01$	0.00%	.09
Independence (Negative Binomial)	74	101.46	.02	90.76%	.10
R (equidistant scores)	60	291.73	$< .01$	73.44%	.16
R (midpoint scores)	60	313.13	$< .01$	71.49%	.16
$RC(1)$	56	249.40	$< .01$	77.29%	.04
$RC(2)$	39	52.17	.08	95.26%	.02

$RC(M)$ association models to higher dimensions. Graphical models for log-linear models are introduced in this section, followed by graphs for $RC(M)$ association models. Lastly, we add more variables to the graphs to represent situations where we have moderate to very high dimensional tables (i.e., large numbers of items).

4.1 Graphs for Log-linear Models

A graph consists of nodes, which for us are variables or items, and edges or lines connecting nodes indicating possible (non-directional) dependency between variables. For example, consider a three-dimensional $J_1 \times J_2 \times J_3$ contingency table, cross-classifying the categorical variables Y_1, Y_2, Y_3.

Figure 3 contains four simple graphs showing the relationship between Y_1, Y_2 and Y_3. In this chapter, discrete variables are represented by boxes. The absence of a line connecting two variables indicates that the two variables are independent conditional on the rest of the graph. Graph 3(a) does not contain any edges and this graph represents complete independence. The presence of a line between two variables only indicates that they *may* be dependent conditional on the rest of the graph. Figure 3(b) represents a log-linear model of joint independence between Y_2 and Y_1 & Y_3, and graph 3(c) represents a log-linear model of conditional independence between Y_1 and Y_2 given Y_3.

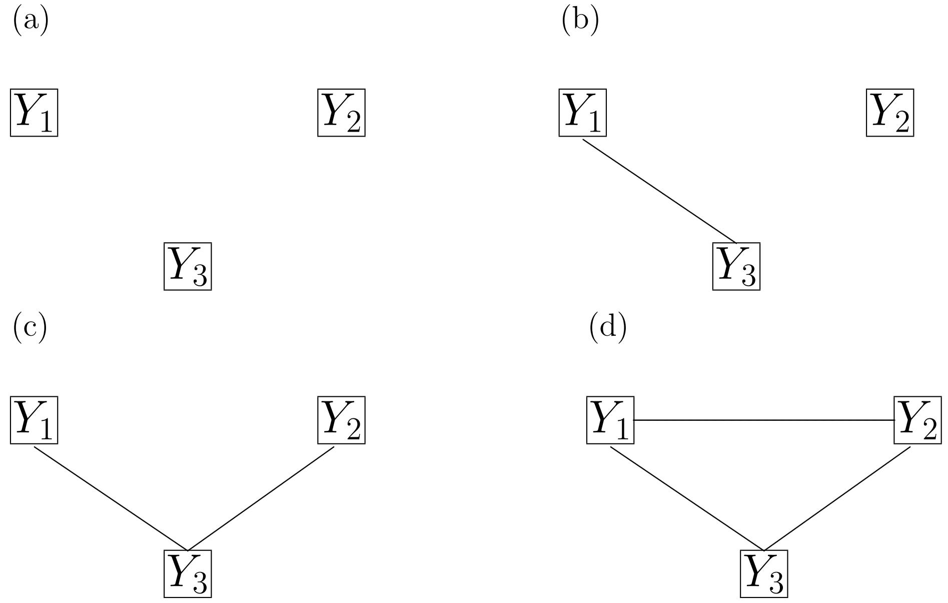


Figure 3: Graphical models corresponding to log-linear models of (a) complete independence, (b) joint independence, (c) conditional independence, and (d) 3-way interaction.

Graphical models (a), (b) and (c) are collapsible over variables. For example, in (b), we can collapse the data over Y_2 and this does not change the dependency between Y_1 and Y_3; that is, we can simply analyze the marginal relationship between Y_1 and Y_3. For model (c), conditional independence of Y_1 and Y_2 given Y_3, we can collapse over Y_2 to study the relationship between Y_1 and Y_3, and collapse over Y_1 to study the relationship between Y_2 and Y_3. Any model that has some form of (conditional) independence can be collapsed over some set of variables (items); however, this is not the case for graph (d).

Figure 3 (d) is a model of conditional dependence; that is, none of variables are independent conditional on the rest of the graph. This graph is a representation of a log-linear model with all 2-way interactions between pairs of variables and a model with all 2-way interactions and a 3-way interaction (i.e., a saturated model). For every model there is a unique graph, but every graph with edges (dependencies) can represent multiple models. This yields an ambiguity regarding the complexity of the interaction structure. In this chapter, we use graphs to represent theory and take the most complex model implied by a graph. For example, Figure 3 (d), which is a complete graph[2], represents the log-linear model with all 2-way interactions

[2] A complete (sub)graph in one where all variables are directly related to each other.

and a 3-way interaction.

We can obtain graphical representations for our models such that there is more of a one-to-one correspondence between graphs and models. Consider the simpler case of 2 categorical variables. In Figure 4, the graphs (a) and (d) represent log-linear models of complete independence and dependence, the latter being a saturated log-linear model. As commented on in Section 3, in this case we have $(J_1 - 1)(J_2 - 1)$ degrees of freedom with which to represent the dependency; however, we may not need all of these degrees of freedom. There are models in between independence and dependence, which we discuss in Section 3. We introduce an unobserved continuous variable in our graphs, which are represented by the circles in graphs 4 (b) and (c). The categorical variables are now conditionally independent given the latent continuous variable(s). Consider graph (b) in Figure 4. If we collapse over the continuous variable, we will produce an association between the categorical variables ([46]). The model for observed data is one of dependence. Graph (b) is a representation of the LL, U, R, C, and $RC(1)$ models. The differences depend on whether the scale values are set equal to specific values or are estimated. For the LL model, both ν_{1j_11} and ν_{2j_21} are set equal to specific values, for the U models both ν_{1j_11} and ν_{2j_21} are set to equally spaced scores, for the R (or C) model one set of scores (e.g., ν_{ij_i1}) is set to specific values and the other set (e.g, ν_{2j_21}) is estimated, and for the $RC(1)$ model, both ν_{1j_11} and ν_{2j_21} are estimated. Graph (c) is a representation of $RC(M)$ association model previously introduced in Section 3 and are discussed below in more detail below.

4.2 Graphs of the $RC(M)$ Association Model

Figure 4 (b) is a graphical representation of models for two variables corresponding to the U, LL, R, C and $RC(1)$ association models, and Figure 4 (c) is the representations for the $RC(M)$ model. To represent the AMs, we have added a continuous variable that is unobserved or latent. These continuous variables are represented by the circles. Goodman ([28]) first mentioned that a latent variable may underlie data fit by an $RC(1)$ model, but he never expanded on this. We provide explicit details about a possible underlying or latent variable model and use this to generalize the $RC(M)$ model to high dimensional tables.

Models can be “read” from the graphs. As an example, consider the $RC(1)$ association models represented by graph (b) in Figure 4. All models

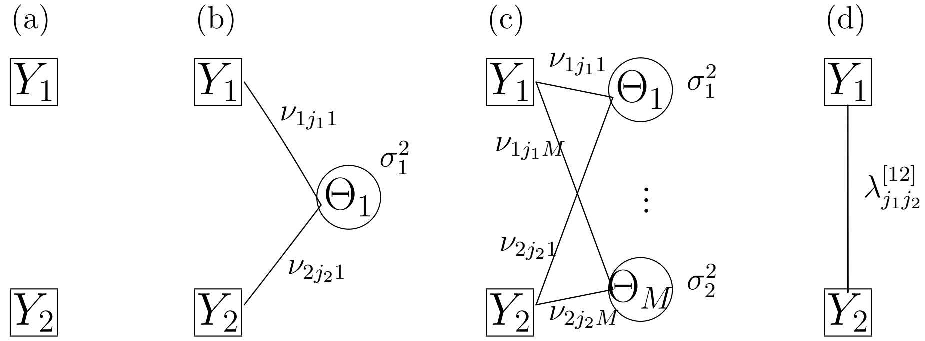


Figure 4: Graphs for log-linear models for 2-way tables where (a) is a log-linear model of independence, (b) is the RC(1) association model, (c) is the RC(M) association model, and (d) is a saturated log-linear model.

for data include a parameter to ensure probabilities sum to 1 (i.e., λ), and include marginal effect terms for each categorical variables (i.e., $\lambda_{j_i}^{[i]}$ and $\lambda_{j_k}^{[k]}$). For the interaction, the lines connecting unobserved continuous and observed discrete variables are labeled by the category scale values, and the latent variable Θ_1 is labeled by σ_1^2. The observed interaction between Y_1 and Y_2 equals the product of parameters on the path between Y_1 and Y_2; that is, $\nu_{1j_11}\sigma_1^2\nu_{2j_21}$. Likewise, the interaction between Y_1 and Y_2 represented by Figure 4 (c) is $\sum_m \sigma_m^2 \nu_{1j_1m}\nu_{2j_2m}$.

The AMs in Figures 4 (b) and (c) are models of conditional independence: the (observed) categorical variables are independent given values on the unobserved continuous variables. The number of Θ_m's corresponds to the dimensionality of the $RC(M)$ model, which should not be confused with the dimension of a cross-classification (i.e., the number of variables). Since the Θ_ms are continuous, if we collapse over the Θ_ms, we may observe a dependency between the categorical variables. On the contrary, we cannot collapse over one categorical variable to study the relationship between the other categorical variable and the continuous variable. According to theory on graphical models, the graphs for the $RC(M)$ models are not collapsible ([46], [20]); however, this is a property that does not hold in a strong sense, as will be shown in Section 5 in the context of higher-dimensional models.

To derive the algebraic model from the graph, we need two assumptions in addition to conditional independence. First, the observed data $\boldsymbol{y}$ come from a multinomial distribution, which as mentioned above is a common assumption for way tables of frequencies. This assumption is not restrictive, because for inferential purposes, the three standard sampling schemes for contingency tables (multinomial, product multinomial (i.e. independent multinomials in each row or column), and independent Poisson in each cell) are equivalent. We must also assume that the latent variables follow a (multivariate) normal distribution where the mean and variance are conditional on the response patterns (i.e., cells of the table); that is,

$$\boldsymbol{\theta} \mid \boldsymbol{y} \sim MVN(\boldsymbol{\mu}_{\boldsymbol{y}}, \boldsymbol{\Sigma}_{\boldsymbol{y}}).$$

Justification for the assumption of a conditional Gaussian distribution for $\boldsymbol{\theta}$ can be found in Chang [12, 13] and [44]. The association parameters of the $RC(M)$ model are the elements of $\boldsymbol{\Sigma}_{\boldsymbol{y}}$. Typically, we assume a homogeneous conditional covariance matrix, i.e., $\boldsymbol{\Sigma}_{\boldsymbol{y}} = \boldsymbol{\Sigma}$. Previously, we discussed the orthogonality identification constraint on the $\boldsymbol{\nu}_{im}$s for $M > 1$, which requires that $\boldsymbol{\Sigma}$ is a diagonal matrix, $\boldsymbol{\Sigma} = \text{diag}(\sigma_1^2, \ldots, \sigma_M^2)$. The conditional mean of θ_m is the sum of the category scale values that are directly related to θ_m weighted by σ_m^2; that is,

$$\boldsymbol{\mu}_{\boldsymbol{y}} = \left(\sigma_1^2 \sum_i \nu_{ij_i1},\ \sigma_2^2 \sum_i \nu_{ij_i2}\ , \ldots,\ \sigma_M^2 \sum_i \nu_{ij_iM}\right)' \tag{8}$$

The $RC(M)$ association models do not include values of the θ_ms; however, the models do include parameters that give us the distributional parameters of $\boldsymbol{\theta}|\boldsymbol{y}$.

5 High Dimensional Tables

High dimensional tables are common, especially when considering questions on surveys, items on psychological scales, or items on educational tests. Two problems faced with analyzing high dimensional tables are the large numbers of (i) 2-way interactions, and (ii) cells. For example, 20 five category items, there are $20(19)/2 = 190$ different 2-way interactions and $5^{20} = 9.536743e + 13$ cells in the cross-classification of the items. To deal with the problem of large numbers of interactions, we generalize of the AMs to large numbers of variables. For the second problem where the table is large and data are sparse, we use pseudo-likelihood estimation. In this section, we tackle both problems and discuss the connection between AMs and IRT models.

To generalize the association models to high dimensional cross-classifications, we start with graphs and subsequently discuss the algebraic model. We continue to only consider two-way interactions, because item response models using the standard assumption that $f(\boldsymbol{\theta})$ is multivariate normal implies only two-way interactions between items. We will discuss the similarities and differences with respect to $RC(M)$ association model, as well as explicitly show the correspondence of association model parameters and common IRT models.

5.1 Graphs for High Dimensional Association Models

For high-dimensional tables, we simply add variables to the graphs, as in Figure 5. Figure 5 has three examples of possible graphs for 6 items. Graph 5 (a) is similar to an $RC(1)$ model, except instead of 2 categorical variables we have 6. In all graphs in Figure 5, the categorical variables are conditionally independent given the unobserved continuous variable(s); however, the latent variables can be dependent. The covariance between latent variables θ_m and $\theta_{m'}$ conditional on the observed variables is equal to $\sigma_{mm'}$. We have changed our notation slightly and are using σ_{mm} rather than σ_m^2 for variances (i.e., association parameters).

If categorical variables are discrete measures of underlying continuous variables, then it would stand to reason that the scale values for the variables are the same over the interactions; that is, the scale values would be homogeneous. For example, in graph 5 (a) the interaction between, say variables Y_i and Y_k, would be represented by $\sigma_{11}\nu_{ij_i1}\nu_{kj_k1}$ and the interaction between Y_i and Y_ℓ would be $\sigma_{11}\nu_{ij_i1}\nu_{\ell j_\ell 1}$, both of which involve ν_{ij_i1} and σ_{11}.

Just as we replaced two-way interaction parameters in a log-linear model for 2 items by products of association parameters and scale values to get an $RC(M)$ model, we do the same for association models for high dimensional tables. The interactions between the categorical variables are the products of labels of the paths between them. For example, the interactions between variables Y_1 and Y_6 for graphs in Figure 5 are

$$\begin{array}{ll} \sigma_{11}\nu_{1j_11}\nu_{6j_61} & \text{for graph (a)} \\ \sigma_{12}\nu_{1j_11}\nu_{6j_62} & \text{for graph (b)} \\ \sigma_{13}\nu_{1j_11}\nu_{6j_63} & \text{for graph (c).} \end{array}$$

The latter two involve covariances between the latent variables.

5.2 Algebraic Details and Properties

The most general case where each item is directly related to each of the latent variables and all latent variables are related to each other leads to the following complex association model:

$$P(\boldsymbol{y}) = \exp\left[\lambda + \sum_i \lambda_{j_i}^{[i]} + \sum_i \sum_{k>i} \sum_m \sum_{m' \geq m} \sigma_{mm'}\nu_{ij_im}\nu_{kj_km'}\right]. \tag{9}$$

This model has an intercept, all main effects, and all possible two-factor interactions, where the interactions have a multiplicative structure. For the models to be equivalent to a hierarchical log-linear model of all two-factor interactions would require the number of terms (dimension) for the $RC(M_{ik})$ interaction term of every pair of items Y_i and Y_k, for $i, k = 1, \ldots, I$, to equal $M_{ik} = \min(I_i, I_k) - 1$.

A variety of more parsimonious models with a special structure for the associations among the variables of sound interpretation can be obtained by considering smaller values for the rank of the interaction terms or/and homogeneity of scores across interaction terms. Furthermore higher-order interactions having multiplicative terms among scores for more than two variables are possible. For the case of three-factor interactions and related references we refer to [40, Sections 6.7, 6.8.1].

These complex models require a considerable number of identification constraints; therefore, for the sake of discussion, we restrict our attention to models where each item is related to only one latent variable, which means that all interaction terms are of RC(1) type. The simple structures shown in Figure 5 imply that each item has only one set of ν_{ij_im}s that are not all equal to zero. If Y_i is not related to a Θ_m, then $\boldsymbol{\nu}_{im} = \mathbf{0}$. For example, in Figures 5 (b) and (c), there is no edge between Y_1 and Θ_2 so $\nu_{1j_12} = 0$, $j_1 = 1, \ldots, J_1$.

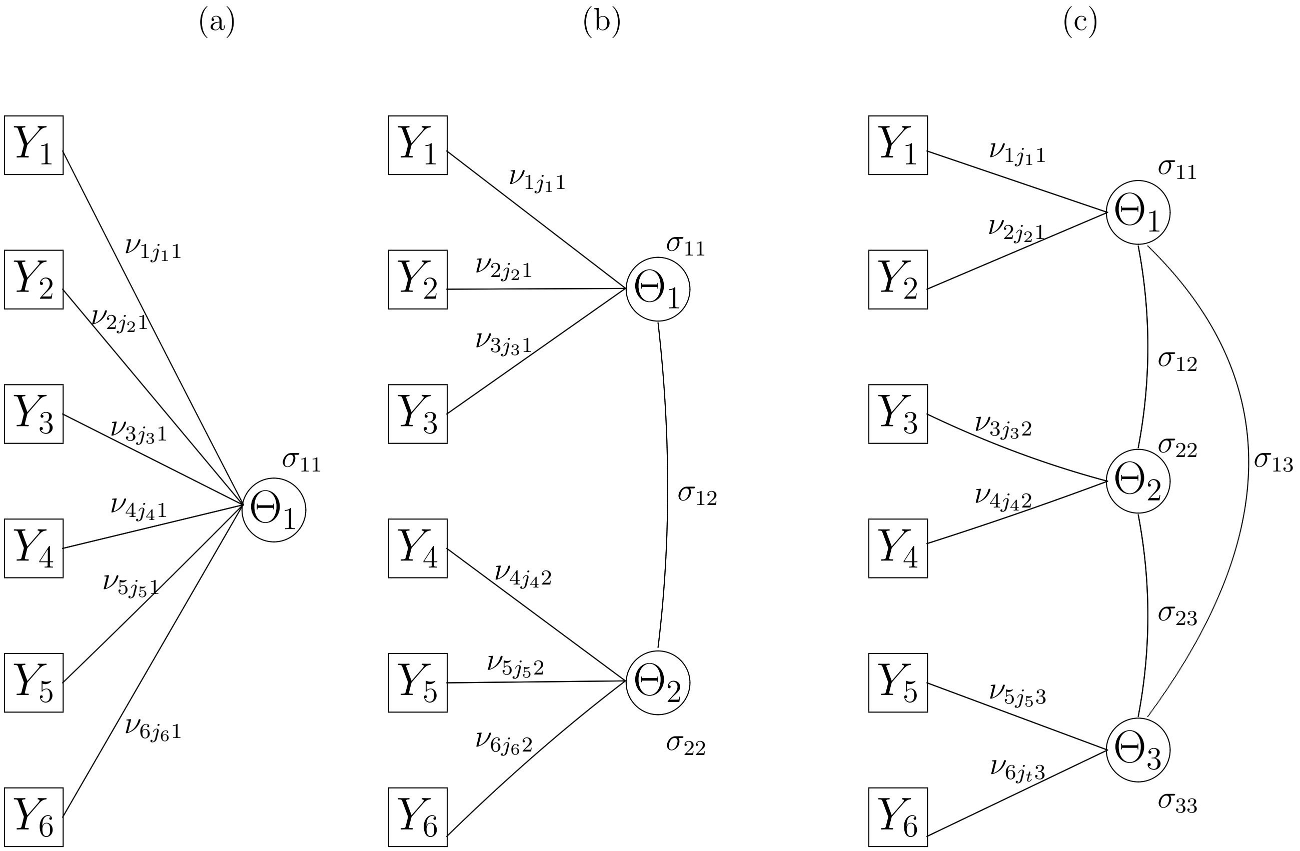


Figure 5: Graphs for log-multiplicative association models for 1, 2 and 3 continuous latent variables (circles) and six observed categorical variables (squares).

In AM models, we can have additional edges between say Y_1 and other Θs; however, the simple structures discussed here prove to be sufficient for many applications.

Association model (9) can be derived from the same assumptions as the $RC(M)$ models: $\boldsymbol{y}$ is multinomial, $\boldsymbol{\theta}$ is conditional Gaussian[3], and conditional independence of items given $\boldsymbol{\theta}$. One difference is that the conditional covariance matrix does not have to be diagonal ([3]). As a result, the means within response patterns also include responses to other variables; namely,

$$E(\theta_m|\boldsymbol{y}) = \sigma_{mm}\left(\sum_i \nu_{ij_im}\right) + \sum_{m'>m} \sigma_{mm'}\left(\sum_k \nu_{kj_km'}\right). \qquad (10)$$

For simple structures, items Y_i load on latent variable θ_m and Y_k load on $\theta_{m'}$. Estimates of θ_m are based not only on the items directly related to θ_m, as in (8), but also those that are indirectly related through $\theta_{m'}$. When $\sigma_{mm'} \neq 0$, measurement can be improved and become more precise by including multiple correlated latent variables ([64, 16]). In addition to the derivation based on statistical graphical models, model (9) can be derived from an IRT perspective ([36, 35, 14, 4, 2]), conditional specification of models [4, 2]), a theory of ferrimagnetism ([45, 49]), distance based models ([17, 18, 19]), others.

To facilitate the discussion of the models, we use the following 2 dimensional model for 4 categorical variables where variables Y_1 and Y_2 are directly related to θ_1 and variables Y_3 and Y_4 are directly related to θ_2:

$$\begin{aligned} P(\boldsymbol{y}) &= \exp[\lambda + \sum_{i=1}^{4} \lambda_{j_i}^{[i]} + \sigma_{11}(\nu_{1j_11}\nu_{2j_21}) + \sigma_{22}(\nu_{3j_32}\nu_{4j_42}) \\ &\quad +\sigma_{12}(\nu_{1j_11}\nu_{3j_32} + \nu_{1j_11}\nu_{4j_42} + \nu_{2j_21}\nu_{3j_32} + \nu_{2j_21}\nu_{4j_42})]. \end{aligned} \qquad (11)$$

A log-linear model with all 2-way interactions for 4 five category items could require $(6 \times 5 \times 5) = 150$ parameters to completely represent the dependencies in data; whereas, the association model with homogeneous scales across interactions would have at most 23 parameters[4]. The difference between the number of parameters of log-linear and AMs increases exponentially for more items and categories per item.

[3] The marginal distribution of $\boldsymbol{\Theta}$ is a mixture of Gaussian distributions.

[4] The number of unique parameters for the log-linear model equals $(6 \times 4 \times 4) = 96$ and that for the association model equals $(4 \times 4) + 1 = 17$.

Unlike AMs for two-way tables which require more than two categories per variable, this is not the case for the higher-dimensional models. For example, model (11) and models that correspond to graphs in Figure 5 (a), (b) and (c) can be fit to binary variables ([3, 4]).

The identification constraints on the location of marginal effect terms and the scale values are analogously to the $RC(1)$ model (e.g., $\sum_{j_i} \lambda^{[i]}_{j_i} = 0$ and $\sum_{j_i} \nu_{ij_im} = 0$) and just one scaling constraint is required for each latent variable. For example, in (11), possible scaling constraints can be either

$$\sigma_{mm} = 1 \quad \text{for all } m,$$

or

$$\sum_{j_i} (\nu_{ij_im})^2 = 1 \qquad \text{for one } i \text{ per } \theta_m \text{ for all } m \ ,$$

but not both, analogous to the $RC(1)$ model. In example (11), if $\sigma_{11} = \sigma_{22} = 1$, then we cannot linearly transform the ν_{ij_im}s without changing the values of the interaction terms. If we fix the variances and re-scale ν_{1j_11} such that $\sum_{j_i} \nu^2_{ij_im} = 1$, the interaction between variables does not necessarily remain the same. Placing scaling constraints on both of the σ_{mm} and ν_{ij_im} is a restriction that impacts the goodness-of-fit of the model. For example, if we set variance to $\sigma_{11} = 1$ and re-scale the ν_{1j_11}, then the interaction between Y_1 and Y_2 changes,

$$\sigma_{11}\nu_{1j_11}\nu_{2j_21} \neq 1(\nu_{1j_11}/c)\nu_{2j_21} = (1/c)\nu_{1j_11}\nu_{2j_21},$$

where $c = \sqrt{\sum_{j_1}(\nu_{1j_11})^2}$. To achieve equality, either $\sum_{j_1} \nu^2_{ij_1m} \neq 1$ or $\sigma_{11} = 1/c$. Whether the scaling constraint is put on σ_{mm} or scale values is more a matter of convenience. For example, for estimation of models for our example, we found it more convenient to set $\sigma_{mm} = 1$; however, after the model has been fit, we can switch to $\sum_{j_i} \nu^2_{ij_im} = 1$ and adjust σ_{mm} (and the $\sigma_{mm'}$) and other scale values. We did the latter in a simulation study reported below on culpability where we needed to separate the effects of the strength and structure of the association.

In the standard AMs framework, (11) is an AM having RC(1) type two-factor interactions and every variable has homogeneous scores (i.e., the same scores across all interaction terms involved)

$$\begin{aligned} P(\boldsymbol{y}) &= \exp[\lambda + \sum_{i=1}^{4} \lambda_{j_i}^{[i]} + \phi_{12}\nu_{1j_1 1}\nu_{2j_2 1} + \phi_{34}\nu_{3j_3 2}\nu_{4j_4 2} \\ &\quad +\phi_{13}\nu_{1j_1 1}\nu_{3j_3 2} + \phi_{14}\nu_{1j_1 1}\nu_{4j_4 2} + \phi_{23}\nu_{2j_2 1}\nu_{3j_3 2} + \phi_{24}\nu_{2j_2 1}\nu_{4j_4 2}], \end{aligned} \tag{12}$$

with the additional constraint on certain intrinsic association parameters $\phi_{13} = \phi_{14} = \phi_{23} = \phi_{24} = \sigma_{12}$ (notice that $\phi_{12} = \sigma_{11}$ and $\phi_{34} = \sigma_{22}$). Such constraints are unusual for standard AMs, but are found in applications to square tables (rows and columns are the same categories) and are linked to latent variables models (e.g., IRT models) later in Section 5.4. In applications with all variables (items) being measured on the same scale, we find homogeneity constraints on the scores for each variable and dimension (i.e., $\nu_{ij_i m} = \nu_{kj_k m}$ where $j_i = j_k$) that result in symmetric interaction terms.

To understand the physical interpretation of this constraint, consider (12) under the additional assumption that the scores of all variables are known, equidistant for successive categories (i.e. $\nu_{i(j_i+1)m} - \nu_{ij_i m} = c_i$, for all $j_i = 1, \ldots, J_i - 1$, with $m = 1$ for $i = 1, 2$ and $m = 2$ for $i = 3, 4$), which means that we assume U-type structures for all interactions. In particular for the (Y_1, Y_2) partial table when $Y_3 = j_3$ and $Y_4 = j_4$ we have

$$\begin{aligned} \theta_{j_1 j_2 | i_3 i_4}^{[12]} &= \exp\left(\frac{\pi_{i_1,i_2,i_3,i_4}\pi_{i_1+1,i_2+1,i_3 i_4}}{\pi_{i_1,i_2+1,i_3,i_4}\pi_{i_1+1,i_2,i_3 i_4}}\right) \\ &= \exp\left(\phi_{12}(\nu_{1(j_1+1)1} - \nu_{1j_1 1})(\nu_{2(j_2+1)1} - \nu_{2j_2 1})\right) = \exp\left(\phi_{12}c_1c_2\right) = \theta^{[12]}, \end{aligned} \tag{13}$$

while for the other partial tables, the $\theta^{[ik]}$'s, $i, k = 1, \ldots, 4$ with $i \neq k$, are defined analogously. Consequently, the conditional local ORs in every partial table (Y_i, Y_k) are all equal to $\theta^{[ik]}$, for all values of j_i and j_k (uniform) but also across all levels of the other items (homogeneous). Thus the underlying model is the homogeneous U model (see [40, Section 6.7]). Notice that due to the sum to zero constraints satisfied by the scores, $c_i \neq c_k$ if $J_i \neq J_k$. For the special case of $J_i = J$, $i = 1, \ldots, 4$, it holds $c_i = c$ and the additional equality constraint among the ϕ parameters above leads to $\theta^{[13]} = \theta^{[14]} = \theta^{[23]} = \theta^{[24]}$, hence to equality of the corresponding conditional local ORs.

A difference in terms of identification constraints with respect to AMs for two-way tables with M latent variables (i.e. $RC(M)$ models), is that in models of type (11) with more than two variables, each variable is directly related to only one of the M latent variables; whereas, under $RC(M)$ each variable is related to all M of them. The orthogonality constraints that are required for the $RC(M)$, are not required for (11) and $\mathbf{\Sigma}$ can have non-zero off diagonals. This is not true for all versions of (9), in particular if every variable is directly related to each and every Θ_M, then an orthogonality is required as well as for underlying bi-factor structures.

An alternative version of (9) that has the same identification constraints as (11) may have all variables directly related to each of the M latent variables, except one per latent variable. The variables that are related to just one M "anchor" the rotation. Similarly, in a factor analysis/IRT model framework, parameter constraints are imposed to uniquely identify the model parameters. In a factor model with M latent variables, M^2 constraints are required to obtain a unique solution and avoid the rotational indeterminancy issue. Among the constraints are those that set the scale of the latent variable. Similarly to what it has been said above for $RC(M)$ models, the scale of a latent variable is set either by standardizing the latent variable assuming that is has zero mean and unit variance in the population or by forcing its scale to be the same as one of the observed variables. Usually, the variable that best represents the latent variable has its factor loading set equal to one. The selected variable is known as a "reference" variable. Setting the scale of the latent variables to one takes care of M of the required restrictions. The additional ones are imposed on the loading and factor covariance matrices (e.g. diagonal factor covariance matrix, certain loadings are set to zero). In exploratory factor analysis, the required restrictions can be imposed on any of the parameters. Those restrictions will produce an arbitrary set of factors which can be then rotated to another set of factors that have better interpretability. In confirmatory factor analysis, the constraints are driven by the investigator's research hypothesis. A useful constraint that eases the interpretation of the factors is to consider that each latent variable has at least one item that loads solely on that factor (i.e. setting specific elements of the loading matrix to zero) ([38, 39]). Returning to the AMs framework, anchoring one item (i.e., $\sum_j \nu_{ijm}^2 = 1$ and $\nu_{ijm'} = 0$ for $m' \neq m$) that best represents the latent variable is a key to fitting non-simple structure models.

With the $RC(M)$ association model, we cannot collapse over an item and study the relationship between the other item and a latent variable, because then we would have only one observed variable. This is not true in a strong sense for more than 2 items. As mentioned previously, ν_{ij_im} represents the structure and σ_{mm} represents the strength of the relationship between variables. This leads to a semi-collapsible situation. This is illustrated for the

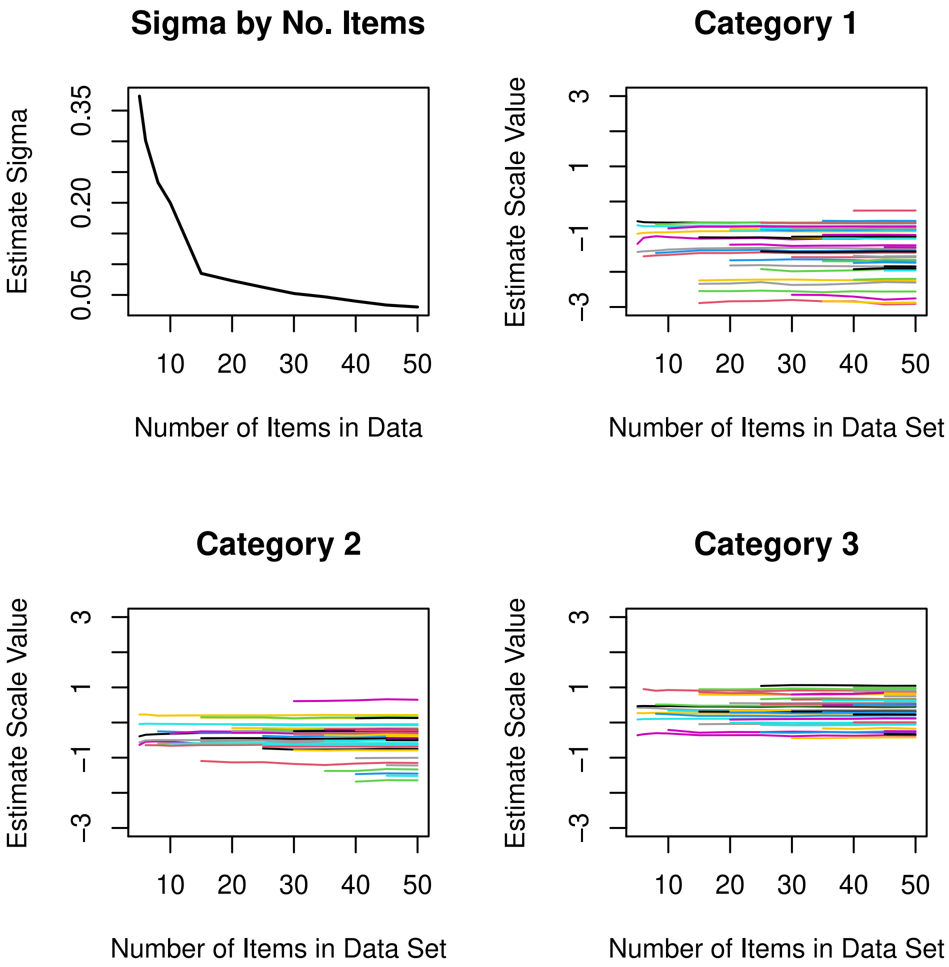


Figure 6: Mean estimated σ_{11} and ν_{ij_i1} parameters from fitting a unidimensional nominal model to data simulated from a nominal item response model for $n = 500$ and four category items, with the number of items varying up to 50.

uni-dimensional model, such as Figure 5 (a). If we have 50 items but drop or collapse over one of them, the structure of the relationship between an item and the latent variables should not change but the strength of associations does change ([5]). To illustrate this property, 100 data sets were simulated for 50 four category items. A nominal IRT model was used to simulate the data where the slopes (scale values) were drawn from a $N(0, 1.5)$, and the location (marginal effects) were drawn from a $N(0, 2)$. Values of θ for each of $n = 1000$ observation for each of the 100 replications were drawn from a $N(0, 1)$. For cases where a simulated item did not have observations in all categories, a new data set was randomly created. The model was fit to data sets with 50 to 15 items dropping 5 items at a time, and 14 to 5 items dropping one item at a time. The scaling identification constraint was placed on the first item (i.e., $\sum_{j=1}(\nu_{1j1})^2 = 1$) and σ_{11} was estimated.

In Figure 6, the means over replications of association parameters and scale values are plotted by the number of items in the data set. The estimated σ_{11}s are larger for small numbers of items and asymptotes down to 0 for large numbers of items. Recall that σ_{11} is the conditional variance of θ_1 within a response pattern $\boldsymbol{y}$ (i.e., a cell in a cross-classification of items). When the number of possible patterns is very large (infinity), only one person can fall into a cell of the table, so the variance necessarily equals 0. Also in Figure 6 are the means of the estimated scale values where each line in the figures corresponds to a different item. The scale values plots are only given for three of the four categories because $\nu_{i4_i1} = -\sum_{j_i=1}^{3} \nu_{ij_i1}$. The lines are essentially flat. In other words, we can collapse over categorical variables and the structure between variables remains the same. The only thing that changes is the strength of the relationship between observed variables (items), and between items and latent variables. This also holds for mulidimensional models ([5]). The association between two observed variables, say Y_1 and Y_2, is represented in the AM by $\sigma_{mm'}\nu_{1j_1m}\nu_{2j_2m'}$ (for $m = m$ and $m \neq m'$). As variables are dropped σ_{mm} and $\sigma_{mm'}$ both get larger, but the $\nu_{ij_im'}$ stay essentially the same. Therefore, associations are larger for fewer items but the structure remains the same.

5.3 Pseudo-likelihood Estimation

Large numbers of variables result in large, sparse cross-classifications, in which case maximum likelihood estimation becomes computationally infeasible. An alternative is pseudo-likelihood estimation (PLE), which takes a large complex problem and reduces it to a number of simpler and smaller problems ([8, 9, 6, 27]). A joint distribution can be specified by a set of conditional distributions ([26]) where the conditional distributions are compatible and consistent with the joint distribution and imply a unique model for the joint distribution. Rather than maximize the likelihood of the joint distribution, PLE for AMs maximize the product of the (log) likelihoods of one variable conditional on the rest. Pseudo-likelihood estimators are asymptotically consistent and normal [6, 27, 37]). The conditional distributions of (9) for one item given responses to all other items is a discrete choice model or conditional multinomial logistic regression model. The conditional distribution for item i for individual s is

$$\begin{aligned} P(Y_{is} = j|\boldsymbol{y}_{-i,s}) &= \frac{\exp[\lambda^{[i]}_{j_i} + \nu_{ijm}\sum_{k\neq i}\sum_{m'}\sigma_{mm'}\nu_{kj_km'}]}{\sum_{j=1}^{J}\exp[\lambda^{[i]}_{j_i} + \nu_{ij_im}\sum_{k\neq i}\sum_{m'}\sigma_{mm'}\nu_{kj_km'}]} \qquad (14)\\ &= \frac{\exp[\lambda^{[i]}_{j_i} + \nu_{ijm}\tilde{\theta}_{-i,ms})]}{\sum_{j=1}^{J}\exp[\lambda^{[i]}_{j_i} + \nu_{ij_im}\tilde{\theta}_{-i,ms}]} \qquad (15)\\ &= \frac{\exp[\lambda^{[i]}_{j_i} + \sum_{m'}\sigma_{mm'}\breve{\theta}_{im's}]}{\sum_{j=1}^{J}\exp[\lambda^{[i]}_{j_i} + \sum_{m'}\sigma_{mm'}\breve{\theta}_{im's'}]} \qquad (16) \end{aligned}$$

where $\boldsymbol{y}_{-i,s}$ are the responses by person s to all items except item i, and j_k are the categories chosen by individual s. The predictor variable in (15), $\tilde{\theta}_{-i,ms}$, is the weighted sum of person s's scores on all items except item i; that is,

$$\tilde{\theta}_{-i,ms} = \sum_{k\neq i}\sum_{m'}\sigma_{mm'}\nu_{kj_km'}.$$

Note that $\tilde{\theta}_{-i,ms}$ depends on individual s. Using this value for $\tilde{\theta}_{-i,ms}$, we can get estimates of λ_{ij_i} and ν_{ij_im} by fitting (15) to the data for item i.

There are multiple $\sigma_{mm'}$s that need to be estimated, one $\breve{\theta}_{im's}$ for each $\sigma_{mm'}$. The $\breve{\theta}_{im's}$ equal a different weighted sum of persons s's scale values for $k \neq i$, specifically,

$$\breve{\theta}_{im's} = \nu_{ijm}\sum_{k\neq i}\nu_{kj_km'} \ ,$$

where the sub-script j_k indicates the categories chosen by s on item k. The $\breve{\theta}_{im's}$ differ over individuals, items, and categories. The slopes in (16) are the same over individuals and items. If we had estimates of $\breve{\theta}_{im's}$, estimates of $\lambda_{j_i}^{[i]}$ and $\sigma_{mm'}$ can be obtained by fitting (16) to a data for just item i; however, the $\sigma_{mm'}$ must be the same over items. We need to also estimate all possible $\sigma_{mm'}$s. Fitting only (16) to one item's data does not yield all possible $\sigma_{mm'}$'s. For example if $m = 1$, then only $\sigma_{1m'}$s would be estimated but not say σ_{23}. To impose the restriction on $\sigma_{mm'}$s over items and estimate all of them, we vertically concatenate or "stack" the data and fit a single discrete choice model to the stacked data. In the stacked data set, there are blocks of J_i lines for each item and each individual.

Of importance is the recognition that if we have the conditionals in (14) for every item, the set of models are compatible and consistent with a joint distribution for all the items. The set actually over determines the joint distribution and thus requires restrictions on the parameters. The restrictions are that the terms that represent the interaction of i and k are the same whether of i is modeled as a function of k or k as a function of i. These terms equal $\nu_{ij_im}\sigma_{mm'}\nu_{kj_km'}$, and since $\sigma_{mm'} = \sigma_{m'm}$, the restriction is met. The set of fully conditional distributions given by (14) uniquely imply the AM in (9) for the joint distribution of $\boldsymbol{Y}$ ([4, 2, 56] and references therein).

For uni-dimensional models, ν_{ij_im} and $\lambda_{j_i}^{[i]}$ are estimated by fitting model (15) to the data for item i using the current estimates of the scale values for all $k \neq i$ and the $\sigma_{mm'}$ parameters to compute the predictor variable $\tilde{\theta}_{-i,ms}$. This is done successively for each item and fitting of the model to item data is iterated until convergence is achieved. For multidimensional models, estimates of $\sigma_{mm'}$s are obtained by fitting (16) to the stacked data set using current estimates of scale values to compute the $\breve{\theta}_{im's}$ values. For uni-dimensional models, only item parameters are estimated; whereas, for multi-dimensional models, the algorithm iterates between up-dating ν_{ij_im} parameters and $\sigma_{mm'}$ parameters. If fixed scores are input (e.g., $\nu_{ij_im} = 0, 1, \ldots, (J_i - 1)$), then model (16) is only fit once.

We maximize the pseudo-likelihood function by fitting discrete choice models to data using MLE. In **R**, discrete choice models can be fit using `mlogit` ([15]), `mnlogit` ([34])), `mclogit` ([22]), and others. Due to the data manipulation required and iterative nature of the PLE algorithm, PLE for log-multiplicative association models has been implemented in the **R** package `pleLMA` (Anderson, 2020). The package `mnlogit` is used in `pleLMA`, because it is efficient and can handle large data sets. Alternative packages, `IssingSampling` ([23]) and `plRasch` ([1]), both implement pseudo-likelihood estimation but they are more limited especially in terms of models for multicategory data and the estimation of category scale values. A version of the `pleLMA` package is included in the supplemental material along with data and code.

PLE for estimation of AMs has been extensively studied for small problems and yields estimates for both ν_{ijm} and λ_{ij} that are nearly identical to MLE values. Paek ([54, 55]) simulated data from (M)IRT models for different numbers of categories (3, 4, 5), different numbers of items (4, 6, 20, 50), 1 to 4 dimensional models, and different sample sizes (200, 500, 1000). For small numbers of items and uni-dimensional models, she found correlations between parameter estimates from MLE and PLE equal to .999 to 1.000, and for multi-dimensional models most correlations were greater than .980. For larger problems where MLE was not possible, data were simulated from an (M)IRT model and results were compared. Paek ([54, 55]) found that PLE estimates recovered the parameters used to simulate the data, were unbiased and had small root mean squared errors. This was true for different numbers of categories, different numbers of items, 1 to 4 dimensional models, and different sample sizes.

Alternative tools for model assessment are required, because the data for high-dimensional tables is sparse. Additionally, we do not obtain fitted values for response patterns (i.e., cells in the table), because estimating the λ-parameters is computationally and numerically challenging even given estimates of all other parameters. Some alternative methods are described Section 6 and others are illustrated in the context of our example; however, conclude this section by briefly describing the connections between the AMs and (M)IRT models.

5.4 Connection to IRT Models

The conditional model in (15) has the same form as the nominal response model, including all of it's special cases (e.g., models in the Rasch family, the two-parameter logistic model, the generalized partial credit model (GPCM)). The mathematical equivalence between AMs and (M)IRT models can be proven formally ([56, 5, 45, 49]). From (15), the $\tilde{\theta}_{-i,ms}$ is person s's value on the latent variable based on all items except i; however, after fitting a model we would use (10) to estimate the mean given a response pattern. The marginal effect parameters are sometimes referred to as "difficulty" or location parameters, and rather than denoted by $\lambda_{j_i}^{[i]}$, they are usually represented by b_{ij}. The scale values ν_{ijm} are slopes on the latent variables and are "discrimination" parameters, often denoted as a_{im}. The following restrictions on the category scale values lead to common IRT models:

$$\begin{array}{llll} \text{Nominal:} & \nu_{ijm} = a_{ijm} & \text{no restrictions} & \\ \text{GPCM:} & \nu_{ijm} = a_{im} x_j & \text{where } x_j = \text{fixed scores} & (17) \\ \text{Rasch:} & \nu_{ijm} = x_j & \text{where } x_j = \text{fixed scores.} & \end{array}$$

The fixed scores, x_j, are typically set to equally spaced values or consecutive integers. Note that the conditional (partial) odds ratios are functions of the association parameters and category scale values. The conditional odds ratio for items i and k for the nominal model equals

$$\exp[\sigma_{mm'}(\nu_{ijm} - \nu_{ij'm})(\nu_{k\ell m'} - \nu_{kj\ell' m'})] \ .$$

When the x_js equal consecutive integers as is typical for GPCM and Rasch models, the local partial OR (13) for models in the Rasch family reduce to $\exp(\sigma_{mm'})$, since $c_i = 1$. Instead of just one value for local ORs in the 2-way table case, there is one for each latent variable and one for each pair of latent variables for a total of $M(M-1)/2 + M$ local conditional ORs. Regardless of the number of variables, the number of these ORs depends on the number of latent variables. For the GPCM with consecutive integers, the local conditional ORs equal $\exp(\sigma_{mm'} a_{im} a_{km'})$; that is, (13) for items i and k with $c_i = a_{im}$ and $c_k = a_{km'}$.

In the AM framework, there is flexibility in setting the x_js, which can be set to of non-equally spaced values and different values over items. These possibilities yield item response models that deviate from the traditional Rasch and GPCM models. If the ordering of the response options is not clear, the category scale values from the nominal model can reveal the proper ordering and whether the spacing between category scale values is approximately equal. The scale values from nominal models can show whether a GPCM is plausible. Alternatively, models can be constructed where some items follow a GPCM and others a nominal model. There is great flexibility in crafting a model for data.

6 Sampling Properties

Let denote with $\boldsymbol{\omega}$ the parameter vector corresponding to the fitted model. For example for model (11),

$$\boldsymbol{\omega}' = (\lambda, \boldsymbol{\lambda}^{[1]}, \boldsymbol{\lambda}^{[2]}\boldsymbol{\lambda}^{[3]}, \boldsymbol{\lambda}^{[4]}, \boldsymbol{\nu}_{11}, \boldsymbol{\nu}_{21}, \boldsymbol{\nu}_{3}, \boldsymbol{\nu}_{42}, \sigma_{11}, \sigma_{22}, \sigma_{12}) \ ,$$

where $\boldsymbol{\lambda}^{[i]}$ is a vector with elements $\lambda_{j_i}^{[i]}$, and $\boldsymbol{\nu}_{im}$ is a vector with elements ν_{ij_im}. From the theory of composite likelihood estimators (pseudo-likelihood), it holds that $\sqrt{N}\,(\hat{\boldsymbol{\omega}}_{PL} - \boldsymbol{\omega}) \xrightarrow{d} \mathcal{N}\left(0, G^{-1}(\boldsymbol{\omega})\right)$, where $G(\boldsymbol{\omega})$ is the Godambe information matrix [48, 62], (also known as the sandwich information matrix) given by

$$G(\boldsymbol{\omega}) = H(\boldsymbol{\omega})J^{-1}(\boldsymbol{\omega})H(\boldsymbol{\omega}),$$

where

$$H(\boldsymbol{\omega}) = E\left\{-\frac{\partial^2}{\partial\boldsymbol{\omega}'\partial\boldsymbol{\omega}} pl(\boldsymbol{\omega};\boldsymbol{y})\right\},$$

$$J(\boldsymbol{\omega}) = Var\left\{\frac{\partial}{\partial\boldsymbol{\omega}'} pl(\boldsymbol{\omega};\boldsymbol{y})\right\},$$

and $pl(\boldsymbol{\omega};\boldsymbol{y})$ is the log pseudo-likelihood function. $H(\boldsymbol{\omega})$ and $J(\boldsymbol{\omega})$ can be estimated by:

$$\hat{H}(\hat{\boldsymbol{\omega}}_{PL}) = -\frac{1}{N}\,\frac{\partial^2}{\partial\boldsymbol{\omega}'\partial\boldsymbol{\omega}} pl\left(\boldsymbol{\omega};(\boldsymbol{y}_1,\ldots,\boldsymbol{y}_N)\right)\Big|_{\boldsymbol{\omega}=\hat{\boldsymbol{\omega}}_{PL}} \tag{18}$$

and

$$\hat{J}(\hat{\boldsymbol{\omega}}_{PL}) = \frac{1}{N}\sum_{n=1}^{N}\left(\frac{\partial}{\partial\boldsymbol{\omega}'} pl\left(\boldsymbol{\omega};\boldsymbol{y}_n\right)\Big|_{\boldsymbol{\omega}=\hat{\boldsymbol{\omega}}_{PL}}\right)\left(\frac{\partial}{\partial\boldsymbol{\omega}'} pl\left(\boldsymbol{\omega};\boldsymbol{y}_n\right)\Big|_{\boldsymbol{\omega}=\hat{\boldsymbol{\omega}}_{PL}}\right)' \ , \tag{19}$$

respectively.

7 Evaluation and Testing

The pseudo likelihood (PL) estimation framework used here falls within the composite likelihood (CL) framework which is used for approximating complex full likelihoods. The inference part under CL requires certain modifications and corrections similar to the ones needed for misspecified models [53]. Overall goodness of fit test statistics (e.g. likelihood ratio, Wald and score test) and model selections criteria (e.g. AIC and BIC) can be derived under the CL estimation framework. Adjusted Wald, score and likelihood ratio test statistic for overall fit and nested models under the CL framework have been developed for models for multivariate clustered data, time series data and structural equation models [47, 27, 53, 62, 41]. Moreover, the model selection criteria AIC and the BIC are appropriately adjusted to hold under CL

7.1 Composite likelihood likelihood ratio test for overall fit

The fit of the model can be assessed by constructing a likelihood ratio test for testing $H_0 : \pi_r = \pi_r(\omega)$ against $H_1 : \pi_r$ subject to $\sum \pi_r = 1$, where ω is a vector of all independent parameters, r runs over all possible response patterns (cells of the contingency table), and π_r is the probability of response pattern r. In particular, $\pi_r(\omega)$ is defined by a model such as (6) or (11). The maximum of log-likelihood ($\ln L$) under H_0 and multinomial sampling is

$$\ln L_0 = \sum_r n_r \ln \hat{\pi}_r = N \sum_r p_r \ln \hat{\pi}_r \ , \ \ \hat{\pi}_r = \pi_r(\hat{\boldsymbol{\omega}})$$

and the maximum of $\ln PL$ under H_1 (saturated model) is

$$\ln L_1 = \sum_r n_r \ln p_r = N \sum_r p_r \ln p_r \ ,$$

where n_r is the number of times response pattern r occurs in the sample, $p_r = n_r/N$ and N is the sample size. The likelihood ratio (LR) test statistic is

$$\chi^2_{\mathrm{LR}} = 2 \sum_r n_r (\ln p_r - \ln \hat{\pi}_r) = 2N \sum_r p_r (\ln p_r - \ln \hat{\pi}_r) \ \ . \tag{20}$$

Under H_0 we need to work out its distribution. If full information ML is used then this is distributed approximately as χ^2 with degrees of freedom equal to the number of independent response patterns minus one minus the number of elements of $\boldsymbol{\omega}$.

Alternatively, one can use the goodness-of-fit (GF) test statistic

$$\chi^2_{\mathrm{GF}} = \sum_r [(n_r - N\hat{\pi}_r)^2/(N\hat{\pi}_r)] = N \sum_r (p_r - \hat{\pi}_r)^2/\hat{\pi}_r \ . \tag{21}$$

Both statistics (20) and (21) have the same asymptotic distribution under H_0.

In principle, these tests are possible to use with full information ML (FIML). They cannot be used with the pseudo-likelihood approach because this does not maximize an overall likelihood function, so the $\hat{\pi}_r$ are not directly computed. In practice, however, these tests do not work well because in real data there are often many zero and small frequencies n_r which will distort the approximation to the chi-square distribution [57].

Nevertheless, under the pseudo-likelihood estimation framework the pseudo-likelihood ratio test (PLRT) is written as

$$\chi^2_{\text{PLLT}} = 2 \times (pl(\hat{\boldsymbol{\omega}}; \boldsymbol{y}) - pl(\tilde{\boldsymbol{\omega}}; \boldsymbol{y})), \tag{22}$$

where $pl(\hat{\boldsymbol{\omega}}; \boldsymbol{y})$ and $pl(\tilde{\boldsymbol{\omega}}; \boldsymbol{y})$ are the log pseudo-likelihood values under the alternative and null hypothesis respectively.

It has been shown that the asymptotic distribution of composite likelihood (pseudo-likelihood) ratio statistic is a weighted sum of χ^2_1 distribution [47, 27, 53, 62, 41]. We leave the development and studying of the performance of PLRT for testing overall fit and nested association models for future research.

7.2 Composite Likelihood Model Selection Criteria

Based on the results of [63], the Akaike pseudo-likelihood (PL) information criterion, AIC_{PL} for the CL framework is defined as:

$$AIC_{PL} = -pl\left(\hat{\boldsymbol{\omega}}_{PL}; \mathbf{y}\right) + tr(\hat{J}(\hat{\boldsymbol{\omega}}_{PL})\hat{H}^{-1}(\hat{\boldsymbol{\omega}}_{PL})), \tag{23}$$

and, based on the results found in [25], the PL Bayesian information criterion, BIC_{PL}, is defined as:

$$BIC_{PL} = -2pl\left(\hat{\boldsymbol{\omega}}_{PL}; \mathbf{y}\right) + tr(\hat{J}(\hat{\boldsymbol{\omega}}_{PL})\hat{H}^{-1}(\hat{\boldsymbol{\omega}}_{PL})) \times \log N\,, \tag{24}$$

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

8 Example

The data used here, the DASS data (retrieved July, 2020 from OpenPsychometrics.org), consist of responses collected during the period of 2017 – 2019 to 42 items, and of the 38,776 respondents, only a random sample of 1,000 were used in this example. The items were presented online to respondents in a random order. The items included in the DASS data are from scales designed to measure depression (d1–d14), anxiety (a1–a13), and stress (s1–s15). For each item, respondents were asked to consider the last week when making their responses using the following categories:

1. Did not apply to me at all
2. Applied to me to some degree, or some of the time
3. Applied to me to a considerable degree, or a good part of the time
4. Applied to me very much, or most of the time

The items are given in the appendix and in the online supplemental material, along with the data and **R** code used to fit the models to the data.

We used pseudo-likelihood estimation in this example with a relatively strong convergence criterion. We deem that a model has converged if the item with the largest change in the maximum likelihood between iterations is less than $1e-6$, which also yields changes in many parameters on the order of $1e-10$. The convergence information is given in Table 3, number of iterations ("#iter") and the value of the convergence criterion. The Rasch and independence models were only fit once; therefore, we report the convergence information from the `mnlogit` output from the stacked regression. Parameters estimates were close to the final estimations in approximately 5 iterations and the algorithm achieves convergence in less than or equal to 15 iterations. The x_j values for each item for the Rasch and GPCM models and the starting values for the ν_{ijm} parameters for the Nominal model were -0.1035098, -0.03450328, 0.03450328, and 0.1035098; that is, they sum to zero and are equally spaced.

An independence log-linear model was fit to the data as a baseline model. One and three-dimensional models corresponding to Rasch, GPCM, and Nominal models were fit to the data. Table 3 contains basic summary statistics for each model, including the number of unique parameters estimated ('# of params), the maximum of the log of the pseudo-likelihood (MLPL) function, and pseudo-likelihood information criteria, AIC and BIC (smaller is better). As expected, the uni-dimensional models fit considerably worse than the 3-dimensional models and will not be considered further. Among the 3 dimensional models, the Rasch model is not selected whether using the AIC (tends to select more complex models) or the BIC (tends to select simpler models). The $M = 3$ Nominal model has the smallest AIC_{pl} and the GPCM has the smallest BIC_{pl}. We will further study the results of the 3-dimensional Nominal and GPCM models.

Table 3: Global summary statistics and convergence information for models fit to the DASS data.

Model	M	# of parms	MLPL	Fit Statistics AIC_{PL}	BIC_{PL}	Convergence criterion	#iter
Independence	1	126	-56146	56272	113162	0	5
Rasch	1	127	-44609	44736	90095	0	6
GPCM	1	168	-44240	44408	89641	2.5e-07	14
Nominal	1	252	-44069	44321	89879	1.6e-07	14
Rasch	3	132	-42529	42661	85969	4.4e-07	6
GPCM	3	171	-42258	42429	85698	3.6e-07	14
Nominal	3	255	-42030	42285	85822	2.5e-07	15

8.1 Measures of Item Fit for the DASS Data

The analyses in this section are a combination of statistics and graphics at the item level. In Table 4 are the maximum of the likelihoods for each item from fitting models to each item in the PLE algorithm. These are given for the Nominal model and GPCM along with the differences between the models' values. These differences (i.e., Δ or -2Δ) do not meet the regularity conditions for these to be chi-square distributed because the values of the predictor variables are different for the GPCM and Nominal models (i.e., different data). However, Δ still provides information regarding which models are better fitting particular items. The sum over items of the maximum likelihoods in Table 4 equals a model's MLPL. Table 5 further summarizes the item fit statistic and contains the proportion of items within a scale that fall within ranges of the maximum likelihood values. From Tables 3 and 5, in general the items from the nominal model have larger values than items fit by the GPCM (larger is better). Furthermore, the depression items tend to be fit better than the items from the anxiety scale, and the items from the stress scale are the worse fit. Based on the difference in Table 4, some items appear to be fit equally well by the GPCM and Nominal model. In particular, the Δ's for d14, a4, and a9 equal 1.85, 1.37, and 1.77, respectively;

however, items d6, d7 and a10 all have the largest Δ values, which suggests that the Nominal model should be used (at least for these items).

Table 4: Item statistics: Values of maximums of the likelihoods (components of the PLML) for each item and the Nominal and GPCM, and the differences (Δ) between the log-likelihoods.

item	GPCM	Nominal	Δ	item	GPCM	Nominal	Δ	item	GPCM	Nominal	Δ
d1	-964.52	-957.28	7.23	a1	-1197.71	-1193.60	4.12	s1	-983.09	-974.35	8.74
d2	-1062.83	-1058.89	3.94	a2	-982.74	-974.77	7.97	s2	-1051.49	-1048.81	2.68
d3	-910.01	-905.11	4.90	a3	-973.90	-964.58	9.32	s3	-1041.61	-1036.94	4.67
d4	-844.14	-839.26	4.88	a4	-1010.04	-1008.67	1.37	s4	-1016.15	-1013.58	2.57
d5	-976.96	-969.36	7.60	a5	-964.33	-956.77	7.56	s5	-1002.18	-995.89	6.28
d6	-891.23	-878.79	12.44	a6	-1097.10	-1092.13	4.98	s6	-1163.45	-1161.15	2.30
d7	-806.19	-795.27	10.93	a7	-1014.66	-1008.84	5.82	s7	-1141.78	-1136.96	4.82
d8	-1009.82	-1001.83	7.99	a8	-815.02	-809.09	5.92	s8	-1020.17	-1014.21	5.96
d9	-969.42	-966.66	2.76	a9	-1118.39	-1116.62	1.77	s9	-1037.74	-1032.84	4.90
d10	-1028.59	-1024.36	4.23	a10	-968.84	-956.47	12.37	s10	-1046.11	-1042.29	3.82
d11	-884.02	-877.86	6.16	a11	-985.39	-977.32	8.07	s11	-1090.11	-1086.84	3.27
d12	-944.48	-936.61	7.87	a12	-1090.40	-1087.02	3.38	s12	-1103.96	-1101.73	2.23
d13	-872.29	-865.68	6.61	a13	-1012.65	-1006.15	6.50	s13	-1010.43	-1008.02	2.41
d14	-1068.67	-1066.82	1.85					s14	-1072.32	-1069.73	2.59
								s15	-1013.22	-1011.09	2.12

Table 5: Summary of proportion of items fit in terms of ranges of the values of the items' maximum of the log-likelihoods.

		Range of log-likelihoods			
Model	Scale	> -799	-800 to -899	-900 to -999	< -1000
Nominal	Depression	.07	.29	.36	.29
	Anxiety	.00	.08	.38	.54
	Stress	.00	.00	.13	.87
GPCM	Depression	.00	.36	.36	.29
	Anxiety	.00	.08	.38	.54
	Stress	.00	.00	.07	.93

The difference between the GPCM and Nominal models is that the former has linear restrictions on the scale values. To determine whether this restriction is reasonable, we first examine statistics and then graphics. For the nominal model, a measure how strongly an item is related to the latent variable that is is directly related to is η_i ([3])

$$\eta_{im} = \sqrt{\sum_{j_i} \nu^2_{ij_im}}$$

. When the location identification constraint is $\sum_j \nu_{ijm} = 0$, η_{im} is proportional to the standard deviation of ν_{ijm}s. Alternatively, we can fit the GPCM model to the data and examine the $\hat{a}_{im}$ parameters, which when ν_{ijm} are equally spaced will be highly correlated with η_{im}s. In our example, $r(\eta_{im}, \hat{a}_{im}) = .996$, which suggests that the ν_{ijm} maybe equally spaced and the x_j's used to fit the GPCM model are reasonable for the data. Computing η_{im} or a_{im} only requires fitting one model. The η_{im}s and $\hat{a}_{im}$s are given in Table 6. Whether using η_{im} or $\hat{a}_{im}$, the items that are most strongly related to their respective latent traits are d4, d7, and a10, which indicate that both models are identifying the same items are being highly related to the latent variable and therefore to each other. These statistics indicate the magnitude of association between items within a scale. For example, among the depression items, the relationship between d4 ("I felt sad and depressed") and d7 ("I felt that life wasn't worthwhile") is larger than that from any other two items, and the smallest is between a1 ("I was aware of dryness of my mouth") and a6 ("I felt scared without any good reason"). Comparing across scales, it appears that the items on the depression scale are more highly related to the depression trait and between each other (mean $\hat{a}_i = 5.08$), followed by anxiety items related to the anxiety trait (mean $\hat{a} = 4.14$). The least strongly related to the latent trait and among each other are the stress items (mean $\hat{a} = 4.08$).

Table 6: Item statistics: The slopes $\hat{a}_{im}$ from the GPCM and the η_{im} statistics from the Nominal models, which reflect the strength of the relationship between the items and the latent variables, as well as between items themselves.

item	a_{i1}	η_{i1}	item	a_{i2}	η_{i2}	item	a_{i3}	η_{i3}
d1	4.77	0.77	a1	2.14	0.34	s1	5.03	0.83
d2	3.59	0.57	a2	4.21	0.68	s2	4.20	0.66
d3	5.71	0.89	a3	4.17	0.64	s3	4.21	0.66
d4	6.60	1.05	a4	5.67	0.88	s4	4.70	0.74
d5	4.63	0.74	a5	3.65	0.61	s5	4.46	0.73
d6	5.62	0.91	a6	2.41	0.37	s6	2.62	0.41
d7	7.45	1.19	a7	5.24	0.81	s7	2.96	0.47
d8	4.29	0.69	a8	3.67	0.63	s8	4.58	0.72
d9	4.73	0.74	a9	3.28	0.51	s9	4.21	0.68
d10	4.00	0.63	a10	6.00	0.94	s10	4.24	0.68
d11	5.80	0.89	a11	5.48	0.85	s11	3.66	0.57
d12	5.03	0.78	a12	4.10	0.64	s12	3.37	0.52
d13	5.56	0.86	a13	3.79	0.58	s13	4.88	0.76
d14	3.32	0.54				s14	3.72	0.59
						s15	4.35	0.67

To further investigate whether the GPCM or Nominal models are better for particular items, we examine the scale value estimates from the nominal model to see if they are indeed linear with respect to equally spaced numbers. Estimated scale values from the nominal model (solid circles and lines) can be plotted against integers with linear regression drawn (dashed lines) in the same plot. Examples for four items are given in Figure 7. The categories for all items are clearly ordinal and increase with values of the integers. The values for aggression item a9 (upper left) are coincident with the regression line, which from Table 4 has $\Delta = 1.77$. The scale values for the other items a10, d7 and d8, deviate from their regression lines and had Δ values of 12.37, 10.93, and 7.99, respectively. Scale values for item a9 and possibly item d8 might be satisfactorily modeled using equally spaced category scores as in the GPCM. Another aspect to consider is slope of the lines, which would correspond to a_{im} in a GPCM model. Among these four items, a9 (smallest slope) appears to be more weakly related to it's the latent variable; whereas, item d7 (the steepest slope) is the most strongly related to it's latent variable. These results further confirm our conclusions based on statistics in Table 6.

The last analyses looks at the correspondence between data and fitted values (probabilities). In logistic regression with continuous predictors and in IRT, the continuous values can be collapsed into groups or bins. Estimates of θ_m were computed using (10) and then grouped into 10 categories. The observed proportions within a group who select a category within a group and the fitted probabilities for the group were plotted against the mean of the continuous $\hat{\theta}_m$ for the groups. Two examples of such plots are given in Figure 8 where the data are points and lines are fitted values from the Nominal model (one line per category). Item d7 has the largest log-likelihood value in Table 4 and the largest η_{im} and $\hat{a}_{im}$ in Table 6. This appears to be the item fit best according to our statistics and there is a close correspondence between the fitted probabilities (lines) and the observed proportions (points). Item s6 has the smallest log-likelihood and one of the smallest η_i and $\hat{a}_i$, which indicate that this items in one of the fit worse by the model. The model for item s6 under predicts the first (squares) category and last (diamonds) category, which further confirms that this item is not fit well by the Nominal model. Item s6 is not fit well by the nominal model and will not fair any better under a GPCM.

Computing $\hat{\theta}_m$ using (10) makes use of responses to all items where items were weighted by the conditional covariances. Since we set $\sigma_{mm} = 1$ for identification, we actually estimated conditional correlation matrices between traits within response pattern. These estimated conditional correlation matrices from the Nominal and GPCM models are very similar,

$$\hat{\Sigma}_{nom} = \begin{pmatrix} 1.000 & 0.038 & 0.094 \\ 0.038 & 1.000 & 0.290 \\ 0.094 & 0.290 & 1.000 \end{pmatrix} \quad \text{and} \quad \hat{\Sigma}_{gpc} = \begin{pmatrix} 1.000 & 0.047 & 0.099 \\ 0.047 & 1.000 & 0.299 \\ 0.099 & 0.299 & 1.000 \end{pmatrix},$$

where the subscript *nom* is for the Nominal model and *gpc* is for the GPCM. The conditional correlations between depression and anxiety and between

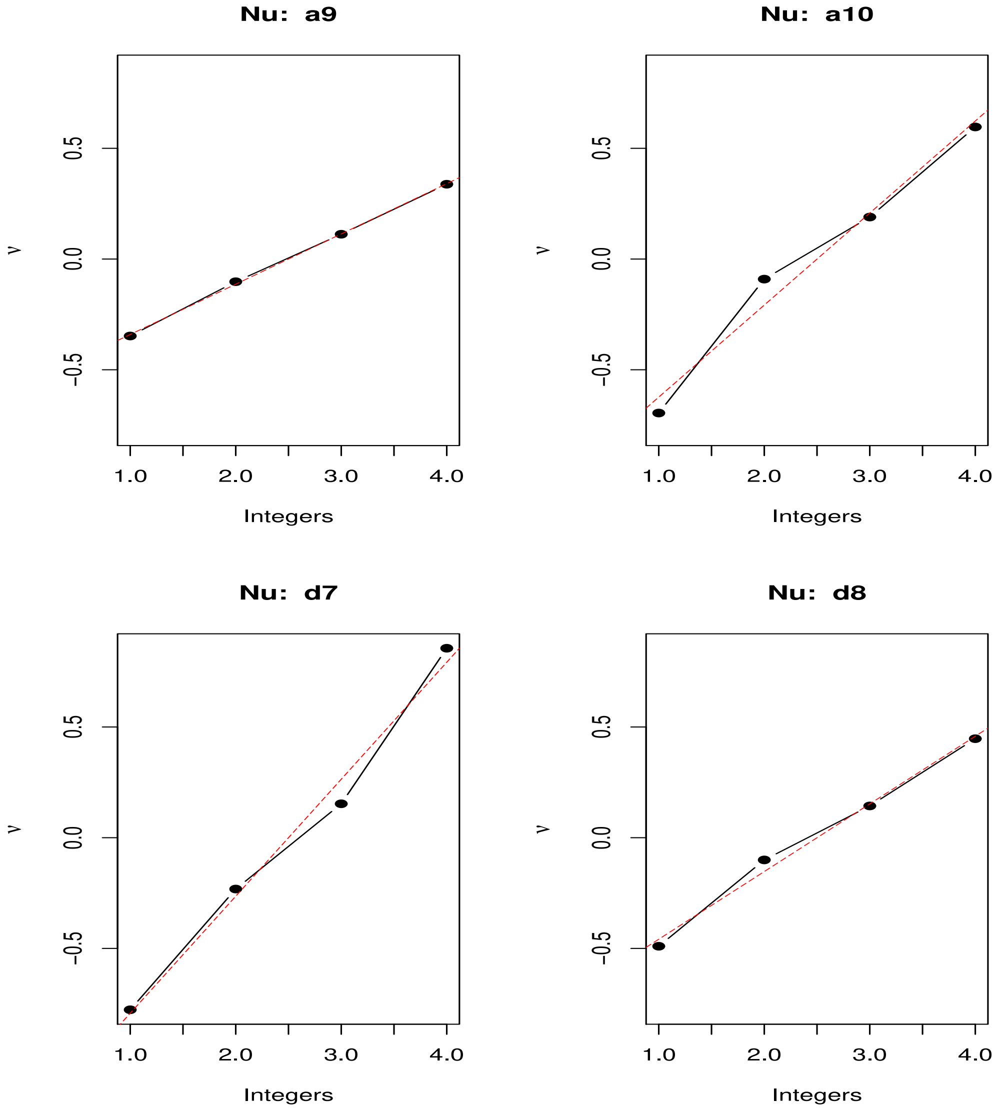
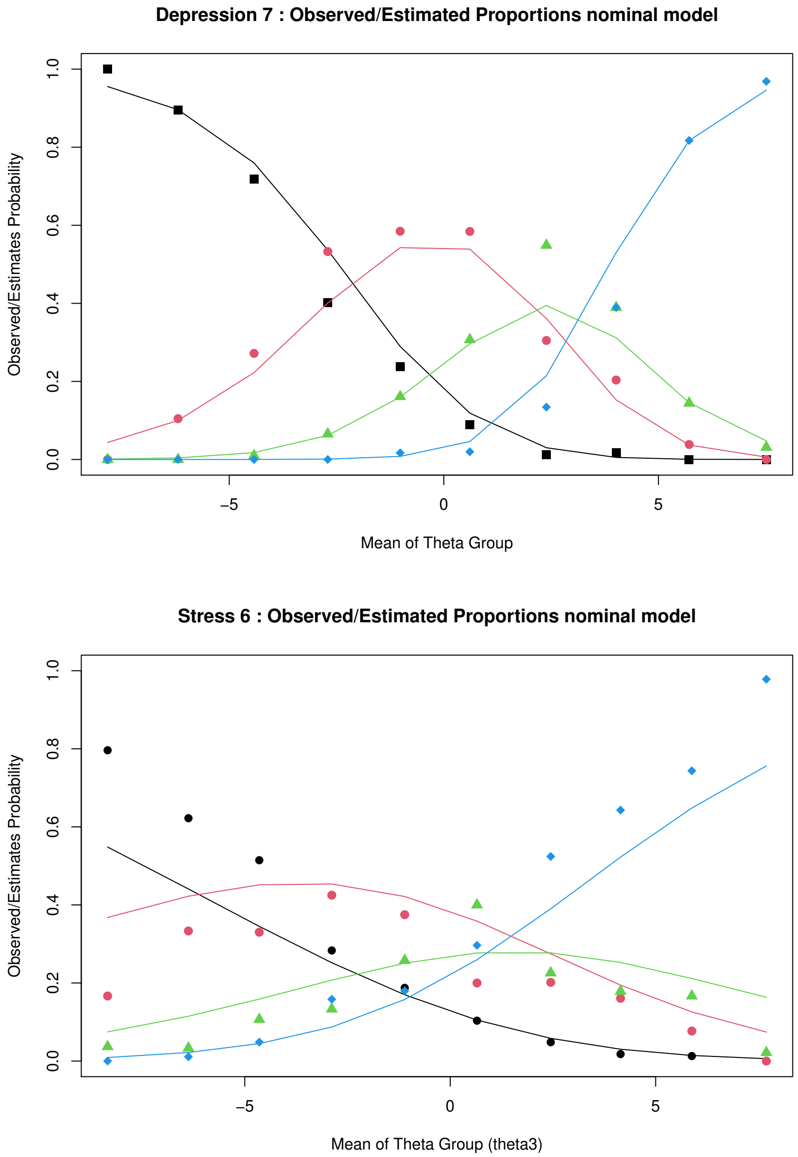


Figure 7: Estimated scale values ν_{ijm} (solid points and lines) from the Nominal response model for two aggression items (top) and two depression items (bottom) plotted against integers with linear regression lines (dashed lines).



Depression 7 : Observed/Estimated Proportions nominal model

Stress 6 : Observed/Estimated Proportions nominal model

Figure 8: For depression item d7 (top) and stress item s6 (bottom), observed proportions (points) and fitted probabilities (lines) are plotted against the mean of θ_m where estimate of $\hat{\theta}_m$ has been collapsed in to groups. Symbols for categories are 1=squares, 2=dots, 3=triangles, and 4=diamonds

depression and stress appear relatively small.

Small conditional correlations do not imply that the marginal correlations are small. The marginal correlation matrices between $\hat{\theta}_m$ from the Nominal and GPCM model are

$$\hat{R}_{nom} = \begin{pmatrix} 1.000 & 0.774 & 0.827 \\ 0.774 & 1.000 & 0.954 \\ 0.827 & 0.954 & 1.000 \end{pmatrix} \quad \text{and} \quad \hat{R}_{gpc} = \begin{pmatrix} 1.000 & 0.781 & 0.830 \\ 0.781 & 1.000 & 0.957 \\ 0.830 & 0.957 & 1.000 \end{pmatrix}.$$

Table 7: Alternative models for DASS data with $M = 2$ where stress and anxiety are one scale, and $M = 3$ where $\sigma_{12} = 0$.

				Fit Statistics		Convergence	
Model	M	#parm	MLPL	AIC_{PL}	BIC_{PL}	criteria	#iter
Rasch	2	129	-42840	42969	86571	4.4e-07	6
GPCM	2	169	-42490	42659	86147	3.2e-07	13
Nominal	2	253	-42265	42518	86278	5.1e-07	13
Rasch	3	131	-42536	42667	85983	4.2e-07	6
GPCM	3	170	-42268	42438	85711	3.2e-07	14
Nominal	3	254	-42036	42290	85827	6.7e-07	15

The order in terms of magnitude of the marginal and conditional correlations have the same pattern (i.e., largest is for anxiety and stress, and the smallest is for depression and anxiety), but are considerably larger than the conditional values.

The large marginal correlations between anxiety and stress suggest that perhaps these are not distinct constructs and a two dimensional model maybe sufficient. The data were re-analyzed using a two dimensional Rasch, GPCM and nominal model, which each has 2 fewer parameters. The statistics for these models are reported in Table 7, but the new models fit the data worse than our original three dimensional models (i.e., orginal models have smaller AIC_{PL} and BIC_{PL}. This results occurred because the conditional correlations (i.e., .290 and .299) are relatively small. It is important to point out that We do not set the marginal correlations, but rather the conditional correlations (or covariances). If a conditional correlation is close to 1, then a 2-dimensional model might be better than a 3-dimensional one.

In conclusion, our analysis confirmed our conjecture that the items represent three correlated constructs, as well as the excepted ordering of the category scores. For some items, the relative spacing between items is roughly equal but not for all, which suggest that the nominal model is the best model. We also detected some items that did not fit the data very well (e.g., s6). The items on the depression scale are more closely related to the latent variable of depression, and thus they are also more closely related to each other. The stress items have weaker association with the stress latent variable and also have weaker associations between the stress items themselves. Due to small values of $\hat{\sigma}_{mm'}$ for depression and anxiety and for depression and stress, the estimated value of depression depends mostly on the responses to the depression items. On the other hand, the larger value of $\hat{\sigma}_{mm'}$ for stress and anxiety, indicate that each provide more information in the estimating of values on the stress and anxiety constructs.

9 Conclusion/Discussion

When there are interactions between categorical variables, the AMs presented in this chapter are just one way to describe the nature and strength of associations. Other possibilities not covered in this chapter and often missing in the literature on AMs include (multiple) correspondence analysis ([33, 32]), optimal scaling ([50]), and dual scaling ([52, 51]), which are all scaling methods that in the case of 2-way tables are all essentially the same and yield very similar results. For a history of these methods see [50]). These scaling methods are data analytic techniques without distributional assumptions and statistical tests of model goodness-of-fit to the data. Other related methods that are statistical models are canonical correlation models ([30]) and latent class models. For 2-way tables, latent class models with 2 latent classes and the correlation models yield similar results.

The AMs discussed here provide useful representations of interactions between categorical variables; however, they have also been derived from an underlying theoretical model (e.g., IRT). Although we focus on models with an underlying simple structure, the log-multiplicative AMs afford more complex structures, including models where items load on multiple latent variables more exploratory analysis and bi-factor structures. We can also add covariates to the AMs. Pseudo-likelihood estimation can be used for these more complex structures.

The data analysis examples illustrate how measures of fit such as item log-likelihood differences, transformed scores, fitted proportions can be used to check item misfit, and strength of an item in measuring a latent variable. Furthermore, AM provide information about the arbitrary selected scores when choosing the response categories of an item. This is also what IRT modelling tries to achieve by estimating discrimination coefficients for each item or each score in the nominal case. Goodness-of-fit tests and model selection criteria can be developed under the pseudo-likelihood estimation framework presented here to test overall fit and select among nested and non-nested models.

The connection between IRT and association models provides multiple insights on the same data analysis problem, i.e. on how to model and interpret associations depending the aim of our analysis. The availability of statistical software and the extension to high-dimensional tables for multi-category variables is a very useful tool to data analysts who want to have the flexibility of choosing and estimating a suitable model to high dimensional data.

Appendix A: R Code for Estimating Models

To facilitate the use of the models and for reproducible we have included **R** code and data for results presented in this chapter. Some are below and all are available on the online book supplemental material.

2-way Tables

The code used to produce results presented for the MOOC example for the is available on the book's online supplemental material, as well a function that computes various fit statistics reported.

Higher Way Tables using pleLMA package

To aid in fitting the AM models for moderate to large tables for dichotomous or multi-category variables described in this paper, we have written a package that uses pseudo-likelihood estimation to fit models to data. Note that "phi" in the `pleLMA` package is what we have called "Σ" in this chapter.

```
# Install and load
```

```
R/pleLMA_0.1.0.tar.gz
library(pleLMA)

# Set up data
data(dass)
inData <- dass

# Trait by trait adjacency matrix
inTraitAdj <- matrix(c(1,1,1, 1,1,1, 1,1,1), nrow=3 ,ncol=3)

# Item by trait adjacency matrix
d <- matrix(c(1, 0, 0),nrow=14,ncol=3,byrow=TRUE)
a <- matrix(c(0, 1, 0),nrow=13,ncol=3,byrow=TRUE)
s <- matrix(c(0, 0, 1),nrow=15,ncol=3,byrow=TRUE)
das <- list(d, a, s)
inItemTraitAdj  <- rbind(das[[1]], das[[2]], das[[3]])

# Fit models with defaults
# independence
ind <- ple.lma(inData, inItemTraitAdj, inTraitAdj, model.type="independence")

# rasch
r3 <- ple.lma(inData, inItemTraitAdj, inTraitAdj, model.type="rasch")

# gpcm
g3 <- ple.lma(inData, inItemTraitAdj, inTraitAdj, model.type="gpcm")

# nominal
n3 <- ple.lma(inData, inItemTraitAdj, inTraitAdj, model.type="nominal")

# some functions for Nominal and GPCM models
# -- output available
summary(n3)

# -- summary of model fit, input, and global statistics
summaryModel(n3)

# -- more details on convergence
```

```
converged <- convergence.stats(n3$item.log, n3$nitems, n3$nless)

iterationPlot(history=n3$item.log, n3$nitems, n3$ncat, n3$nless,
              n3$ItemNames)

iterationPlot(history=n3$phi.log, n3$nitems, n3$ncat, n3$nless,
             ItemNames=n3$ItemNames, Maxnphi=n3$Maxnphi,
             PhiNames=n3$PhiNames)

# --- For plot of scale values versus integers for Nominal models
scalingPlot(n3)

# -- Matrix of max loglike and estimated parameters for each item
n3$estimates

# -- Estimated conditional correlation matrix
n3$Phi.mat

# -- Estimate of latent variables
theta.n3 <- theta.estimates(n3, inData, scores=n3$estimates)
```

Appendix B: DASS Data

For each item, respondents were asked consider the last week when making their responses using the rating scale

1. Did not apply to me at all
2. Applied to me to some degree, or some of the time
3. Applied to me to a considerable degree, or a good part of the time
4. Applied to me very much, or most of the time

Depression Scale:

d1 I couldn't seem to experience any positive feeling at all.

d2 I just couldn't seem to get going.

d3 I felt that I had nothing to look forward to.

d4 I felt sad and depressed.

d5 I felt that I had lost interest in just about everything.

d6 I felt I wasn't worth much as a person.

d7 I felt that life wasn't worthwhile.

d8 I couldn't seem to get any enjoyment out of the things I did.

d9 I felt down-hearted and blue.

d10 I was unable to become enthusiastic about anything.

d11 I felt I was pretty worthless.

d12 I could see nothing in the future to be hopeful about.

d13 I felt that life was meaningless.

d14 I found it difficult to work up the initiative to do things.

Anxiety Scale:

a1 I was aware of dryness of my mouth.

a2 I experienced breathing difficulty (eg, excessively rapid breathing, breathlessness in the absence of physical exertion).

a3 I had a feeling of shakiness (eg, legs going to give way).

a4 I felt that I was using a lot of nervous energy.

a5 I had a feeling of faintness.

a5 I perspired noticeably (eg, hands sweaty) in the absence of high temperatures or physical exertion.

a6 I felt scared without any good reason.

a7 I had difficulty in swallowing.

a8 I was aware of the action of my heart in the absence of physical exertion (eg, sense of heart rate increase, heart missing a beat).

a9 I felt I was close to panic.

a10 I felt terrified.

a11 I was worried about situations in which I might panic and make a fool of myself.

a12 I experienced trembling (eg, in the hands).

Stress Scale:

s1 I found myself getting upset by quite trivial things.

s2 I tended to over-react to situations.

s3 I found it difficult to relax.

s4 I found myself in situations that made me so anxious I was most relieved when they ended.

s5 I found myself getting upset rather easily.

s6 I found myself getting impatient when I was delayed in any way (eg, elevators, traffic lights, being kept waiting).

s7 I felt that I was rather touchy.

s8 I found it hard to wind down.

s9 I found that I was very irritable.

s10 I found it hard to calm down after something upset me.

s11 I feared that I would be thrownnoff by some trivial but unfamiliar task.

s12 I found it difficult to tolerate interruptions to what I was doing.

s13 I was in a state of nervous tension.

s14 I was intolerant of anything that kept me from getting on with what I was doing.

s15 I found myself getting agitated.

References

[1] Carolyn J. Anderson, Zhusan Li, and Jeoren J.K. Vermunt. Estimation of models in a rasch family for polytomous items and multiple latent variables. *Journal of Statistical Software*, 2007.

[2] Carolyn J. Anderson, Jay V. Verkuilen, and Buddy Peyton. Modeling polytomous item responses using simultaneously estimated multinomial logistic regression models. *Journal of Educational and Behavioral Statistics*, 35:422–452, 2010.

[3] Carolyn J. Anderson and Jeroen J.K. Vermunt. Log-multiplicative association models as latent variable models for nominal and/or ordinal data. *Sociological Methodology*, 30:81–121, 2000.

[4] Carolyn J. Anderson and Hsiu-Ting Yu. Log-multiplicative association models as item response models. *Psychometrika*, 72:5–23, 2007.

[5] Carolyn J. Anderson and Hsiu-Ting Yu. Theoretical and empirical properties of log-multiplicative association models as multidimensional nominal item response models. *Manuscript*, 2021.

[6] Barry C. Arnold and David Straus. Pseudolikelihood estimation: Some examples. *The Indian Journal of Statistics*, 53:233–243, 1991.

[7] Mark Becker. On the bivariate normal distribution and association models for ordinal categorical data. *Statistics & Probability Letters*, 8:435 – 440, 1989.

[8] Julian Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B*, 36:192–225, 1974.

[9] Julian Besag. Statistical analysis of non-lattice data. journal of the royal statistical society. *Journal of the Royal Statistical Society: Series D (The Statistician*, 24:179–195, 1975.

[10] Suma Bhat, Carolyn J. Anderson, Wes Crues, Lawrence Angrave, Najmuddin Shaik, and Genevieve G.M. Hendricks. Know your audience: Who is served and their engagement levels in moocs. *Manuscript*, 2020.

[11] Milan Bouchet-Valat, Heather Turner, Michael Friendly, Jim Lemon, and Cabor Csardi. *Package 'logmult'*, 2020. R package version 0.7.21.

[12] Hua-Hua Chang. The asymptotic posterior normality of the latent trait for polytomous IRT models. *Psychometrika*, 58:445—463, 1986.

[13] Hua-Hua Chang and William Stout. The asymptotic posterior normality of the latent trait in an IRT model. *Psychometrika*, 58:37–52, 1003.

[14] Yunxio Chen, Xiaoou Li, Jingchen Liu, and Zhiliang Ying. Robust measurement via a fused latent variable and graphical item response theory model. *Psychometrika*, 85:538–562, 2018.

[15] Yves Croissant. Estimation of random utility models in R: The `mlogit` package. *Journal of Statistical Software*, 95(11):1–41, 2020.

[16] Jimmy de la Torres, Hau Song, and Yuan Hong. Comparison of four methods of IRT subscoring. *Applied Psychological Measurement*, 35:296–316, 2001.

[17] Mark de Rooij. The analysis of change, netwon's law of gravity and association models. *Journal of the Royal Statistical Society: Statistics in Society, Series A*, 171:137–157, 2007.

[18] Mark de Rooij. Ideal point discriminant analysis revisited with a special emphasis on visualization. *Psychometrika*, 74:317–330, 2009.

[19] Mark de Rooij and Willem Heiser. Graphical representations and odds ratios in a distance-association model for the analysis of cross-classified data. *Psychometrika*, 70:99–122, 2005.

[20] David Edwards. *Introduction to Graphical Models*. Springer, NYC, 2nd edition, 2000.

[21] Fred Van Eeuwijk. Multiplicative interaction in generalized linear models. *Biometrics*, 51:1017–1032, 1995.

[22] Martin Elff. Multinomial logit models, with or without random effects or overdispersion. *Journal of Statistical Software*, 2020.

[23] Sasha Epskamp. *Package 'Ising Sampler'*, 2020. R package version 0.2.1.

[24] Stephen E. Fienberg and Alessandro Rinaldo. Maximum likelihood estimation in log-linear models. *The Annals of Statistics*, 40:996–1023, 2012.

[25] Xin Gao and Peter X.-K. Song. Composite likelihood Bayesian information criteria for model selection in high dimensional data. *Journal of the American Statistical Association*, 105(492):1531–1540, 2010.

[26] Andrew Gelman and Terence P. Speed. Characterizing a joint distribution by conditionals. *Journal of the Royal Statistical Association: Series B*, 55:185–188, 1993.

[27] Helena Geys, Geert Molenberghs, and Louise M. Ryan. Pseudolikelihood modeling in multivariate outcomes in developmental toxicology. *Journal of the American Statistical Association*, 94:734–745, 1999.

[28] Leo A. Goodman. Simple models for the analysis of association in cross-classifications having ordered categories. *Journal of the American Statistical Association*, 74(367):537–552, 1979.

[29] Leo L. Goodman. Association Models and the Bivariate Normal for Contingency Tables With Ordered Categories. *Biometrika*, 68:347–355, 1981.

[30] Leo L. Goodman. The analysis of cross-classified data having ordered and/or unordered categories: Association models, correlation models, and asymmetry models for contingency tables with or without missing entries. *The Annals of Statistics*, 75:1–24, 1985.

[31] Leo L. Goodman. Some useful extensions of the usual correspondence analysis approach and the usual log-linear models approach in the analysis of contingency tables. *International Statistical Review*, 54:243–270, 1986.

[32] Michael Greenacre. *Correspondence Analysis in Practice*. Chapman & Hall/CRC Interdisciplinary Statistics. CRC Press, 2017.

[33] Michael Greenacre and Jorg Blasius. *Multiple Correspondence Analysis and Related Methods*. Chapman & Hall/CRC Statistics in the Social and Behavioral Sciences. CRC Press, 2006.

[34] Asad Hasan, Zhiyu Wang, and Alireza S. Mahani. Fast estimation of multinomial logit models: `R` package `mnlogit`. *Journal of Statistical Software*, 75:1–24, 2016.

[35] David J. Hessen. Fitting and testing conditional multinomial partial credit models. *Psychometrika*, 77:693–709, 2012.

[36] Paul W. Holland. The dutch identity: A new tool for the study of item response models. *Psychometrika*, 55:5–18, 1990.

[37] International Machine Learning Society. *Distributed Parameter Estimation via Pseudo-likelihood*, 2012.

[38] Karl G. Jöreskog. A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, 34:183–202, 1969.

[39] Karl G. Jöreskog. Structural equation models in the social sciences: Specification, estimation and testing. In Karl G. Jöreskog and Dag Sörbom, editors, *Advances in factor analysis and structural equation models*, pages 105–127. Cambridge, Mass.: Abt Books, 1979.

[40] Maria Kateri. *Contingency Table Analysis: Methods and Implementation Using R*. Birkhäuser/Springer, New York, 2014.

[41] Myrsini Katsikatsou and Irini Moustaki. Pairwise likelihood ratio tests and model selection criteria for structural equation models with ordinal variables. *Psychometrika*, 81:1046–1068, 2016.

[42] Henk Kelderman. Multidimensional rasch models for partial-credit scoring. *Applied Psychological Measurement*, 20:1–10, 1996.

[43] Henk Kelderman and Carl P. M. Rijkes. Loglinear multidimensional IRT models for polytomous scored items. *Psychometrika*, 59:149–176, 1994.

[44] Mia Kornely and Maria Kateri. Asymptotic posterior normality of multivariate latent traits in an IRT model. *Manuscript submitted for publication*, 2020.

[45] Joost Kruis and Gunter Maris. Three representations of the ising model. *Scientific Reports*, 6:1–10, 2016.

[46] Steffen L. Lauritzen. *Graphical Models*. Oxford, Oxford, 1996.

[47] Kung-Yee Liang and Steven G. Self. On the asymptotic behaviour of the pseudolikelihood ratio test statistic. *Journal of the Royal Statistical Sociery, Series B*, 58:785–796, 1996.

[48] Bruce G. Lindsay. Composite likelihood methods. In N. U. Prabhu, editor, *Statistical Inference from Stochastic Processes*, pages 221–239. Providence, RI: Americal Mathematical Society, 1988.

[49] M. Marsman, D. Borsboom, J. Kruis, S. Epskamp, R. van Bork, L.J. Waldorp, H.L.J. van der Maas, and G. Maris. An introduction to network psychometrics: Relating ising network models to item response theory models. *Multivariate Behavioral Research*, 53:15–35, 2018.

[50] Jaquelene Muelman, Anita J. Van Der Kooij, and Willem Heiser. Principal components analysis with nonlinear optimal scaling transformations for ordinal and nominal data. In David Kaplan, editor, *The SAGE Handbook of Quantitative Methodology for the Social Sciences*, chapter 3, pages 49–72. SAGE Publications, 2004.

[51] Shizuhiko Nishisato. Dual scaling. In David Kaplan, editor, *The SAGE Handbook of Quantitative Methodology for the Social Sciences*, chapter 1, pages 2–24. SAGE Publications, 2004.

[52] Shizuhio Nishisato. *Elements of Dual Scaling*. Psychology Press, 2004.

[53] Luigi Pace, Alessandra Salvan, and Nicola Sartori. Adjusting composite likelihood ratio statistics. *Statistica Sinica*, 21:129–148, 2011.

[54] Youngshil Paek. *Pseudo-Likelihood Estimation of Multidimensional Item Response Theory Model*. PhD thesis, University of Illinois, Urbana-Champaign, 2016.

[55] Youngshil Paek and Carolyn J. Anderson. Pseudo-likelihood estimation of multidimensional response models: Polytomous and dichotomous items. In Andries van der Ark, Marie Wiberg, Steven A. Culpepper, Jeffrey A. Douglas, and Wen-Chung Wang, editors, *Quantitative Psychology — The 81st Annual Meeting of the Psychometric Society*, pages 21–30. Springer, NYC, 2017.

[56] Psychometric Society. *Network Multidimensional Item Response Models: Beyond Simple Structure*, 2017.

[57] Mark Reiser and Maria VandenBerg. Validity of the chi-square test in dichotomous variable factor analysis when expected frequencies are small. *British Journal of Mathematical and Statistical Psychology*, 47:85–107, 1994.

[58] Dror Rom and Sanat K. Sarkar. Approximating probability integrals of multivariate normal using association models. *Journal of Statistical Computation and Simulation*, 35(1-2):109–119, 1990.

[59] Yoshio Takane, Hamparsum Bozdogan, and Tadashi Shibayama. Ideal point discriminant analysis. *Psychometrika*, 52:371–392, 1987.

[60] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017.

[61] Heather Turner and David Firth. *Generalized nonlinear models in R: An overview of the gnm package*, 2020. R package version 1.1-1.

[62] Cristiano Varin, Nancy Reid, and David Firth. An overview of composite likelihood methods. *Statistica Sinica*, 21:5–42, 2011.

[63] Cristiano Varin and Paolo Vidoni. A note on composite likelihood inference and model selection. *Biometrika*, 92:519–528, 2005.

[64] Wen-Chung Wang, Po-Hsi Chen, and Ying-Yao Cheng. Improving measurement precision of test batteries using multidimensional item response models. *Psychological Methods*, 9:116–136, 2004.

[65] Yuchung J. Wang. The probability intergrals of bivarite normal distributions: A contingency table approach. *Biometrika*, 74:185–190, 1987.

[66] Yuchung J. Wang. Multivariate normal integrals and contingency tables with ordered categories. *Psychometrika*, 62:267–284, 1997.

[67] Thomas W. Yee. *VGAM: Vector Generalized Linear and Additive Models*, 2020.