Journal of Statistical Software

May 2007, Volume 20, Issue 10. *http://www.jstatsoft.org/*

Marginal Maximum Likelihood Estimation of Item Response Models in R

Matthew S. Johnson
Baruch College, The City University of New York

Abstract

Item response theory (IRT) models are a class of statistical models used by researchers to describe the response behaviors of individuals to a set of categorically scored items. The most common IRT models can be classified as generalized linear fixed- and/or mixed-effect models. Although IRT models appear most often in the psychological testing literature, researchers in other fields have successfully utilized IRT-like models in a wide variety of applications. This paper discusses the three major methods of estimation in IRT and develops R functions utilizing the built-in capabilities of the R environment to find the marginal maximum likelihood estimates of the generalized partial credit model. The currently available R packages **ltm** is also discussed.

Keywords: item response theory, partial credit model, two-parameter logistic model, mixed effects models, marginal maximum likelihood, Gauss-Hermite quadrature.

1. Introduction to item response theory models

Item response theory (IRT) models are a class of statistical models used by researchers to describe the response behaviors of individuals to a set of categorically scored items. The most common IRT models can be classified as generalized linear fixed- and/or mixed-effect models. Although IRT models appear most often in the psychological testing literature, researchers in other fields have successfully utilized IRT-like models in a wide variety of applications. Fienberg, Johnson, and Junker (1999) employ an item response model for population size estimation when the assumption of homogeneous capture probabilities fails. Sinharay and Stern (2002) use an item response model to investigate whether the clutch, or family a baby turtle belongs to plays any role in whether or not the turtle survives. The USDA utilizes an IRT model for the measurement of a construct they call *Food Insecurity*, a measure of one's ability to obtain enough food to live a healthy life. The edited volume *Rasch Measurement*

in Health Sciences (Bezruczko 2005) discusses the use of IRT models in a number of health science disciplines.

To formalize the item response problem, let X_{vi} be the score of individual $v \in \{1, \ldots, N\}$ to item $i \in \{1, \ldots, J\}$, scored on a discrete scale from $m = 0, \ldots, K_i$. Further let $P_{im}(\theta_v) \equiv Pr\{X_{vi} = m \mid \theta_v\}$, denote the mth *category response function* for item i. When item i is dichotomous we often denote the first category response function by $P_i(\theta) \equiv P_{i1}(\theta)$ and the 0^{th} category response function by $P_{i0}(\theta) = 1 - P_i(\theta)$; $P_i(\theta)$ is often called the *item response function* (IRF) for item i or the *item characteristic curve* for item i.

A number of item response models exist in the statistics and psychometric literature for the analysis of multiple discrete responses. The models typically rely on the following assumptions:

- *Unidimensionality (U):* There is a one-dimensional, unknown quantity associated with each respondent in the sample that describes the individuals propensity to endorse the items in the survey (or exam). Let θ_v denote the propensity of individual v.
- *Conditional Independence (CI):* Given an individual's propensity θ, the elements of the item response vector for respondent v, $\boldsymbol{X}_v = (X_{v1}, \ldots, X_{vJ})^\top$, are independent.
- *Monotonicity (M):* $Pr\{X_{vi} > t \mid \theta_v\}$ is a non-decreasing function of an individual's propensity θ_v, for all i and all $t \in \mathbb{R}$. Respondents with high propensities are more likely to endorse items than those with low propensities.

In educational testing psychometricians often refer to the propensity θ_v as the latent ability, or proficiency of individual v.

The monotonicity assumption (M) allows us to use the observed item response vector for individual v ($\boldsymbol{x}_v$) as repeated measures of the latent variable θ. In fact, in the dichotomous case, under the conditions U, CI and M the *total score* for individual v, defined as $X_{v+} = \sum_{i=1}^{J} X_{vi}$ has a monotone likelihood ratio in θ (Grayson 1988; Huynh 1994). That is

$$\frac{Pr\{X_{v+} > s \mid \theta\}}{Pr\{X_{v+} > r \mid \theta\}} \text{ is increasing in } \theta \text{ for } s > r,$$

and the score X_{v+} consistently orders individuals by their latent variable θ

This paper discusses the use of R (R Development Core Team 2007) for the estimation of item response models. Section 2 reviews models for the analysis of polytomous item responses. Section 3 discusses the most common methods for the estimation of such item response models. Section 4 describes the **ltm** and **gpcm** packages for the estimation of item response models; **gpcm** contains functions for the estimation of the generalized partial credit model (described below) and **ltm** (Rizopoulos 2006) is a multi-purpose package for the estimations of latent trait models, including the graded response model described below. Section 5 demonstrates the use of the packages to simulate and analyze data, and analyzes a data set from the USDA's survey of food insecurity. The paper concludes in Section 6 with a discussion of the limitations of the two packages.

2. Item response models

2.1. Models for polytomous data

Partial credit models

The (non-parametric) partial credit model (np-PCM; Hemker, Sijtsma, Molenaar, and Junker 1997) assumes that the adjacent category logits are monotone increasing functions of the latent propensity θ, that is, they assume

$$\log\left\{\frac{P[X_{vi}=m|\theta, X_{vi}\in\{m-1,m\}]}{P[X_{vi}=m-1|\theta, X_{vi}\in\{m-1,m\}]}\right\}=\psi_{im}(\theta)$$

is an increasing function for all $i\in\{1,\ldots,J\}$ and $m\in\{1,\ldots,K_i\}$; by definition $\psi_{i0}(\theta)=0$ for all i and all θ.

The adjacent-category logit functions $\psi_{im}(\theta)$ describe how the probability of responding in category m changes relative to the probability of responding in category $m-1$ as a function of θ. Let β_{im} denote the point that satisfies $\psi_{im}(-\beta_{im})=0$. Then an individual with propensity $\theta=-\beta_{im}$ has equal probabilities of responding in categories m and $m-1$ on item i. Because of the monotonicity of the adjacent category function ψ_{im} any individual with propensity $\theta>(<)-\beta_{im}$ is more (less) likely to respond in category m than category $m-1$.

The category-response functions (CRFs) resulting from the definition above are

$$P_{im}(\theta)=\frac{\exp\left\{\sum_{k=0}^{m}\psi_{ik}(\theta)\right\}}{\sum_{h=0}^{K_j}\exp\left\{\sum_{k=0}^{h}\psi_{ik}(\theta)\right\}} \tag{1}$$

The parametric generalized partial credit model (GPCM; Muraki 1992) assumes that the adjacent category logit function $\psi_{im}(\theta)$ is a linear function of θ; the parametrization utilized herein defines

$$\begin{aligned}\psi_{im}(\theta) &= \alpha_i\theta+\beta_i+\delta_{im}\\ &= (\theta,1,\boldsymbol{e}_m^\top)^\top\begin{pmatrix}\alpha_i\\ \beta_i\\ \boldsymbol{\delta}_i\end{pmatrix},\end{aligned} \tag{2}$$

where $\boldsymbol{e}_m$ is a K_i-vector with a one in the mth position and zeroes in the remaining K_i-1 positions. Throughout the paper, we refer to the parameters α_i, β_i, and the vector $\boldsymbol{\delta}_i$ as the slope, intercept, and item-step parameters for item i. The original partial credit model introduced by Masters (1982) assumed that the slopes were equal across items, i.e., $\alpha_1=\alpha_2=\cdots=\alpha_J$.

Graded response models

The partial credit models are by no means the only models available for the analysis of polytomous item response data. One popular alternative class of models is the graded response

models, which assume that the log-odds of scoring m or higher on item i is an increasing function of the latent propensity θ. The parametric graded response model (GRM; Samejima 1969) assumes that these log-odds are linear in the latent trait, i.e.,

$$\log\left\{\frac{Pr[X_{vi} \geq m \mid \theta]}{Pr[X_{vi} < m \mid \theta]}\right\} = \alpha_i\theta + \beta_{im}.$$

Unlike the partial credit model the graded response model requires that the item-category parameters β_{im} are ordered by the category index m, $\beta_{i1} < \beta_{i2} < \cdots < \beta_{iK_i}$. The R package **ltm** has the capability to estimate the graded response function under various constraints on the item parameters.

Sequential scale models

The sequential scale models assume that the continuation ratio logits are increasing functions of the propensity θ. The parametric sequential scale model (SSM) assumes that the continuation ratio logits are linear in the propensity score:

$$\log\left\{\frac{Pr[X_{vi} \geq m \mid \theta, X_{vi} \geq m-1]}{Pr[X_{vi} = m-1 \mid \theta, X_{vi} \geq m-1]}\right\} = \alpha_i\theta + \beta_{im},$$

Assuming that the continuation logits are linear results in the following item-category response functions:

$$P_{im}(\theta_v) = \frac{\exp\{\sum_{\ell=1}^{m} \alpha_i\theta_v + \beta_{i\ell}\}}{\prod_{\ell=1}^{m}(1 + \exp\{\alpha_i\theta_v + \beta_{i\ell}\})}$$

Tutz (1990) introduced the sequential or step-wise Rasch model, which is a sequential scale model that assumes that the slopes are constant across item, i.e., $\alpha_1 = \alpha_2 = \cdots = \alpha_J$.

Relationships between the models

The three classes of parametric IRT models above, namely the GPCM, GRM, and SSM models are disjoint classes of models. Therefore, the parameters derived from one of the models cannot be mapped to meaningful parameters in one of the other classes of models. For a good review of the relationships among these three classes of models see van der Ark (2001) and the references therein.

2.2. Models for dichotomous data

The two-parameter logistic model

When items are dichotomous ($K_j = 1$), the partial credit model, graded response model, and rating scale model all reduce to the two-parameter logistic model (2PL; Birnbaum 1968). Specifically the 2PL models the item response function of a two-parameter logistic model $P_j(\theta) \equiv P_{j1}(\theta)$ as

$$\begin{aligned} \text{logit}\{P_j(\theta)\} &= \alpha_j\theta + \beta_j \\ P_j(\theta) &= \frac{1}{1 + \exp\{-\alpha_j\theta - \beta_j\}} \end{aligned} \tag{3}$$

The slope parameter, sometimes called the discrimination of the item, is a measure of how much information an item provides about the latent variable θ. As $\alpha \to \infty$ the item response function approaches a step function with a jump at β_j; such item response functions are sometimes referred to as Guttman items (Guttman 1950).

The Rasch model

The one-parameter logistic model, or Rasch model (Rasch 1960) assumes that the item slopes are constant across items, i.e., $\alpha_1 = \alpha_2 = \cdots = \alpha_J = \alpha$, and therefore is the dichotomous version of Masters' partial credit model.

The slope parameter α in the Rasch model can be fixed to some arbitrary value without affecting the likelihood as long as the scale of the individuals' propensities is allowed to be free. Common values for the discrimination are $\alpha = 1$ and $\alpha = 1.7$, which is used so that the item response function is similar to the normal CDF (the standard deviation of the logistic distribution is $\frac{\pi}{\sqrt{3}} \approx 1.8$ and a MacLauren expansion yields the approximation $\text{logit}\{\Phi(x)\} \approx 1.6x$) .

Another attractive property of the Rasch model is that the raw score $X_{v+} = \sum_{i=1}^{J} X_{vi}$ is a minimal sufficient statistic for the individual propensity parameter θ_v. In fact, the Rasch model is the only dichotomous item response model for which there exists a one-dimensional minimal sufficient statistic for the propensity parameter Andersen (1977).

Three parameter logistic model

The response functions $P_i(\theta) \to 1$ as $\theta \to \infty$ and $P_i(\theta) \to 0$ as $\theta \to -\infty$ for both the Rasch and 2PL models. However, for multiple choice test items, cognitive theory suggests that when an examinee does not know the correct response, the individual will guess. In situations where guessing is possible, the assumption $\lim_{\theta \to -\infty} P_i(\theta) = 0$ is not a reasonable assumption of the cognitive process the model is attempting to measure. For this reason Birnbaum (1968) developed a generalization of the 2PL that allows the IRF $P_i(\theta)$ to have a lower asymptote different from zero. The generalization is

$$P_i(\theta) = \gamma_i + \frac{1 - \gamma_i}{1 + \exp\{\alpha_i(\beta_i - \theta)\}} \tag{4}$$

The 3PL assumes that the examinee knows the correct answer of the item with probability equal to (3) or guesses the item correctly with probability γ_i.

The 3PL model may be useful in applications other than educational testing. In many attitudinal surveys, there are items for which it makes sense to assume that all individuals have a probability that is bounded below by some non-zero number γ, regardless of the individual's propensity.

2.3. Non-parametric IRT models

Many researchers have suggested using the total score x_{v+} as the independent variables in a non-parametric logistic regression as a way to examine the shape of the unknown response function $P_i(\theta)$. Ramsay (1991), for example, uses Kernel regression as a way to estimate $P_i(\theta)$. Although Douglas (1997) shows that this method consistently estimates both the shape of the item response function and the rank order of examinees, the method does not work well for small data sets.

Ramsay and Abrahamowicz (1989) and Winsberg, Thissen, and Wainer (1984) on the other hand suggest methods for the estimation of the non-parametric response function $P_{im}(\theta)$, which utilizes B-splines to model model the adjacent-category logit ψ_{im}. Although the B-spline item response model is likely too complicated to use operationally, it can be utilized to examine the appropriateness simpler item response models.

3. Estimation

Estimating the model parameters for any item response model requires additional thought about the items and the respondents participating in the survey. Basic estimation techniques for item response models assume that the individuals participating in the survey are independent of one another and that items behave in the same way for all individuals (i.e. there is no differential item functioning present).

There are four basic techniques for the estimation of item response models: joint maximum likelihood, conditional maximum likelihood, marginal maximum likelihood, and Bayesian estimation with Markov chain Monte Carlo. All four basic estimation methods rely heavily on the assumption that individuals are independent of one another, and that the item responses of a given individual are independent given that individual's propensity score θ_v. Under the assumption of conditional independence the joint probability of the item response vector $\boldsymbol{x}_v$ conditional on θ_v is

$$L_i(\theta \mid \boldsymbol{x}_v, \boldsymbol{\phi}) = Pr\{\boldsymbol{x}_v \mid \theta_v, \boldsymbol{\phi}\} = \prod_{i=1}^{J} Pr\{X_{vi} = x_{vi} \mid \theta_v, \boldsymbol{\phi}_i\}, \tag{5}$$

where $\boldsymbol{\phi}_i$ is the vector of all item parameters for item i. For example, the likelihood for propensity θ under the 2PL model, where $\boldsymbol{\phi_i} = (\alpha_i, \beta_i)^\top$, is:

$$L_v(\theta_v \mid \boldsymbol{x}_v, \boldsymbol{\phi}) = \frac{\exp\{\theta_v \sum_i x_{vi}\alpha_i - \sum_i x_{vi}\alpha_i\beta_i\}}{\prod_i [1 + \exp\{\alpha_i(\theta_v - \beta_i)\}]}$$

The following sections describe the four basic methods for the estimation of item response models.

3.1. Joint maximum likelihood

The joint maximum likelihood (JML) estimation procedure treats both item parameters (e.g. β_i) and propensities θ_v as unknown, but fixed model parameters. Under the JML procedure the $N \times J$ item responses are essentially treated as the observational units in the analysis. The JML procedure estimates the item parameters ($\boldsymbol{\phi}$) and examinee abilities by maximizing $L(\boldsymbol{\phi}, \boldsymbol{\theta}; \boldsymbol{X}) = \prod_v L_v(\theta_v \mid \boldsymbol{x}_v, \boldsymbol{\phi})$ with respect to $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ simultaneously.

The model is not identified, which means there is no unique solution to the maximization. A unique solution does exist if further constraints are placed on the parameters of the model. For two parameter models like the 2PL, two constraints are necessary: a location constraint, and a scale constraint. The location constraint can be made by constraining either a single propensity or difficulty to some fixed number, or by constraining the average propensity or difficulty to some fixed number (typically zero). The scale constraint can be made by forcing the product of the discrimination parameters to one (i.e. $\prod_i \alpha_i \equiv 1$).

One of the problems with JML estimates in models similar to IRT models is that the estimates are inconsistent (Neyman and Scott 1948; Andersen 1970; Ghosh 1995). In terms of IRT models, this means that no matter how many individuals are included in the sample, the estimates for the item parameters may still be biased.

3.2. Conditional maximum likelihood

Andersen (1970) suggests an alternative method for maximum likelihood estimation of the Rasch model. His method conditions on the vector of raw scores $X_{v+} = \sum_i X_{vi}$, which is a sufficient statistic for the propensities of individuals in the sample:

$$Pr\{\boldsymbol{x}_v \mid \theta_v, \boldsymbol{\phi}, X_{v+} = r_v\} = \frac{\exp\{-\sum_i x_{vi}\beta_i\}}{\sum_{\{\boldsymbol{y}:\sum_i y_i = r_v\}} \exp\{-\sum_i y_i\beta_i\}},$$

where $\boldsymbol{r} = (r_1, \ldots, r_N)^\top$ denotes the vector of observed raw scores. The quantity above does not depend on the value of the individual's propensity θ. The conditional maximum likelihood estimates the item parameters by maximizing the *conditional* likelihood $L(\boldsymbol{\phi} \mid \boldsymbol{X}, \boldsymbol{r}) = \prod_i Pr\{\boldsymbol{x}_i \mid \boldsymbol{\psi}, s_i\}$. The paper by Mair and Hatzinger (2007) in this volume discusses the use of R for conditional maximum likelihood estimation in IRT.

Although Andersen (1970) shows that conditional maximum likelihood estimates for the item difficulties are consistent, an *ad hoc* procedure must be implemented to estimate the propensities of individuals. In addition, the conditional maximum likelihood method only works when there is a simple sufficient statistic like the raw score for the Rasch model. However, as noted earlier more complex IRT models, including the 2PL, do not have simple sufficient statistics.

3.3. Marginal maximum likelihood

Marginal maximum likelihood (MML) takes a different approach to removing the propensities from the likelihood. Unlike joint maximum likelihood estimation techniques, which treat each of the $N \times J$ item responses as separate observational units, the marginal technique treats only the N individuals as the observational units. To accomplish this the MML technique assumes that the propensities are random effects sampled from some larger distribution, denoted $F(\theta)$. The distribution may or may not have support on the whole real line. When the distribution $F(\cdot)$ is discrete, we typically call the resulting model a *ordered latent class model*. *Latent variable models* usually refer to models where $F(\cdot)$ is continuous.

Integrating the random effects (i.e. propensities) out of the individual likelihoods defined in (5) defines the marginal probability of observing the item response vector $\boldsymbol{x}_i$,

$$Pr\{\boldsymbol{x}_v \mid \boldsymbol{\phi}\} = \int_\Theta L_i(\theta \mid \boldsymbol{x}_v, \boldsymbol{\phi}) dF(\theta). \tag{6}$$

Taking the product of the probabilities in (6) over individuals v defines the marginal likelihood of the item parameter vector $\boldsymbol{\phi}$

$$L(\boldsymbol{\phi} \mid \boldsymbol{X}) = \prod_v Pr\{\boldsymbol{x}_v \mid \boldsymbol{\phi}\},$$

which is is maximized with respect to the item parameters $\boldsymbol{\phi}$ to derive the MML estimates. Like the JML estimation method, location and scale constraints are required to identify the

model. The constraints can either be placed on the mean and standard deviation of the propensity distribution F or on the item parameters.

The propensity distribution F is now a part of the IRT model and care must be taken when choosing the parametric form of F. Typically IRT modelers assume that the distribution F is the normal distribution with mean zero and standard deviation one. However, the normal distribution does not necessarily work for all applications.

Another possible way to get around the difficulty of defining the distribution F is to assume some non- or semi-parametric form. For example, the analysis of the National Assessment of Educational Progress, a large scale educational survey, assumes examinee propensities θ_v are independently and identically distributed according to a discrete distribution on 41 equally spaced points from -4 to 4 with unknown mass. That is, the probability mass function for the propensity θ is

$$f_\Theta(t) = \begin{cases} p_t & \text{if } t \in \{-4, -3.8, \ldots, 4\} \\ 0 & \text{otherwise} \end{cases}$$

where $\sum_t p_t = 1$. For this distribution of propensities the marginal probability in (6) becomes

$$Pr\{\boldsymbol{x}_v \mid \boldsymbol{\phi}\} = \sum_{t \in \{-4,\ldots,4\}} Pr\{\boldsymbol{x}_v \mid t, \boldsymbol{\phi}\} p_t.$$

The masses p_t are estimated simultaneously with the item parameters $\boldsymbol{\phi}$. Mislevy and Bock (1982) and Muraki and Bock (1997) provide more information on this estimation technique.

In addition to requiring numerical methods to accomplish the maximization of the likelihood, the MML technique also requires numerical integration techniques to approximate the integral in (6).

3.4. Bayesian estimation with Markov chain Monte Carlo

The Bayesian method for estimation of IRT models is similar to the marginal likelihood technique described in the previous section. However, in addition to assuming a mixing distribution for the propensities, Bayesian analysis places a prior distribution on each of the model parameters. It is also possible to simultaneously estimate posterior quantities for both the items and the respondents in the data set.

One of the shortcomings of a Bayesian analysis of an IRT model is that numerical integration techniques must be used to approximate the posterior distributions (Patz and Junker 1999). The numerical method, called Markov chain Monte Carlo (MCMC), can be quite time consuming for large data sets, and requires extreme care to make sure that the resulting approximations to the posterior distribution are valid.

4. Marginal estimation of item response models

The R package **ltm** (Rizopoulos 2006) contains a large number of functions for the analysis of item response data, including functions for estimation and inference for the Rasch, 2PL, and GRM models described above. The **gpcm** developed here is a relatively basic package designed to produce parameter estimates and standard errors for the generalized partial credit model.

Both packages rely on the expectation-maximization (EM; Dempster, Laird, and Rubin 1977) algorithm for marginal maximum likelihood estimation of their models. Below I review the EM algorithm as it relates to the generalized partial credit model and give some details about the **gpcm** package.

4.1. An EM algorithm for the GPCM

The EM algorithm is a numerical method for the maximization of a likelihood that depends on missing, or latent data. For latent variable models such as the GPCM and IRT models in general, we can think of the latent propensity θ_v, for each individual v, as the missing data. Combining the vector of all missing data $\boldsymbol{\theta}$ with the observed data matrix $\boldsymbol{X}$ produces the "complete" data $\boldsymbol{Z} = (\boldsymbol{X}, \boldsymbol{\theta})$. In the E-step the expected value of the log-likelihood of the parameters given the complete data is calculated, with the expectation being taken over the conditional distribution of the missing data θ, given the observed data $\boldsymbol{X}$. The M-step maximizes the objective function resulting from the E-step.

Below I summarize the the EM algorithm for the generalized partial credit model. See Muraki (1992) for more detail on implementing the EM algorithm for the GPCM.

For the generalized partial credit model, the complete data log-likelihood is

$$\begin{aligned} \ell_c(\boldsymbol{\phi}|\boldsymbol{X}, \boldsymbol{\theta}) &= \sum_{v=1}^{N} \left\{ f(\theta_i) + \sum_{i=1}^{J} \left\{ \sum_{m=1}^{x_{vi}} \psi_{im}(\theta_v) + \ln P_{i0}(\theta_v) \right\} \right\} \\ &= \sum_{v=1}^{N} \left\{ f(\theta_v) + \sum_{i=1}^{J} \left\{ \boldsymbol{y}_{vi}^{\top} \boldsymbol{\psi}_i(\theta_v) + \ln P_{i0}(\theta_v) \right\} \right\}, \end{aligned} \tag{7}$$

where $f(\theta_v)$ is the density of the prior (mixing) distribution assumed for the latent propensities (we assume that all parameters of f are known). The vector $\boldsymbol{y}_{vi}$ is a K_i-vector with

$$y_{vim} = \begin{cases} 1 & \text{if } x_{vi} \geq k \\ 0 & \text{if } x_{vi} < m \end{cases}$$

and

$$\begin{aligned} \boldsymbol{\psi}_i(\theta) &= \left[\begin{array}{ccc} \theta \mathbf{1}_{K_i} & \mathbf{1}_{K_i} & \boldsymbol{C}_{K_i} \end{array} \right] \left(\begin{array}{c} \alpha_i \\ \beta_i \\ \boldsymbol{\delta}_i \end{array} \right) \\ &= \boldsymbol{D}(\theta) \boldsymbol{\phi}_i; \end{aligned}$$

$\boldsymbol{\phi}_i$ denotes the vector of item parameters for item i, $\boldsymbol{\phi}$, without an index, denotes the vector of all model parameters, and $\boldsymbol{\psi}_i(\theta)$ denotes the vector of adjacent-category logit functions of θ. The matrix $\boldsymbol{C}_{K_i}$ is a $K_i \times (K_i - 1)$ matrix of contrasts of the form

$$C_{K_i} = \left[\begin{array}{c} \boldsymbol{I}_{K_i - 1} \\ -\mathbf{1}_{K_i - 1}^{\top} \end{array} \right],$$

which ensures identifiability of the model.

The E-step

Let $\boldsymbol{\phi}^{(s)}$ denote the approximation of the maximum likelihood estimates of the model parameters at step s of the EM algorithm. To update the approximation at step $s+1$, the E-step defines the objective function $Q(\boldsymbol{\phi}|\boldsymbol{\phi}^{(s)})$ by calculating the expected value of the complete log-likelihood in (7) with respect to the conditional (posterior) distribution of the vector of the propensities $\boldsymbol{\theta}$ given the observed data $\boldsymbol{X}$. Under the assumption that the individuals in the study are independent the expectation simplifies to

$$Q(\boldsymbol{\phi}|\boldsymbol{\phi}^{(s)}) = \sum_{v=1}^{N} \int \left\{ f(t) + \sum_{i=1}^{J} \left\{ \boldsymbol{y}_{vi}^{\top} \boldsymbol{D}(t) \boldsymbol{\phi}_i + \ln P_{i0}(t) \right\} \right\} dF_{\theta|\boldsymbol{x}}(t|\boldsymbol{x}_v, \boldsymbol{\phi} = \boldsymbol{\phi}^{(s)}),$$

where the conditional posterior distribution of θ given the response vector $\boldsymbol{x}_v$, $F_{\theta|\boldsymbol{x}}(t|\boldsymbol{x}_v, \boldsymbol{\phi} = \boldsymbol{\phi}^{(s)})$, is calculated assuming the current estimate of the item response parameters, $\boldsymbol{\phi}^{(s)}$.

The integration required is analytically intractable and therefore a numerical method to approximate it is required. R provides a number of resources for approximating integrals. The **gpcm** package described below uses Gauss-Hermite quadrature to approximate the integrals. The package **gpcm** utilizes the function `gauss.quad.prob` available from the **stamod** package.

The M-step

The maximization step of the EM algorithm maximizes the objective function $Q(\boldsymbol{\phi}|\boldsymbol{\phi}^{(s)})$ with respect to the parameter vector $\boldsymbol{\phi}$. The objective function is maximized by solving the system of equations

$$\nabla_{\boldsymbol{\phi}} Q(\boldsymbol{\phi}|\boldsymbol{\phi}^{(s)}) = \mathbf{0}.$$

The gradient function is

$$\nabla_{\boldsymbol{\phi}} Q = \begin{pmatrix} \nabla_{\boldsymbol{\phi}_1} Q \\ \vdots \\ \nabla_{\boldsymbol{\phi}_J} Q \end{pmatrix},$$

where $E_{\theta|\boldsymbol{x}}$ denotes the expected value operator over the conditional distribution of θ given the response vector $\boldsymbol{X}$,

$$\nabla_{\boldsymbol{\phi}_i} Q = \sum_{v=1}^{N} E_{\theta|\boldsymbol{x}} \left[D^{\top}(\theta) \left(\boldsymbol{y}_{vi} - \boldsymbol{W}_{K_i} \boldsymbol{P}_i(\theta) \right) \middle| \boldsymbol{x}_v, \boldsymbol{\phi}^{(s)} \right],$$

and $\boldsymbol{W}_{K_i}$ is the $K_i \times K_i$ matrix with ones on the diagonal and the upper triangle, and zeroes in the lower triangle:

$$\boldsymbol{W}_{K_i} = \begin{pmatrix} 1 & 1 & 1 & \cdots & 1 \\ 0 & 1 & 1 & \cdots & 1 \\ 0 & 0 & \ddots & & 1 \\ & & & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & 1 \end{pmatrix}$$

It is analytically intractable to solve the system of equations required to optimize the objective function Q. We therefore are required to solve the system numerically. The **gpcm** and **ltm** packages use the Newton-Raphson algorithm to approximate the maximizing values of the

parameter vectors. The method requires the Hessian matrix of second derivatives of the objective function Q. For the generalized partial credit model, the Hessian can be calculated separately for each item j, because of the assumption of conditional independence of the items given the latent propensity θ. The Hessian matrix corresponding to item j is the $(K_j + 1) \times (K_j + 1)$ matrix

$$\nabla^2_{\boldsymbol{\psi}_j} Q = -\sum_{v=1}^{N} E_{\theta|\boldsymbol{x}} \left[\boldsymbol{D}^\top(\theta) \boldsymbol{W}_{K_i} \left(\text{diag}(\boldsymbol{P}_i) - \boldsymbol{P}_i(\theta) \boldsymbol{P}_i^\top(\theta) \right) \boldsymbol{W}_{K_i}^\top \boldsymbol{D}(\theta) \right]$$

An updated parameter vector for item i is obtained by setting

$$\boldsymbol{\phi}_i^{(s+1)} = \boldsymbol{\phi}_i^{(s)} - \left[\nabla^2_{\boldsymbol{\phi}_i} Q \right]^{-1} \left[\nabla_{\boldsymbol{\phi}_i} Q \right]$$

The algorithm could iterate this step until the objective function Q is maximized, and then recalculate a new Q function via the E-step. The approach taken in the **gpcm** package takes only a single step before updating the Q function. Because of the high computational cost of recalculating the gradient vector and Hessian matrix, this approach was found to be more computationally efficient.

Approximating standard errors

The variances of the maximum likelihood estimates are approximated by inverting the observed Fisher information matrix. The observed Fisher information matrix for the item parameters $\boldsymbol{\phi}_i$ of item i can be written

$$\mathcal{I}(\hat{\boldsymbol{\phi}}_i) = -\nabla^2_{\boldsymbol{\phi}_i} Q - E_{\boldsymbol{\theta}|\boldsymbol{x}} \left[(\nabla_{\boldsymbol{\phi}_i} \ell_c)(\nabla_{\boldsymbol{\phi}_i} \ell_c)^\top \right]$$

The vector $\nabla_{\boldsymbol{\phi}_i} \ell_c$ is function of the vector $\boldsymbol{\theta}$ all N latent propensities and the notation $E_{\boldsymbol{\theta}|\boldsymbol{x}}$ illuminates the fact that expectation is taken with respect to the entire vector of propensities.

The observed Fisher information for the entire set of item parameters $\boldsymbol{\phi}$, across all items, is not block-diagonal. However, the implementation in **gpcm** does not calculate the cross-item information, and therefore the complete covariance matrix cannot be obtained. The standard errors reported are derived by inverting the item-specific information matrix defined above.

4.2. The gpcm package

The **gpcm** package is an R package for the estimation of the generalized partial credit model. The package contains one data set, named `foodsec`, described in greater detail below, and eight functions:

- `gpcm`: The main function of the package is the `gpcm` function, which estimates the parameters of the GPCM and approximates the standard errors of those estimates assuming one of the distributions handled by the **stamod** function `gauss.quad.prob` (normal, uniform, beta, and gamma). The algorithm used is an iterative one, where, at first, parameter estimates are updated via the EM algorithm described above. Once the difference in parameter estimates reaches a pre-determined tolerance level (set by the option `tol1`), the algorithm utilizes the negative of the Fisher information to obtain updated parameter values, i.e.,

$$\boldsymbol{\phi}^{(s+1)} = \boldsymbol{\phi}^{(s)} + [\mathcal{I}(\boldsymbol{\phi})]^{-1} [\nabla Q].$$

The algorithm stops after the difference between iterations is below a pre-determined tolerance level (`tol2`).

At the time of this article `gpcm` could not handle missing values. However, work is underway to handle data that is missing at random.

- `update.gpcm.em` and `update.gpcm.nr` are the functions that update the parameter estimates. The '.em' function uses the EM algorithm to update parameter values, the '.nr' function replaces the Hessian in the EM algorithm with the negative of the approximate Fisher information.
- `gpcm.p` calculates the the category response functions for the generalized partial credit model. The function handles a vector of propensities (θ) and can handle several items.
- `plot` method for `gpcm` objects: It plots the estimated category response functions or item response functions based on the estimated item parameters from the gpcm function.
- `print` method for `gpcm` objects: It simply prints the item parameter estimates obtained from `gpcm`.
- `rgpcm` simulates individuals' propensities from a normal distribution and then simulates their responses to items defined by the item parameters input by the user.
- `logLik` method for `gpcm` objects: It extracts the value of the log-likelihood from a `gpcm` object.

5. Examples

5.1. Analysis of simulated data

We begin by simulating the responses of $N = 1000$ individuals to $J = 10$ items, where item parameters have been simulated from appropriate normal distributions; the vector `a` is the vector of item slopes, `b` is the vector of item intercepts, and `d` is a matrix of item-step parameters.

```
R> library("gpcm")
Loading required package: statmod
R> a <- rnorm(10, 1.5, 0.3)
R> b <- rnorm(10, 0, 1)
R> d <- matrix(rnorm(40), 10, 4)
```

The matrix of item-step parameters is 10×4. However, if not all items have five score categories, we must first set the appropriate columns of the matrix to `NA`. An `NA` in the matrix indicate that the category associated with that column is not possible for the item designated by that row of the matrix. In our simulated data, the first two items are dichotomous, so the the last three columns of the first two rows of `d` must be set to `NA`.

```
R> is.na(d[1:2,2:4]) <- TRUE
```

The third and fourth items are trichotomous.

```
R> is.na(d[3:4,3:4]) <- TRUE
```

The fifth and sixth have four categories.

```
R> is.na(d[5:6,4]) <- TRUE
```

The remaining items have five categories. Finally, we force the item-step parameters to sum to zero.

```
R> d <- d - apply(d, 1, mean, na.rm = TRUE)
```

To investigate the shape of the category response functions we create a list containing a `data.frame` named `est` that holds the parameter values and then invoke `plot.gpcm` by issuing the command `plot`.

```
R> parms <- list(est = data.frame(cbind(a,b,d)))
R> class(parms) <- "gpcm"
R> plot(parms)
R> plot(parms,plot.type = "irf")
```

The first `plot.gpcm` plots the category response functions. The second invocation plots the expected score function or item response function,

$$E[X_j|\theta] = \sum_{k=1}^{K_j} kP_{jk}(\theta)$$

For sake of brevity we do not include the graphics here.

Use `print` (`print.gpcm`) to print out the item parameters

```
R> parms
       a      b      V3     V4     V5    V6
1  1.620 -0.135  0.0000     NA     NA    NA
2  1.703 -0.294  0.0000     NA     NA    NA
3  1.010  1.881  0.4244 -0.424     NA    NA
4  1.246  0.871  0.6313 -0.631     NA    NA
5  1.595  1.399  0.0410 -0.155  0.114    NA
6  1.378  0.690 -0.5429  1.536 -0.993    NA
7  1.670 -0.381 -0.7683 -1.211  1.082 0.897
8  0.929 -0.524  0.7124 -1.322  0.301 0.309
9  1.372  1.002  1.0860 -1.229 -0.328 0.470
10 1.092 -0.158 -0.1687  0.653 -1.315 0.830
```

We simulate the data by invoking the `rgpcm` function.

```
R> Y <- rgpcm(1000, a, b, d)
```

To estimate the model parameters of the GPCM for the simulated data set, we use the `gpcm` function

```
R> (Y.gpcm <- gpcm(Y))
        Slope Intercept   Cat.1   Cat.2    Cat.3 Cat.4
Item 1   1.66    -0.122  0.0000      NA       NA    NA
Item 2   1.73    -0.162  0.0000      NA       NA    NA
Item 3   1.08     1.884  0.4266 -0.4266       NA    NA
Item 4   1.26     0.842  0.6280 -0.6280       NA    NA
Item 5   1.76     1.434 -0.0581  0.0563  0.00183    NA
Item 6   1.40     0.719 -0.6667  1.6377 -0.97098    NA
Item 7   1.62    -0.363 -0.9969 -0.7613  0.92697 0.831
Item 8   1.00    -0.554  0.7222 -1.1128  0.10692 0.284
Item 9   1.41     0.970  0.9942 -1.2378 -0.35188 0.595
Item 10  1.13    -0.145 -0.2950  0.6941 -1.28271 0.884
```

The standard errors of the estimates are contained in the matrix `Y.gpcm$se`.

The default usage of `gpcm` uses 41 quadrature points to approximate the necessary integrals. Experience suggests that using too few quadrature points leads to slope estimates that are biased; specifically, using too few quadrature points leads to slope estimates that are too low on average.

We can estimate the GPCM assuming a number of different mixing distribution. The **statmod** function `gauss.quad.prob` allows for normal, uniform, beta, and gamma mixing distributions. The code below compares the fits for various choices of the mixing distribution

```
R> logLik(Y.gpcm)
'log Lik.' -8668.953 (df=38)
R> logLik(gpcm(Y, prior.dist = "uniform"))
'log Lik.' -8691.243 (df=38)
R> logLik(gpcm(Y, prior.dist = "beta", a = 2, b = 2))
'log Lik.' -8675.137 (df=38)
R> logLik(gpcm(Y, prior.dist = "beta", a = 20, b = 20))
'log Lik.' -8668.863 (df=38)
R> logLik(gpcm(Y, prior.dist = "gamma", a = 1, b = 1))
'log Lik.' -8777.399 (df=38)
```

The uniform mixing distribution clearly does not fit this data well, which is not surprising considering the true propensities were generated from a normal distribution. The Beta(2,2) and Gamma(1,1) mixing distributions also fit significantly worse. The Beta(20,20) distribution fits slightly better than the normal distribution. This is not surprising considering how closely a normal distribution can approximate a Beta(20,20) distribution.

Finally we examine the expected *a posteriori* estimates of the propensity scores. These are defined as the posterior mean of the latent propensity given the individuals response vector $\boldsymbol{x}_i$ and the maximum likelihood estimates of the item parameters $\boldsymbol{\psi}$,

$$EAP_i = E_{\theta|\boldsymbol{x}}[\theta|\boldsymbol{x}_i, \hat{\boldsymbol{\psi}}],$$

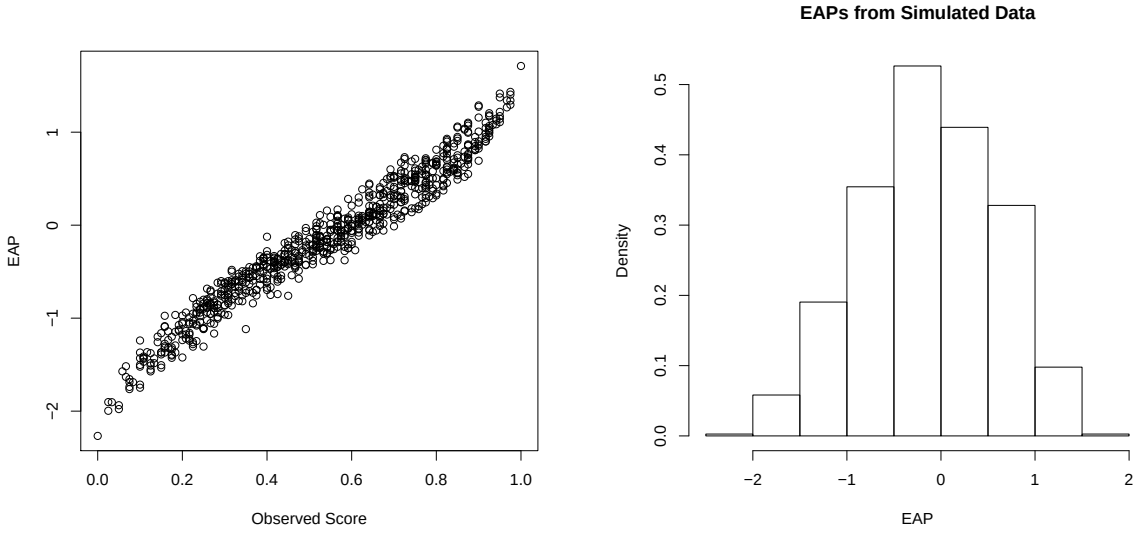

Figure 1: Plots examining the EAPs from the simulated data set. The figure in the left panel is a plot of the EAPs by the observed scores. The right panel contains a histogram of the EAPs.

which is approximated with Gauss-Hermite quadrature. The EAP estimates for each unique response vector are contained in the vector `Y.gpcm$eap`.

To compare the EAPs to the observed scores we begin by calculating the observed scores and then plotting the EAPs against the observed score.

```
R> max.score <- apply(Y.gpcm$data.unique, 2, max, na.rm = TRUE)
R> prop.score <- t(Y.gpcm$data.unique)/max.score
R> obs.score <- apply(prop.score, 2, mean)
R> plot(obs.score, Y.gpcm$eap, xlab = "Observed Score", ylab = "EAP")
```

The resulting scatterplot is contained in the left panel of Figure 1. The right panel of Figure 1 contains a histogram representing the distribution of EAPs in the data set.

```
R> hist(Y.gpcm$eap, freq = FALSE, main = "EAPs from Simulated Data",
+    xlab = "EAP")
```

5.2. The USDA's Food Insecurity scale

The United States Department of Agriculture's (USDA) US Food Security Measurement Project administers a battery of survey questions as a supplement to the Current Population Survey (CPS) in an attempt to estimate the proportion of the population that is (in)secure in their ability to obtain "enough food for an active, healthy life." Responding households without children are classified into one of three food security levels (food secure, food insecure without hunger, food insecure with hunger) based on their responses to the ten items listed in Table 1.

1. We worried whether our food would run out.	6. Adult ate less than they felt they should.
2. Food bought didn't last.	7. Adult hungry but couldn't afford to eat.
3. Adult unable to eat balanced meals.	8. Adult lost weight because couldn't afford to eat enough.
4. Adult cut size or skipped meals.	9. Adult didn't eat for an entire day because couldn't afford to eat.
5. ... this happened in three or more months over the past year.	10. ... this happened in three or more months.

Table 1: The ten food security items. Respondents were asked if statements 1-3 were "often," "sometimes," or "never" true over the last twelve months. For the remaining items, except 5 and 10, respondents responded by answering "yes" or "no". Follow-up items 5 and 10 asked respondents whether the preceding statement was true in "only 1 or 2 months," "some months but not every month," or " almost every month."

The USDA utilizes the Rasch model to measure the unidimensional latent construct "food insecurity" by dichotomizing all of the items listed in Table 1. Responses of "often" and "sometimes" to items 1-3 are treated as successes ($x_{ij} = 1$), and "never" is treated as a failure ($x_{ij} = 0$). Follow up questions 5 and 10 are dichotomized by treating responses of "only in 1 or 2 months" as failures. The Food Security Project treats all missing responses as failures ($x_{ij} = 0$) and we will do the same here.

A number of studies have suggested that the unidimensional Rasch model may not be adequate for the ten items listed above. Johnson (2004) fit a 2PL to Food Security data from the 2002 CPS. That paper found that differences in the discrimination parameters were statistically significant. Specifically the paper found that the discrimination parameter for the item that asks if an "Adult was hungry but couldn't afford to eat" was significantly larger than the discrimination parameters for all other items. Johnson (2006) compared the fit of the Rasch model to a free-knot B-spline response model and found that the Rasch model fit significantly worse.

Below we use the data from Johnson (2004) to perform a number of analyses on the 2002 Food Security data with the **ltm** and **gpcm** packages. The data is included in the **gpcm** package and can be accessed by issuing the command `data(foodsec)`. The data set contains the responses of 9804 respondents on the ten questions described in Table 1.

Dichotomous analysis

We begin by performing an analysis of the food security data that dichotomizes the data. The data in `foodsec` is the polytomous data, and, therefore, must be transformed to the dichotomous data utilized by the USDA.

```
R> data("foodsec")
R> food <- foodsec
R> food[,c(5,10)] <- ifelse(food[,c(5,10)]==0,0,1)
R> food <- t(ifelse(t(food)==apply(food,2,max,na.rm=TRUE),1,0))
```

Although **ltm** claims to be able to handle missing data that is missing at random, all three functions: `grm`, `rasch`, and `ltm` were unable to handle the food security data with missing values. So moving forward we will examine the food security data that treats missing data as zeroes ("failures"). This is exactly what the USDA does in analysis.

```
R> food[is.na(food)] <- 0
```

Using the R packages **gpcm** and **ltm** we test the significance of the differences in the item slopes by first fitting the 2PL with the **ltm** functions `ltm`, `grm`, and **gpcm** function `gpcm` and comparing to the fit of the Rasch model, fit using the **ltm** function `rasch`.

To fit the 2PL to the food data we use the `ltm` function as follows:

```
R> food.2pl1 <- ltm(food ~ z1)
```

where `z1` represents the propensity θ. To fit the 2PL using the `grm` function, we must add one to the all responses, because the function requires the smallest category in the data to be one rather than zero.

```
R> food.2pl2 <- grm(food + 1)
```

The `rasch` and `gpcm` functions take the raw matrix of zeroes and ones to fith the Rasch and 2PL models respectfully.

```
R> food.rasch <- rasch(food)
R> food.2pl3 <- gpcm(food)
Error in drop(.Call("La_dgesv", a, as.matrix(b), tol, PACKAGE = "base")) :
system is computationally singular: reciprocal condition number = 1.32921e-16
```

The first thing we notice is that the function `gpcm` produces an error because the function has encountered a nearly singular matrix. So lets compare the results of the other three models

```
R> rbind(logLik(food.2pl1), logLik(food.2pl2), logLik(food.rasch))
          [,1]
[1,] -45604.56
[2,] -45512.70
[3,] -46156.29
```

The first peculiar result is that the two "2PLs" produce different log-likelihoods at the maximum likelihood estimated values. Closer inspection of the parameter estimates reveals that there is something strange going on with the slope parameters of the last two items.

```
R> food.2pl1

Call:
ltm(formula = food ~ z1)

Coefficients:
         Dffclt  Dscrmn
```

```
HESS2     0.784   1.167
HESS3     1.264   1.298
HESS4     1.351   1.212
HESH2    -0.801   1.779
HESHF2    0.043   2.018
HESH3    -1.167   0.820
HESH4     0.715   1.603
HESH5     1.285   1.577
HESSH1    1.232   3.672
HESSHF1   1.382   6.671

Log.Lik: -45604.56

R> food.2pl2

Call:
grm(data = food + 1)

Coefficients:
         Extrmt1  Dscrmn
HESS2      0.781   1.133
HESS3      1.267   1.251
HESS4      1.345   1.184
HESH2     -0.833   1.681
HESHF2     0.037   1.911
HESH3     -1.190   0.812
HESH4      0.702   1.586
HESH5      1.267   1.553
HESSH1     0.936  11.963
HESSHF1    1.252   9.607

Log.Lik: -45512.7
```

The discrimination parameters for the last two items are outliers when compared to the remaining items. Closer inspection reminds us that the last question was a follow up question to the ninth question, and because missing data was converted to 0's, the resulting item responses clearly violate the assumption of conditional independence of the item responses. When small groups of slope parameters have estimated values that are considerably larger than the rest of the items, there is usually a good indication that local dependence may be presence. This is likely what caused `gpcm` to run into problems too. The fifth item is also a follow-up to the fourth question, so we remove the fifth and tenth items and continue our analysis.

```
R> logLik(food.grm <- grm(food[,-c(5,10)]+1))
'log Lik.' -38670.18 (df=16)
R> logLik(food.gpcm <- gpcm(food[,-c(5,10)]))
'log Lik.' -38670.11 (df=16)
```

```
R> logLik(food.gpcm <- gpcm(food[,-c(5,10)], n.quad.pts = 15))
'log Lik.' -38669.93 (df=16)
R> logLik(food.rasch <- rasch(food[,-c(5,10)]))
'log Lik.' -38893.11 (df=9)
```

The default implementation of `gpcm` uses 41 quadrature points and produces a slightly different log-likelihood than the **ltm** function `grm` produced. Running the `gpcm` function with 15 quadrature points (the default number used in `grm`) produces the same value of the log-likelihood. If we use $-2\log(LRT)$ as a test statistic to compare the 2PL fit with the fit of the Rasch model to this data we find

```
R> -2*as.numeric(logLik(food.rasch) - logLik(food.gpcm))
[1] 446.3603
R> pchisq(446.3603, 7)
[1] 1
```

So we have nearly indisputable evidence against the Rasch model in favor of the two-parameter logistic model.

Analysis of the polytomous food security data

It is rather wasteful to collapse the polytomous data into dichotomous data. Here we examine the polytomous food security data and compare the EAPs from the dichotomous analysis to those from the polytomous analysis of the data.

We begin by treating missing values as zeroes (as is done operationally with this data) and fixing the problem with the follow-up questions (questions 5 and 10), by adding the results of those two items to their respective "stem" questions.

```
R> food.poly <- foodsec[,-c(5,10)]
R> food.poly[is.na(food.poly)] <- 0
R> food.poly[,4] <- food.poly[,4] +
+    ifelse(is.na(foodsec[,5]), 0, foodsec[,5])
R> food.poly[,8] <- food.poly[,8] +
+    ifelse(is.na(foodsec[,10]), 0, foodsec[,10])
```

We then compare the fits of the GPCM and GRM for this data

```
R> logLik(food2.gpcm <- gpcm(food.poly))
'log Lik.' -58343.08 (df=23)
R> logLik(food2.grm <- grm(food.poly + 1))
'log Lik.' -58363.02 (df=23)
```

The generalized partial credit model fits this data better than the graded response model. The difference in log-likelihoods is nearly 20, and the models have the same number of parameters.

Some have suggested that the mixing distribution for the food security data should be a right-skewed distribution because of the way that the subjects are sampled for the study. We compare the fits above, which assume a normal mixing distribution to a Gamma(1,1) distribution, which is clearly skewed to the right.

```
R> logLik(food2.gpcm.gam <- gpcm(food.poly, prior.dist = "gamma", a = 1, b = 1))
'log Lik.' -58106.28 (df=23)
```

So, indeed the GPCM with a Gamma(1,1) prior distribution fits the data better than one that assumes a normal prior distribution. Clearly it would be nice to have a method by which we could estimate the prior mixing distribution on the propensities. One approach might allow the weights on the quadrature points to be estimated. Another might model the distribution as a spline function.

Finally we compare the EAP estimates from the 2PL fit of the data to the two GPCM fits of the data. Because the `gpcm` function only gives us EAPs for the unique response patterns, we must first expand those estimates across all 9804 observations so that we can compare across the two treatments of the problem.

```
R> eap.2pl <- as.vector(food.gpcm$eap)
R> names(eap.2pl) <- apply(food.gpcm$data.uniq, 1, paste, collapse = "")
R> eap.2pl <- eap.2pl[apply(food[,-c(5,10)], 1, paste, collapse = "")]

R> eap.gpcm <- as.vector(food2.gpcm$eap)
R> names(eap.gpcm) <- apply(food2.gpcm$data.uniq, 1, paste, collapse = "")
R> eap.gpcm <- eap.gpcm[apply(food.poly, 1, paste, collapse = "")]

R> eap.gpcm.gam <- as.vector(food2.gpcm.gam$eap)
R> names(eap.gpcm.gam) <- apply(food2.gpcm.gam$data.uniq, 1, paste, collapse = "")
R> eap.gpcm.gam <- eap.gpcm.gam[apply(food.poly, 1, paste, collapse = "")]
```

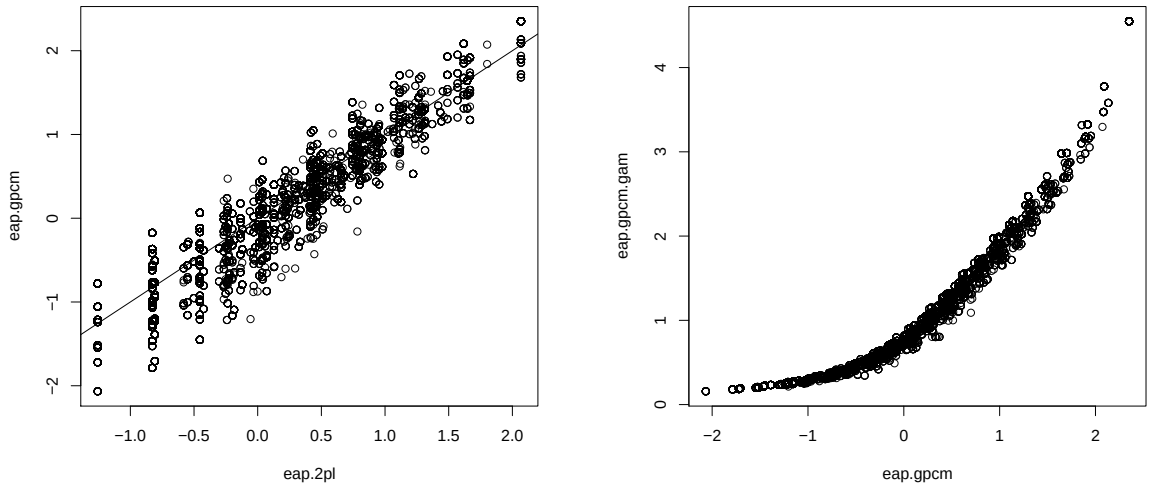

Figure 2: Scatterplots comparing the expected *a posteriori* estimates of the latent propensities from the food security data. The left panel compares the EAPs from the GPCM assuming a normal distribution to the EAPs from a 2PL Analysis. The right panel compares EAPs from two GPCM analyses; one utilizing a normal mixing prior and the other a gamma distribution.

We first compare the the EAPs derived from the 2PL estimation to those from the GPCM estimation assuming a normal mixing distribution.

```
R> plot(eap.2pl, eap.gpcm)
R> abline(a = 0, b = 1)
```

The resulting scatterplot is displayed in the left panel of Figure 2. The relationship is clearly positive and appears to be linear. However, it is clear that the 2PL EAPs tend to be shrunk towards zero more than the GPCM estimates, which results from the fact that the GPCM estimates use more information.

Finally we compare the EAPs from the GPCM assuming a normal mixing distribution to the GPCM assuming a Gamma distribution.

```
R> plot(eap.gpcm, eap.gpcm.gam)
```

The resulting scatterplot appears in the right panel of Figure 2. The relationship between the two sets of EAPs is strong and clearly non-linear, resulting from the difference in assumptions about the shape of the mixing distribution.

6. Discussion

R is an extremely powerful statistical environment. Historically, there was not good, openly-available, R functionality for the analysis of item response data. With the recent advances in the **ltm** package and the **gpcm** package introduced here, it appears that we can expect to see more and more researchers utilizing R for their item response theory analyses.

R has clear advantages over commercial software for IRT analyses, because the user has a single environment in which he/she can complete their entire project. A user of commercial software would have to export their item parameter estimates into another software to produce figures, or non-standard statistics. With **ltm** and/or **gpcm** in R, IRT analysts now have a single environment to complete their entire analysis.

The **ltm** package is quite a bit more sophisticated than the **gpcm** package introduced here. Probably the greatest shortcomings of the **ltm** package is that it only estimates one polytomous item response theory model, the graded response model, and it does not allow users to utilize non-normal mixing distributions. The **gpcm** package is quite limited, in that it only fits the generalized partial credit model.

Moving forward I hope to add functionality to the **gpcm** package to give the user greater control over how the items are constrained. For example, the function should allow users to constrain any linear function of the parameters to some fixed value. Such functionality would allow users to utilize the `gpcm` function to fit the Rasch, and partial credit models by constraining the slopes to be equal.

Extensions to both the **gpcm** and **ltm** packages should allow for the use of more flexible item response functions and prior distributions on the latent propensities. One relatively straightforward extension to the **gpcm** package could allow the adjacent category logits (ψ_{jk}) to be modeled with B-spline functions. The resulting EM algorithm would be only slightly more difficult to implement than the one discussed here.

References

Andersen EB (1970). "Asymptotic Properties of Conditional Maximum Likelihood Estimators." *Journal of the Royal Statistical Society B*, **32**, 283–301.

Andersen EB (1977). "Sufficient Statistics and Latent Trait Models." *Psychometrika*, **42**, 69–81.

Bezruczko N (ed.) (2005). *Rasch Measurement in Health Sciences*. JAM Press, Maple Grove, MN.

Birnbaum A (1968). "Some Latent Trait Models and their Use in Inferring an Examinee's Ability." In F.M. Lord and M.R. Novick (Eds.), *Statistical Theories of Mental Test Scores* (pp. 397–479). Reading, MA: Addison-Wesley.

Dempster A, Laird N, Rubin D (1977). "Maximum Likelihood from Incomplete Data via the EM Algorithm." *Journal of the Royal Statistical Society B*, **39**(1), 1–38.

Douglas J (1997). "Joint Consistency of Nonparametric Item Characteristic Curve and Ability Estimation." *Psychometrika*, **47**, 7–28.

Fienberg SE, Johnson MS, Junker BW (1999). "Classical Multilevel and Bayesian Approaches to Population Size Estimation Using Multiple Lists." *Journal of the Royal Statistical Society A*, **162**(3), 383–405.

Ghosh M (1995). "Inconsistent Maximum Likelihood for the Rasch Model." *Statistics & Probability Letters*, **23**, 165–170.

Grayson DA (1988). "Two Group Classification in Latent Trait Theory: Scores with Monotone Likelihood Ratio." *Psychometrika*, **53**, 383–392.

Guttman L (1950). *Measurement and Prediction, Studies in Social Psychology in World War II*, volume IV, chapter The Basis for Scalogram Analysis, pp. 60–90. University Press, Princeton, NJ.

Hemker B, Sijtsma K, Molenaar I, Junker B (1997). "Stochastic Ordering Using the Latent Trait and the Sum Score in Polytomous IRT Models." *Psychometrika*, **62**, 331–347.

Huynh H (1994). "A New Proof for Monotone Likelihood Ratio for the Sum of Independent Bernoulli Random Variables." *Psychometrika*, **59**, 77–79.

Johnson MS (2004). "Item Response Models and their Use in Measuring Food Insecurity and Hunger." Paper presented at the Workshop on the Measurement of Food Insecurity and Hunger. The National Academy of Science Panel to Review USDA's Measurement of Food Insecurity and Hunger. URL http://www7.nationalacademies.org/cnstat/Item_Response_Models_and_Measuring_Food_Security_Paper.pdf.

Johnson MS (2006). "Modeling Dichotomoous Item Responses with Free-knot Splines." *Computational Statistics & Data Analysis*. In press.

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

Masters GN (1982). "A Rasch Model for Partial Credit Scoring." *Psychometrika*, **47**, 149–174.

Mislevy R, Bock D (1982). **BILOG***: Item Analysis and Test Scoring with Binary Logistic Models*. Scientific Software International, Inc., Lincolnwood, IL. URL http://www.ssicentral.com/.

Muraki E (1992). "A Generalized Partial Credit Model: Application of an EM Algorithm." *Applied Psychological Measurement*, **16**, 159–176.

Muraki E, Bock D (1997). **PARSCALE***: IRT Item Analysis and Test Scoring for Rating Scale Data*. Scientific Software International, Inc., Lincolnwood, IL. URL http://www.ssicentral.com/.

Neyman J, Scott E (1948). "Consistent Estimates Based on Partially Consistent Observations." *Econometrica*, **16**(1), 1–32.

Patz R, Junker BW (1999). "Applications and Extensions of MCMC in IRT: Multiple Item Types, Missing Data, and Rated Responses." *Journal of Educational and Behavioral Statistics*, **24**, 342–366.

Ramsay JO (1991). "Kernel Smoothing Approaches to Nonparametric Item Characteristic Curve Estimation." *Psychometrika*, **56**, 611–630.

Ramsay JO, Abrahamowicz M (1989). "Binomial Regression with Monotone Splines: A Psychometric Application." *Journal of the American Statistical Association*, **84**, 906–915.

Rasch G (1960). *Probablistic Models for Some Intelligence and Attainment Tests*. Nielsen & Lydiche, Copenhagen.

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

Rizopoulos D (2006). "**ltm**: An R Package for Latent Variable Modeling and Item Response Theory Analyses." *Journal of Statitical Software*, **17**(5), 1–25. URL http://www.jstatsoft.org/v17i05/.

Samejima F (1969). "Estimation of Latent Trait Ability Using a Response Pattern of Graded Scores." *Psychometrika Monograph,* No. 17.

Sinharay S, Stern H (2002). "On the Sensitivity of Bayes Factors to the Prior Distribution." *The American Statistician*, **56**, 196–201.

Tutz G (1990). "Sequential Item Response Models with an Ordered Response." *British Journal of Mathematical and Statistical Psychology*, **43**, 39–55.

van der Ark LA (2001). "Relationships and Properties of Polytomous Item Response Models." *Applied Psychological Measurement*, **25**(3), 273–282.

Winsberg S, Thissen D, Wainer H (1984). "Fitting Item Characteristic Curves with Spline Functions." *Technical Report 84-52*, Educational Testing Service, Princeton, NJ.

Affiliation:

Matthew S. Johnson
Department of Statistics & Computer Information Systems
Baruch College, The City University of New York
One Bernard Baruch Way; Box B 11-220
New York, NY 10010, United States of America
E-mail: Matthew_Johnson@baruch.cuny.edu
URL: http://stat.baruch.cuny.edu/~mjohnson

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume 20, Issue 10 *Submitted:* 2006-10-01
May 2007 *Accepted:* 2007-02-22