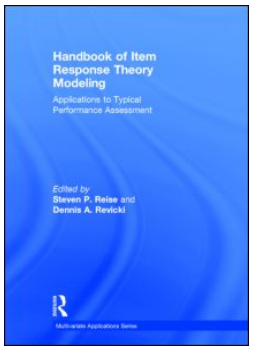
This article was downloaded by: *10.3.98.166*
On: 15 Jun 2018
Access details: *subscription number*
Publisher:*Routledge*
Informa Ltd Registered in England and Wales Registered Number: 1072954 Registered office: 5 Howick Place, London SW1P 1WG, UK



Handbook of Item Response Theory Modeling
Applications to Typical Performance Assessment
Steven P. Reise, Dennis A. Revicki

Modern Approaches to Parameter Estimation in Item Response Theory

Publication details
https://www.routledgehandbooks.com/doi/10.4324/9781315736013.ch3
Li Cai, David Thissen
Published online on: 16 Dec 2014

How to cite :- Li Cai, David Thissen. 16 Dec 2014 *,Modern Approaches to Parameter Estimation in Item Response Theory from:* Handbook of Item Response Theory Modeling,Applications to Typical Performance Assessment Routledge.
Accessed on: 15 Jun 2018
https://www.routledgehandbooks.com/doi/10.4324/9781315736013.ch3

PLEASE SCROLL DOWN FOR DOCUMENT

Full terms and conditions of use: https://www.routledgehandbooks.com/legal-notices/terms.

This Document PDF may be used for research, teaching and private study purposes. Any substantial or systematic reproductions, re-distribution, re-selling, loan or sub-licensing, systematic supply or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The publisher shall not be liable for an loss, actions, claims, proceedings, demand or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.

Downloaded By: 10.3.98.166 At: 03:56 15 Jun 2018; For: 9781315736013, chapter3, 10.4324/9781315736013.ch3

3 Modern Approaches to Parameter Estimation in Item Response Theory

Li Cai and David Thissen

Introduction

Entire volumes (e.g., Baker & Kim, 2004) have been dedicated to the discussion of statistical parameter estimation techniques for item response theory (IRT) models. There has also been much recent development in the technical literature on improved methods for estimating complex IRT models (e.g., Cai, 2010a, 2010b; Edwards, 2010; Rijmen & Jeon, 2013). We offer here a map to help researchers and graduate students understand the fundamental challenges of IRT parameter estimation, and appropriately contextualize the underlying logic of some of the proposed solutions. We assume that the reader is familiar with elementary probability concepts such as prior, posterior, and likelihood, as well as the equations for describing statistical models for categorical observed data, for example, logistic regression. For methodologically inclined readers interested in studying IRT parameter estimation and in trying out some of the approaches discussed here, the combination of conceptual sections and more technical sections should be sufficient as a basis of software implementation.

We do not discuss limited-information estimation methods derived from categorical factor analysis or structural equation modeling (see, e.g., Bolt, 2005), but not because estimators based on polychoric correlation matrices and weighted least squares are not useful. They certainly may be, when conditions are appropriate (see, e.g., Wirth & Edwards, 2007), and for a long time, limited-information methods provided the only practically viable means for conducting formal model appraisal; although on that latter point, the situation has changed dramatically in the past few years (see, e.g., Maydeu-Olivares, 2013). In choosing to focus exclusively on full-information approaches that are based on either likelihood or Bayesian derivations, we believe that we provide readers with insight that tends to be obscured by the technicality that tends to accompany limited-information approaches. That is, the latent variables in IRT models are missing data, and had the latent variable scores been available, estimation for IRT models would have been a rather straightforward task. For the sake of simplicity, we contextualize our discussion with unidimensional logistic IRT models for dichotomously scored outcomes, but the missing data formulation applies generally across a far wider range of statistical modeling frameworks, IRT modeling included.

RESEARCH METHODS

Univariate Logistic Regression

Some Notation

We begin our discussion with a familiar framework, a univariate logistic regression model for dichotomous outcomes. Let there be $j = 1, \ldots, J$ independent cases. For each case, let

Downloaded By: 10.3.98.166 At: 03:56 15 Jun 2018; For: 9781315736013, chapter3, 10.4324/9781315736013.ch3

Y_j denote a binomial random variable with conditional success probability μ_j, that depends on x_j, the value of a fixed and observed covariate/predictor. The number of trials can be understood as a weight variable n_j attached to case j, and Y_j is the number of successes out of n_j independent Bernoulli trials each with success probability μ_j.

We assume that the log-odds of success is described by a linear model

$$\log\left(\frac{\mu_j}{1-\mu_j}\right) = \eta_j = \alpha + \beta x_j, \tag{3.1}$$

in which α and β are the regression intercept and slope parameters, respectively. From Equation (3.1), the conditional probability μ_j may be expressed using the inverse transformation, that is, a logistic cumulative distribution function (CDF):

$$\mu_j = \frac{1}{1+\exp\left(-\eta_j\right)} = \frac{1}{1+\exp\left[-\left(\alpha+\beta x_j\right)\right]}. \tag{3.2}$$

Note that Equation (3.2) resembles the two-parameter logistic IRT model with the key distinction that in IRT, the predictor is a latent variable, whereas x_j is observed.

Given a sample of data, we may write the likelihood function of the regression parameters. The goal is to find the set of parameters that would serve to maximize the likelihood (or log-likelihood) given the observed data. The parameter estimates are the maximum likelihood estimates (MLEs). The nonlinearity of the model implies that direct, analytical solutions such as those found in the case of least squares linear regression analysis are not feasible, and iterative algorithms such as Newton-Raphson or Fisher Scoring must be employed.

Maximum Likelihood Estimation for Logistic Regression

For case j, omitting constant terms, the binomial likelihood function is

$$L\left(\alpha,\beta|y_j,x_j\right) \propto \mu_j^{y_j}\left(1-\mu_j\right)^{n_j-y_j}, \tag{3.3}$$

where μ_j is as in Equation (3.2), and y_i is the realized/observed value of Y_j. Invoking the assumption of independence of observations across cases, for the entire sample, the likelihood function becomes a product of individual likelihood contributions:

$$L\left(\alpha,\beta|\boldsymbol{y},\boldsymbol{x}\right) = \prod_{j=1}^{J} L(\alpha,\beta|y_j,x_j) = \prod_{j=1}^{J} \mu_j^{y_j}\left(1-\mu_j\right)^{n_j-y_j}, \tag{3.4}$$

where the vector $\boldsymbol{y}$ collects together all the observed outcomes, and $\boldsymbol{x}$ contains all the predictor values. At this point it is convenient to take the natural logarithm of the likelihood function in Equation (3.4), and the log-likelihood is a sum of individual case contributions

$$l\left(\alpha,\beta|\boldsymbol{y},\boldsymbol{x}\right) = \log L\left(\alpha,\beta|\boldsymbol{y},\boldsymbol{x}\right) = \sum_{j=1}^{J} y_j \log \mu_j + \left(n_j - y_j\right)\log\left(1-\mu_j\right). \tag{3.5}$$

Downloaded By: 10.3.98.166 At: 03:56 15 Jun 2018; For: 9781315736013, chapter3, 10.4324/9781315736013.ch3

To maximize the log-likelihood, one would need its first-order partial derivatives (also known as the gradient vector). Using the chain rule, we have:

$$\begin{aligned}\frac{\partial l\left(\alpha,\beta|\boldsymbol{y},\boldsymbol{x}\right)}{\partial\alpha}&=\sum_{j=1}^{J}\left(\frac{y_j}{\mu_j}\frac{\partial\mu_j}{\partial\eta_j}\frac{\partial\eta_j}{\partial\alpha}-\frac{\left(n_j-y_j\right)}{1-\mu_j}\frac{\partial\mu_j}{\partial\eta_j}\frac{\partial\eta_j}{\partial\alpha}\right),\\ \frac{\partial l\left(\alpha,\beta|\boldsymbol{y},\boldsymbol{x}\right)}{\partial\beta}&=\sum_{j=1}^{J}\left(\frac{y_j}{\mu_j}\frac{\partial\mu_j}{\partial\eta_j}\frac{\partial\eta_j}{\partial\beta}-\frac{\left(n_j-y_j\right)}{1-\mu_j}\frac{\partial\mu_j}{\partial\eta_j}\frac{\partial\eta_j}{\partial\beta}\right).\end{aligned}\tag{3.6}$$

From Equation (3.2), one can verify a convenient fact about the logistic CDF:

$$\frac{\partial\mu_j}{\partial\eta_j}=\mu_j\left(1-\mu_j\right).$$

Furthermore, the derivatives of the linear function η_j are conveniently:

$$\frac{\partial\eta_j}{\partial\alpha}=\frac{\partial}{\partial\alpha}\left(\alpha+\beta x_j\right)=1,\quad \frac{\partial\eta_j}{\partial\beta}=\frac{\partial}{\partial\beta}\left(\alpha+\beta x_j\right)=x_j.$$

Inserting these identities into Equation (3.6), we see that the expressions simplify considerably and the gradient vector of the log-likelihood is:

$$\boldsymbol{g}\left(\alpha,\beta|\boldsymbol{y},\boldsymbol{x}\right)=\begin{pmatrix}\dfrac{\partial l\left(\alpha,\beta|\boldsymbol{y},\boldsymbol{x}\right)}{\partial\alpha}\\ \dfrac{\partial l\left(\alpha,\beta|\boldsymbol{y},\boldsymbol{x}\right)}{\partial\beta}\end{pmatrix}=\begin{pmatrix}\displaystyle\sum_{j=1}^{J}\left(y_j-n_j\mu_j\right)\\ \displaystyle\sum_{j=1}^{J}\left(y_j-n_j\mu_j\right)x_j\end{pmatrix}.\tag{3.7}$$

Setting these derivatives to zero, the likelihood equations have remarkably direct interpretations. They amount to equating the observed counts in y to the expected counts in μ, summed over the individual contributions. We see that the likelihood equations are, however, nonlinear in α and β. Hence they cannot be solved analytically. We can use the Fisher Scoring method to solve the likelihood equations. To do so, we would need the second-order derivatives of the log-likelihood. Continuing from Equation (3.7), we see that

$$\begin{aligned}\frac{\partial^2 l\left(\alpha,\beta|\boldsymbol{y},\boldsymbol{x}\right)}{\partial\alpha^2}&=\frac{\partial}{\partial\alpha}\sum_{j=1}^{J}\left(y_j-n_j\mu_j\right)=-\sum_{j=1}^{J}n_j\frac{\partial\mu_j}{\partial\eta_j}\frac{\partial\eta_j}{\partial\alpha}=-\sum_{j=1}^{J}n_j\mu_j\left(1-\mu_j\right),\\ \frac{\partial^2 l\left(\alpha,\beta|\boldsymbol{y},\boldsymbol{x}\right)}{\partial\beta^2}&=\frac{\partial}{\partial\beta}\sum_{j=1}^{J}\left(y_j-n_j\mu_j\right)=-\sum_{j=1}^{J}n_j\frac{\partial\mu_j}{\partial\eta_j}\frac{\partial\eta_j}{\partial\beta}x_j=-\sum_{j=1}^{J}n_j\mu_j\left(1-\mu_j\right)x_j^2,\\ \frac{\partial^2 l\left(\alpha,\beta|\boldsymbol{y},\boldsymbol{x}\right)}{\partial\alpha\partial\beta}&=\frac{\partial}{\partial\alpha}\sum_{j=1}^{J}\left(y_j-n_j\mu_j\right)x_j=-\sum_{j=1}^{J}n_j\frac{\partial\mu_j}{\partial\eta_j}\frac{\partial\eta_j}{\partial\alpha}x_j=-\sum_{j=1}^{J}n_j\mu_j\left(1-\mu_j\right)x_j.\end{aligned}\tag{3.8}$$

Downloaded By: 10.3.98.166 At: 03:56 15 Jun 2018; For: 9781315736013, chapter3, 10.4324/9781315736013.ch3

The information matrix, which is minus one times the matrix of second-order derivatives of the log-likelihood function (with the latter known as the Hessian matrix), is equal to

$$-\mathcal{H}\left(\alpha,\beta|\boldsymbol{y},\boldsymbol{x}\right)=-\begin{pmatrix}\dfrac{\partial^2 l\left(\alpha,\beta|\boldsymbol{y},\boldsymbol{x}\right)}{\partial\alpha^2} & \\ \dfrac{\partial^2 l\left(\alpha,\beta|\boldsymbol{y},\boldsymbol{x}\right)}{\partial\alpha\partial\beta} & \dfrac{\partial^2 l(\alpha,\beta|\boldsymbol{y},\boldsymbol{x})}{\partial\beta^2}\end{pmatrix}. \tag{3.9}$$

If we choose as starting values some provisional estimate of intercept and slope, say, α_0 and β_0, and evaluate the gradient and information matrix at these provisional values, we would obtain $\boldsymbol{g}\left(\alpha_0,\beta_0|\boldsymbol{y},\boldsymbol{x}\right)$ and $-\mathcal{H}\left(\alpha_0,\beta_0|\boldsymbol{y},\boldsymbol{x}\right)$. The gradient vector and inverse of the information matrix may be combined to obtain a correction factor so that improved estimates become

$$\begin{pmatrix}\alpha_1\\ \beta_1\end{pmatrix}=\begin{pmatrix}\alpha_0\\ \beta_0\end{pmatrix}+\left[-\mathcal{H}\left(\alpha_0,\beta_0|\boldsymbol{y},\boldsymbol{x}\right)\right]^{-1}\boldsymbol{g}\left(\alpha_0,\beta_0|\boldsymbol{y},\boldsymbol{x}\right).$$

In general, from provisional estimates α_k and β_k, $k = 0, \ldots,$ Fisher Scoring uses the iterations

$$\begin{pmatrix}\alpha_{k+1}\\ \beta_{k+1}\end{pmatrix}=\begin{pmatrix}\alpha_k\\ \beta_k\end{pmatrix}+\left[-\mathcal{H}\left(\alpha_k,\beta_k\,|\,\boldsymbol{y},\boldsymbol{x}\right)\right]^{-1}\boldsymbol{g}\left(\alpha_k,\beta_k\,|\,\boldsymbol{y},\boldsymbol{x}\right), \tag{3.10}$$

to gradually improve the provisional estimates. Under general conditions, the sequence of estimates generated by the Fisher Scoring iterations converges to the MLE as k increases without bounds. At the converged solution, the inverse of the information matrix provides an estimate of the large sample covariance matrix of the parameter estimates.

Item Response Theory Model as Multivariate Logistic Regression

Some Notation

Suppose a hypothetical assessment is made up of $i = 1, \ldots, I$ dichotomously scored items. An item score of one indicates a correct or endorsement response, and zero otherwise. Furthermore, suppose that the assumption of unidimensionality holds for this set of items. Let us use the standard notation of θ_j to denote the latent variable score for individual j. The two-parameter logistic (2PL) item response model specifies the conditional response probability curve (also known as the traceline) of a correct response or endorsement as a function of the latent variable and the item parameters:

$$T_i\left(\theta;\alpha_i,\beta_i\right)=\frac{1}{1+\exp\left[-\left(\alpha_i+\beta_i\theta\right)\right]}, \tag{3.11}$$

where α_i and β_i are the item intercept and slope parameters. The parentheses in $T_i\left(\theta;\alpha_i,\beta_i\right)$ highlight the fact that the response probabilities are conditional on θ_j, and that they also depend on the item parameters. Let Y_{ij} be a Bernoulli (0–1) random variable representing

Downloaded By: 10.3.98.166 At: 03:56 15 Jun 2018; For: 9781315736013, chapter3, 10.4324/9781315736013.ch3

individual i's response to item j, and let y_{ij} be a realization of Y_{ij}. This suggests a formulation of the conditional probability of the event $Y_{ij} = y_{ij}$ similar to Equation (3.4),

$$P\left(Y_{ij} = y_{ij} | \theta; \alpha_i, \beta_i\right) = \left[T_i\left(\theta; \alpha_i, \beta_i\right)\right]^{y_{ij}} \left[1 - T_i\left(\theta; \alpha_i, \beta_i\right)\right]^{1-y_{ij}}. \tag{3.12}$$

Under the assumption of unidimensionality, the latent variable θ alone explains all the observed covariations among the items. In other words, conditionally on θ, the item response probabilities are independent for an individual, that is, the probability of response pattern $\boldsymbol{y}_j = \left(y_{1j}, \ldots, y_{Ij}\right)$ factors into a product over individual item response probabilities:

$$\begin{aligned} P\left(\boldsymbol{y}_j | \theta; \gamma\right) &= \prod_{i=1}^{I} P\left(Y_{ij} = y_{ij} | \theta; \alpha_i, \beta_i\right) \\ &= \prod_{i=1}^{I} \left[T_i\left(\theta; \alpha_i, \beta_i\right)\right]^{y_{ij}} \left[1 - T_i\left(\theta; \alpha_i, \beta_i\right)\right]^{1-y_{ij}}. \end{aligned} \tag{3.13}$$

where on the left-hand side we collect all item intercept and slope parameters into $\boldsymbol{\gamma} = (\alpha_1, \ldots, \alpha_I, \beta_1, \ldots, \beta_I)$, a $2I$-dimensional vector. The joint probability of the observed and latent variables is equal to the product of the conditional probability of the observed variables given the latent variables, times the prior probability of the latent variables:

$$P\left(\boldsymbol{y}_j, \theta; \gamma\right) = \prod_{i=1}^{I} \left[T_i\left(\theta; \alpha_i, \beta_i\right)\right]^{y_{ij}} \left[1 - T_i\left(\theta; \alpha_i, \beta_i\right)\right]^{1-y_{ij}} h(\theta), \tag{3.14}$$

where $h(\theta)$ is prior (population) distribution of the latent variable θ. In IRT applications, it is customary to resolve the location and scale indeterminacy of the latent variable by assuming that the θ's are standard normal, so $h(\theta)$ does not contain free parameters.

From Equation (3.14), a natural derived quantity is the marginal probability of the response pattern, after integrating the joint probability over θ:

$$P\left(\boldsymbol{y}_j; \gamma\right) = \int \prod_{i=1}^{I} \left[T_i\left(\theta; \alpha_i, \beta_i\right)\right]^{y_{ij}} \left[1 - T_i\left(\theta; \alpha_i, \beta_i\right)\right]^{1-y_{ij}} h(\theta) d\theta. \tag{3.15}$$

Unfortunately Equation (3.15) is already the simplest form that we can obtain, given the combination of the IRT model and normally distributed latent variable. Note that the marginal probability does not depend on the unobserved latent variable scores; it is a function solely of the observed item response pattern and the item parameters.

As in the case of logistic regression, we assume the individuals are independent, with latent variable scores sampled independently from the population distribution. Let $\mathbf{Y}$ be a matrix of all observed item responses. If we treat the item responses as fixed once observed, the marginal likelihood function for all the item parameters in $\boldsymbol{\gamma}$, based on observed item response data, can be expressed as:

$$L\left(\gamma | \boldsymbol{Y}\right) = \prod_{j=1}^{J} \int \prod_{i=1}^{I} \left[T_i\left(\theta; \alpha_i, \beta_i\right)\right]^{y_{ij}} \left[1 - T_i\left(\theta; \alpha_i, \beta_i\right)\right]^{1-y_{ij}} h(\theta) d\theta. \tag{3.16}$$

Because the marginal likelihood $L\left(\boldsymbol{\gamma} | \boldsymbol{Y}\right)$ does not depend on the unobserved θ values, it may be referred to as the *observed* data likelihood.

Downloaded By: 10.3.98.166 At: 03:56 15 Jun 2018; For: 9781315736013, chapter3, 10.4324/9781315736013.ch3

Under some circumstances, this likelihood function can be optimized directly, again using Newton Raphson or Fisher Scoring-type algorithms (see, e.g., Bock & Lieberman, 1970), but those circumstances are rather limited. In particular, Bock and Lieberman (1970) noted that this direct approach does not generalize well to the case of many items and many parameters because of computing demands. We would add that even as computers have become faster and storage cheaper, what the direct approach glosses over is a *missing data formulation* of latent variable models that is central to our understanding of IRT and of other modern statistical techniques such as random effects regression modeling, or modeling of survey nonresponse. This missing data formulation was made transparent by Dempster, Laird, and Rubin's (1977) classical paper that coined the term *Expectation-Maximization (EM) algorithm*.

Missing Data Formulation and Fisher's Identity

Implicit in the "observed data" terminology is a realization that θ contains the "missing" data. If we treat the item responses as fixed once observed, and also suppose the latent variable scores were observed, then after some algebra that follows directly from Equation (3.14), we see that the so-called complete data likelihood function of the vector of item parameters is:

$$\begin{aligned} L(\boldsymbol{\gamma}|\mathbf{Y},\boldsymbol{\theta}) &= \left[\prod_{j=1}^{J} h(\theta_j)\right]\left[\prod_{j=1}^{J}\prod_{i=1}^{I}[T_i(\theta_j;\alpha_i,\beta_i)]^{y_{ij}}[1-T_i(\theta_j;\alpha_i,\beta_i)]^{1-y_{ij}}\right] \\ &\propto \prod_{i=1}^{I}\prod_{j=1}^{J}\left[T_i(\theta_j;\alpha_i,\beta_i)]^{y_{ij}}[1-T_i(\theta_j;\alpha_i,\beta_i)\right]^{1-y_{ij}}, \end{aligned} \tag{3.17}$$

where θ is a vector that collects together all J latent variable scores. The proportionality on the second line holds because $h(\theta_j)$ does not depend on item parameters in our model, and given θ_j, it becomes a constant. Had the latent variable scores been observed, Equation (3.17) makes it clear that the complete data likelihood function would be a constant multiple of I item-specific likelihoods, each representing a logistic regression model. Thus the IRT model can be understood as multivariate logit analysis, if one could observe the predictor variable θ.

Of course, the latent variable θ is not observed, but that does not imply the situation is hopeless. Instead, it forces us to pay close attention to the posterior distribution of the latent variable given the observed item responses:

$$P(\theta|\mathbf{y}_j;\boldsymbol{\gamma}) = \frac{P(\mathbf{y}_j,\theta;\boldsymbol{\gamma})}{P(\mathbf{y}_j;\boldsymbol{\gamma})} = \frac{\prod_{i=1}^{I}[T_i(\theta;\alpha_i,\beta_i)]^{y_{ij}}[1-T_i(\theta;\alpha_i,\beta_i)]^{1-y_{ij}}h(\theta)}{\int\prod_{i=1}^{I}[T_i(\theta;\alpha_i,\beta_i)]^{y_{ij}}[1-T_i(\theta;\alpha_i,\beta_i)]^{1-y_{ij}}h(\theta)d\theta}. \tag{3.18}$$

This is an analytically intractable distribution, but it follows directly from the application of the Bayes rule to Equations (3.14) and (3.15). It also has some interesting characteristics that deserve comments. First, given item parameter values and the observed item response pattern y_j, the denominator is a constant that can, in principle, be computed. This is a normalization factor that makes (3.18) a proper probability density function. Second, given item parameter values, the posterior is proportional to the joint distribution in the numerator, which is more tractable than the posterior itself. Third, with the help of the posterior distribution, one may verify that given item parameter values, the following equality holds (it is known as Fisher's Identity; Fisher, 1925) assuming mild regularity conditions:

$$\frac{\partial \log P(\mathbf{y}_j;\boldsymbol{\gamma})}{\partial\boldsymbol{\gamma}} = \int\frac{\partial \log P(\mathbf{y}_j,\theta;\boldsymbol{\gamma})}{\partial\boldsymbol{\gamma}}P(\theta|\mathbf{y}_j;\boldsymbol{\gamma})d\theta. \tag{3.19}$$

Downloaded By: 10.3.98.166 At: 03:56 15 Jun 2018; For: 9781315736013, chapter3, 10.4324/9781315736013.ch3

Fisher's Identity states that the gradient of the observed data log-likelihood *log* L($\boldsymbol{\gamma}|\boldsymbol{Y}$) is equal to the conditional expectation of the gradient of the completed data log-likelihood *log* L($\boldsymbol{\gamma}|\boldsymbol{Y}$, $\boldsymbol{\theta}$) over the posterior distribution of the latent variables given the observed variables. This powerful insight suggests that instead of trying to maximize the observed data likelihood, which is direct but often difficult, one should consider iteratively maximizing the conditional expected complete data likelihood (i.e., the right-hand side of Equation 3.19), which can be an indirect route but more computationally tractable. This is because the complete data model is no more than a set of logistic regressions, which is a problem we already know how to solve. We will demonstrate this argument via two approaches, beginning with Bock and Aitkin's (1981) classical application of the EM algorithm, and then turning to its modern cousin, Cai's (2008) Metropolis-Hastings Robbins-Monro algorithm.

Bock-Aitkin EM Algorithm

Bock and Aitkin (1981) began with the insight that the marginal probability can be approximated to arbitrary precision by replacing the integration with a summation over a set of Q quadrature points over θ:

$$P\left(\boldsymbol{y}_j;\gamma\right) \approx \bar{P}_j = \sum_{q=1}^{Q}\prod_{i=1}^{I}\left[T_i(X_q;\alpha_i,\beta_i)\right]^{y_{ij}}\left[1-T_i(X_q;\alpha_i,\beta_i)\right]^{1-y_{ij}} W_q, \tag{3.20}$$

where X_q is a quadrature point, and W_q is the corresponding weight. In the simplest case, one may take the quadrature points as a set of equally spaced real numbers over an interval that captures sufficiently the probability mass of the population distribution, for example, from −6 to +6 in increments of 0.1, and the corresponding weights as a set of normalized ordinates of the quadrature points from the population distribution $W_q = h\left(X_q\right)\Big/\sum_{q=1}^{Q} h\left(X_q\right)$.

Another important insight of Bock and Aitkin (1981) is that the height of the posterior distribution at quadrature point X_q can be approximated to arbitrary precision as well:

$$P\left(X_q|\boldsymbol{y}_j;\gamma\right) \approx \frac{\left[T_i\left(X_q;\alpha_i,\beta_i\right)\right]^{y_{ij}}\left[1-T_i\left(X_q;\alpha_i,\beta_i\right)\right]^{1-y_{ij}} W_q}{\bar{P}_j}. \tag{3.21}$$

Ignoring constants involving the prior distribution $h\left(\theta_j\right)$ from Equation (3.17), the complete data log-likelihood for the item parameters can be written as:

$$\log L\left(\gamma|\mathbf{Y},\theta\right) = \sum_{j=1}^{J}\sum_{i=1}^{I} y_{ij}\log T_i(\theta_j;\alpha_i,\beta_i) + \sum_{j=1}^{J}\sum_{i=1}^{I}(1-y_{ij})\log\left(1-T_i(\theta_j;\alpha_i,\beta_i)\right). \tag{3.22}$$

Following the logic inspired by the Fisher Identity, the conditional expected complete data likelihood given provisional item parameter values $\boldsymbol{\gamma}^* = (\alpha_1^*,\ldots,\alpha_I^*,\beta_1^*,\ldots,\beta_I^*)$ can be approximated by quadrature, case by case, as follows:

$$\begin{aligned} Q(\gamma|\mathbf{Y};\gamma^*) \approx & \sum_{j=1}^{J}\sum_{q=1}^{Q}\sum_{i=1}^{I} y_{ij} log T_i(X_q;\alpha_i,\beta_i) P(X_q|\boldsymbol{y}_j;\gamma^*) \\ & + \sum_{j=1}^{J}\sum_{q=1}^{Q}\sum_{i=1}^{I}(1-\boldsymbol{y}_{ij}) log(1-T_i(X_q;\alpha_i,\beta_i)) P(X_q|\boldsymbol{y}_j;\gamma^*). \end{aligned} \tag{3.23}$$

Downloaded By: 10.3.98.166 At: 03:56 15 Jun 2018; For: 9781315736013, chapter3, 10.4324/9781315736013.ch3

The third and arguably most important insight from Bock and Aitkin (1981) is that by interchanging the order of summation, they realized that the posterior probabilities can be accumulated over individuals first:

$$\begin{aligned} Q(\boldsymbol{\gamma}|\mathbf{Y};\boldsymbol{\gamma}^*) &\approx \sum_{i=1}^{I}\sum_{q=1}^{Q}\log T_i(X_q;\alpha_i,\beta_i)\left(\sum_{j=1}^{J} y_{ij}P(X_q|\mathbf{y}_j;\boldsymbol{\gamma}^*)\right) \\ &\quad + \sum_{i=1}^{I}\sum_{q=1}^{Q}\log\left(1-T_i(X_q;\alpha_i,\beta_i)\right)\left(\sum_{j=1}^{J}(1-y_{ij})P(X_q|\mathbf{y}_j;\boldsymbol{\gamma}^*)\right) \\ &= \sum_{i=1}^{I}\sum_{q=1}^{Q} r_{iq}\log T_i(X_q;\alpha_i,\beta_i) + \sum_{i=1}^{I}\sum_{q=1}^{Q}\bar{r}_{iq}\log(1-T_i(X_q;\alpha_i,\beta_i)), \end{aligned} \tag{3.24}$$

where $r_{iq} = \sum_{j=1}^{J} y_{ij}P\left(X_q|\mathbf{y}_j;\boldsymbol{\gamma}^*\right)$ is understood as the conditional expected proportion of individuals that respond positively/correctly to item i, and $\bar{r}_{iq} = \sum_{j=1}^{J}\left(1-y_{ij}\right)P\left(X_q|\mathbf{y}_j;\boldsymbol{\gamma}^*\right)$ is the conditional expected proportion of individuals that respond negatively/incorrectly to item i, at quadrature point X_q. Taken together, let $n_{iq} = r_{iq} + \bar{r}_{iq} = \sum_{j=1}^{J}P\left(X_q|\mathbf{y}_j;\boldsymbol{\gamma}^*\right)$ be the conditional expected proportion of individuals at quadrature point X_q, then we have:

$$Q\left(\boldsymbol{\gamma}|\mathbf{Y};\boldsymbol{\gamma}^*\right) \approx \sum_{i=1}^{I}\sum_{q=1}^{Q} r_{iq}\log T_i(X_q;\alpha_i,\beta_i) + (n_{iq}-r_{iq})\log\left(1-T_i(X_q;\alpha_i,\beta_i)\right). \tag{3.25}$$

Equation (3.25) highlights the fact that the conditional expected complete data log-likelihood is a set of I independent logistic regression log-likelihoods, with the quadrature points X_q serving as the predictor values, and weights given by n_{iq}, and r_{iq} serving as the positive outcome "frequency" at X_q. The inner summation over the quadrature points bears striking similarity to the log-likelihood given in Equation (3.5). The only difference is that in standard logistic regression, the weights n_j and number of successes y_j are integers, whereas in the case of Bock-Aitkin EM, n_{iq} and r_{iq} may be fractional and will change from cycle to cycle, given different item parameter values. With the Fisher Scoring algorithm developed in Section 2, optimization of $Q(\boldsymbol{\gamma}|\mathbf{Y};\boldsymbol{\gamma}^*)$ is straightforward, which leads to updated parameter estimates that may be used in the next cycle.

In general, Bock-Aitkin EM (or any EM algorithm) alternates between the following two steps from a set of initial parameter estimates, say $\boldsymbol{\gamma}^{(0)}$, and it generates a sequence of parameter estimates $\boldsymbol{\gamma}^{(0)},\ldots,\boldsymbol{\gamma}^{(k)},\ldots$, where $\boldsymbol{\gamma}^{(k)} = \left(\alpha_1^{(k)},\ldots,\alpha_I^{(k)},\beta_1^{(k)},\ldots,\beta_I^{(k)}\right)$, that converges under some very general conditions to the MLE of $\boldsymbol{\gamma}$ as the number of cycles k tends to infinity (Wu, 1983):

E-step. Given $\boldsymbol{\gamma}^{(k)}$, evaluate the conditional expected complete data log-likelihood $Q\left(\boldsymbol{\gamma}|\mathbf{Y};\boldsymbol{\gamma}^{(k)}\right)$, which is taken to be a function of $\boldsymbol{\gamma}$.

M-step. Maximize $Q\left(\boldsymbol{\gamma}|\mathbf{Y};\boldsymbol{\gamma}^{(k)}\right)$ to yield updated parameter estimates $\boldsymbol{\gamma}^{(k+1)}$. Go back to E-step and repeat. The cycles are terminated when the estimates from adjacent cycles stabilize.

Downloaded By: 10.3.98.166 At: 03:56 15 Jun 2018; For: 9781315736013, chapter3, 10.4324/9781315736013.ch3

The application of the EM algorithm to IRT epitomizes the elegance of the missing data formulation in statistical computing. Finding MLEs in logistic regression analysis is a task that statisticians already know how to do. The goal of the E-step, then, is to replace the missing data with conditional expectations that depend on values of θ, represented using a set of discrete quadrature points. Once the missing data are filled in, complete data estimation can be accomplished with tools that are already available. Leveraging the conditional independence built into the IRT model, the M-step logit analyses can even be run in parallel and the overall demand on computing resources is rather low. Although the EM algorithm is only first-order (linearly) convergent, and may be slow (by optimization researchers' standard), the statistical intuition is simply too elegant to ignore. Thissen (1982) extended the unconstrained Bock-Aitkin EM to handle parameter restrictions and used it to estimate the Rasch IRT model.

Metropolis-Hastings Robbins-Monro Algorithm

Motivations of MH-RM

One issue with Bock-Aitkin EM is that while it deftly handles unidimensional IRT parameter estimation with many items, it does not generalize well to the case of multidimensional IRT. This is because the posterior expectations must be accumulated over grids of quadrature points formed by the direct product of the quadrature rule. Even with a moderate number of quadrature points, the exponentially increasing size of the grid as the number of dimensions increases presents major computational challenges. Adaptive quadrature helps somewhat by requiring fewer points than fixed quadrature rules (see, e.g., Schilling & Bock, 2005), but does not solve the problem completely. Various authors (e.g., Wirth & Edwards, 2007) referred to this as the "challenge of dimensionality." As assessments become more complex, multidimensional IRT models are increasingly in demand, but estimating the item parameters has been difficult.

Cai (2006, 2008, 2010a, 2010b) realized that a solution already resides in Fisher's Identity. It is worth repeating that equation:

$$\frac{\partial \log P\left(\boldsymbol{y}_j;\boldsymbol{\gamma}\right)}{\partial \boldsymbol{\gamma}} = \int \frac{\partial \log P\left(\boldsymbol{y}_j,\theta;\boldsymbol{\gamma}\right)}{\partial \boldsymbol{\gamma}} P\left(\theta \,|\, \boldsymbol{y}_j;\boldsymbol{\gamma}\right) d\theta.$$

Cai reasoned that if one can randomly draw plausible values or imputations of θ from its posterior predictive distribution $P\left(\theta \,|\, \boldsymbol{y}_j;\boldsymbol{\gamma}^*\right)$, with provisional item parameter estimates $\boldsymbol{\gamma}^*$, the right-hand side can be approximated by Monte Carlo, that is,

$$\frac{\partial \log P\left(\boldsymbol{y}_j;\boldsymbol{\gamma}\right)}{\partial \boldsymbol{\gamma}} \approx \frac{1}{M}\sum_{m=1}^{M} \frac{\partial \log P\left(\boldsymbol{y}_j,\theta_{jm};\boldsymbol{\gamma}\right)}{\partial \boldsymbol{\gamma}}, \tag{3.26}$$

where θ_{jm} are the random draws from $P\left(\theta \,|\, \boldsymbol{y}_j;\boldsymbol{\gamma}^*\right)$. Because the cases are independent, we also see that:

$$\begin{aligned}\frac{\partial \log L\left(\boldsymbol{\gamma} \,|\, \boldsymbol{Y}\right)}{\partial \boldsymbol{\gamma}} &= \sum_{j=1}^{J} \frac{\partial \log P\left(\boldsymbol{y}_j;\boldsymbol{\gamma}\right)}{\partial \boldsymbol{\gamma}} \approx \frac{1}{M}\sum_{m=1}^{M}\sum_{j=1}^{J} \frac{\partial \log P\left(\boldsymbol{y}_j,\theta_{jm};\boldsymbol{\gamma}\right)}{\partial \boldsymbol{\gamma}} \\ &= \frac{1}{M}\sum_{m=1}^{M} \frac{\partial \log L\left(\boldsymbol{\gamma} \,|\, \boldsymbol{Y},\boldsymbol{\theta}_{\boldsymbol{m}}\right)}{\partial \boldsymbol{\gamma}},\end{aligned} \tag{3.27}$$

Downloaded By: 10.3.98.166 At: 03:56 15 Jun 2018; For: 9781315736013, chapter3, 10.4324/9781315736013.ch3

where $(\boldsymbol{Y}, \boldsymbol{\theta}_{\mathrm{m}})$ may be taken as the mth complete data set, and $\boldsymbol{\theta}_{\mathrm{m}}$ is the augmented missing data. We end up with the first insight that motivates the MH-RM algorithm: *The Monte Carlo average of complete data log-likelihood gradients gives the same likelihood ascent direction as the observed data log-likelihood gradient vector*.

An immediate problem with the Monte Carlo approximation is that it contains error, and unless the Monte Carlo size M becomes large, the random sampling error obscures the true direction of likelihood ascent. This is a known issue in the context of Monte Carlo EM (Booth & Hobert, 1999), where the solution is to adaptively increase the size of the Monte Carlo sampling, so that increasingly accurate approximations can be found as the estimates converge. Unfortunately, as will be explained later, while computing random draws of the $\boldsymbol{\theta}$'s has become an increasingly manageable task, with help from Markov chain Monte Carlo (MCMC; Hastings, 1970; Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953), it is still a nontrivial matter for fitting of IRT models in practical settings, because of the multitude of nonlinear functions (exponential, for instance) that must be evaluated for the IRT model's likelihood functions. The amount of computing time required to draw the imputations frequently dwarfs the amount of time needed to compute the complete data derivatives by several orders of magnitude. Thus it is necessary, if only for computational efficiency, to find a method that utilizes Monte Carlo sampling effectively.

Cai (2010a) noted that instead of treating the Monte Carlo noise as a nuisance to be contained, it may in fact be employed more productively. By drawing an analogy to the engineering applications of Robbins and Monro's (1951) classical Stochastic Approximation (SA) method, the Monte Carlo noise provides the stochastic excitations that drive an underlying stochastic process. The noise is gradually filtered out with the use of an appropriately chosen sequence of gain constants, as the parameters are recursively updated.

This leads to the second insight that leads to MH-RM: *In Robbins and Monro's context, they were attempting to find roots of noise-corrupted regression functions, where the noise may be due to observational measurement error; in our context, we purposefully inject Monte Carlo noise by imputing the missing data (the latent variable scores), so that we can observe an approximate direction of likelihood ascent*. It is not necessary that the approximate ascent direction be made precise, especially in the beginning stages of the iterative scheme. In fact, it is possible to let M be identically equal to one (a single imputation per iteration) and still obtain a point-wise convergent algorithm to the MLE (see Cai, 2010a, for a proof).

Definition of the Algorithm

With the IRT model, cycle $k+1$ of the MH-RM algorithm consists of three steps:

Imputation. Given provisional parameter estimates $\boldsymbol{\gamma}^{(k)}$ from the previous cycle (or initial parameter values $\boldsymbol{\gamma}^{(0)}$ if this is the first cycle), random samples of the latent variables $\boldsymbol{\theta}_m^{(k+1)}$ are imputed. For each individual, the draws may come from a Metropolis-Hastings sampler that has, as its unique invariant distribution, the posterior predictive distribution $P\left(\theta_j \middle| \boldsymbol{y}_j; \boldsymbol{\gamma}^{(k)}\right)$ of the missing data given the observed data and provisional parameter values. In other words, the complete data sets are formed as $\left(\mathbf{Y}, \boldsymbol{\theta}_m^{(k+1)}\right)$.

Approximation. In the second step, based on the imputed data, the complete data log-likelihood and its derivatives are evaluated so that the ascent directions for parameters can be determined. The complete data gradient (score) function is approximated as:

$$\boldsymbol{s}_{k+1} = \frac{1}{M}\sum_{m=1}^{M} \frac{\partial \log L\left(\boldsymbol{\gamma}^{(k)} \middle| \mathbf{Y}, \boldsymbol{\theta}_m^{(k+1)}\right)}{\partial \boldsymbol{\gamma}}, \tag{3.28}$$

Downloaded By: 10.3.98.166 At: 03:56 15 Jun 2018; For: 9781315736013, chapter3, 10.4324/9781315736013.ch3

Note that for each item, the complete data gradient vector is simply $g\left(\alpha_i^{(k)}, \beta_i^{(k)} \middle| \boldsymbol{y}_i, \boldsymbol{\theta}_m^{(k+1)}\right)$, as defined in Equation (3.7), where $\boldsymbol{y}_i$ is a vector that collects together all J observed responses to item i. At the same time, to improve stability and speed, we also evaluate a Monte Carlo approximation to the conditional expected complete data information matrix:

$$\boldsymbol{H}_{k+1} = -\frac{1}{M}\sum_{m=1}^{M}\frac{\partial^2 \log L\left(\boldsymbol{\gamma}^{(k)} \middle| \mathbf{Y}, \boldsymbol{\theta}_m^{(k+1)}\right)}{\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}'}. \tag{3.29}$$

Again, because of conditional independence given the latent variables in the IRT model, the complete data information matrix is block-diagonal, with each item's information matrix equal to $-\mathcal{H}\left(\alpha_i^{(k)}, \beta_i^{(k)} \middle| \boldsymbol{y}_i, \boldsymbol{\theta}_m^{(k+1)}\right)$, as defined in Equation (3.9).

Robbins-Monro Update. In the third step, Robbins-Monro stochastic approximation filters are applied when updating the estimates of item parameters. First, the Robbins-Monro filter is applied to obtain a recursive stochastic approximation of the conditional expectation of the complete data information matrix:

$$\boldsymbol{\Gamma}_{k+1} = \boldsymbol{\Gamma}_k + \epsilon_k\left(\mathbf{H}_{k+1} - \boldsymbol{\Gamma}_k\right), \tag{3.30}$$

where ϵ_k is a sequence of non-negative *gain constants* such that $\epsilon_k \in (0,1], \sum_{k=0}^{\infty}\epsilon_k = \infty, \sum_{k=0}^{\infty}\epsilon_k^2 < \infty$. Next, we use the Robbins-Monro filter again when updating the parameters:

$$\boldsymbol{\gamma}^{(k+1)} = \boldsymbol{\gamma}^{(k)} + \epsilon_k(\boldsymbol{\Gamma}_{k+1})^{-1}\boldsymbol{s}_{k+1}. \tag{3.31}$$

The iterations are started from some initial parameter values $\boldsymbol{\gamma}^{(0)}$ and terminated when the estimates stabilize. Cai (2008, 2010a) showed that the sequence of parameters converges with probability 1 to a local maximum of $L(\boldsymbol{\gamma}|\mathbf{Y})$. Typically, the sequence of gain constants are taken to be $\epsilon_k = 1/(k+1)$, in which case the initial choice of $\boldsymbol{\Gamma}_0$ becomes arbitrary. Cai (2010a) contains formulas for recursively approximating the parameter error covariance matrix, as well as further discussions about convergence checking.

Implementing the Metropolis-Hastings Sampler

At this point, a critical missing link is a method to draw random values of $\boldsymbol{\theta}_m^{(k+1)}$ from its posterior predictive distribution. Cai (2006) proposed the use of the Metropolis-Hastings method, for several reasons. First, we see from Equation (3.18) that while the posterior predictive distribution is analytically intractable in that it does not belong to any "named" distribution family, it is proportional to the joint probability of observed item responses and latent variables:

$$P\left(\theta \middle| \boldsymbol{y}_j; \boldsymbol{\gamma}^{(k)}\right) \propto \prod_{i=1}^{I}\left[T_i\left(\theta; \alpha_i^{(k)}, \beta_i^{(k)}\right)\right]^{y_{ij}}\left[1 - T_i\left(\theta; \alpha_i^{(k)}, \beta_i^{(k)}\right)\right]^{1-y_{ij}} h(\theta). \tag{3.32}$$

The Metropolis-Hastings method is ideally suited to the task of sampling a posterior when the normalization constant is not readily available. In addition, the right-hand side of Equation (3.32) is the complete data likelihood at $\boldsymbol{\gamma}^{(k)}$, which is evaluated in any event to

Downloaded By: 10.3.98.166 At: 03:56 15 Jun 2018; For: 9781315736013, chapter3, 10.4324/9781315736013.ch3

compute the item gradients and information matrices required in the approximation step of MH-RM. Furthermore, the sampling of the θ values can be accomplished in parallel, as the individual $P\left(\theta_j|\mathbf{y}_j;\boldsymbol{\gamma}^{(k)}\right)$'s are fully independent. Finally, the Monte Carlo approximation in Equation (3.28) remains unbiased even if the draws are not independent, for example from a Markov chain.

Implementing the Metropolis-Hastings method is straightforward. For each individual j, we begin with some initial value of θ_j, say, θ_j^c, and let us call it the current state of θ_j. We now draw a random increment from an independent normal sampler, with mean 0 and standard deviation equal to σ. Let this increment value be denoted δ_j. By adding the increment to the current state, we have produced a proposal for a new state of θ_j: $\theta_j^p = \theta_j^c + \delta_j$. We now evaluate the right-hand side of Equation (3.32) at both current and proposal states, and form the following likelihood ratio:

$$R(\theta_j^p, \theta_j^c) = \frac{\prod_{i=1}^{I}\left[T_i\left(\theta_j^p; \alpha_i^{(k)}, \beta_i^{(k)}\right)\right]^{y_{ij}}\left[1-T_i\left(\theta_j^p; \alpha_i^{(k)}, \beta_i^{(k)}\right)\right]^{1-y_{ij}} h(\theta_j^p)}{\prod_{i=1}^{I}\left[T_i\left(\theta_j^c; \alpha_i^{(k)}, \beta_i^{(k)}\right)\right]^{y_{ij}}\left[1-T_i\left(\theta_j^c; \alpha_i^{(k)}, \beta_i^{(k)}\right)\right]^{1-y_{ij}} h(\theta_j^c)}. \tag{3.33}$$

If $R\left(\theta_j^p, \theta_j^c\right)$ is larger than 1.0, meaning that the proposed move to a new state increased the likelihood relative to the current state, we accept the move and set the proposal state as the new current state. If $R\left(\theta_j^p, \theta_j^c\right)$ is smaller than 1.0, meaning that the proposed move decreased the likelihood, we accept the move with probability equal to the likelihood ratio. This can be accomplished by drawing, independently, a uniform (0,1) random number u_j, and comparing it to $R\left(\theta_j^p, \theta_j^c\right)$. If u_j is smaller than $R\left(\theta_j^p, \theta_j^c\right)$, we accept the proposed move and set the proposal state as the new current state. If u_j is larger than the likelihood ratio, we reject the proposal, and remain at the current state. Iterating this sampler will produce a Markov chain that converges to $P\left(\theta_j|\mathbf{y}_j;\boldsymbol{\gamma}^{(k)}\right)$ in distribution.

As the chain evolves, dependent samples from this chain can be regarded as samples from the target distribution. To avoid excessive dependence on the initial state, one can drop the samples in the so-called burn-in phase of the chain. For the IRT model, experience suggests that this burn-in phase typically amounts to not more than 10 iterations of the Metropolis-Hastings sampler. Of course, this assumes that the chain is appropriately tuned by monitoring the rate of acceptance of the proposed moves and scaling the increment density standard deviation σ up (for decreased acceptance rate) or down (for increased acceptance rate). Roberts and Rosenthal (2001) discussed the statistical efficiency of Metropolis-Hastings samplers, and its relationship to optimal scaling. Asymptotically efficient chains can be obtained by tuning the acceptance rate to around 25 percent.

Application

We analyze a well-known data set (Social Life Feelings), analyzed by Bartholomew (1998), among others, to illustrate the Bock-Aitkin EM and MH-RM algorithms. The data set contains responses from J =1,490 German respondents to five statements on perceptions of social life. The responses were dichotomous (endorsement vs. non-endorsement of the statements). Table 3.1 presents the $2^5 = 32$ response patterns and their associated observed response frequencies.

Let us first examine the most frequently encountered response pattern (0, 1, 1, 0, 0), wherein 208 respondents endorsed items 2 and 3 and none of the others. Following the logic of Bock-Aitkin EM, we must first choose a set of quadrature points for

Downloaded By: 10.3.98.166 At: 03:56 15 Jun 2018; For: 9781315736013, chapter3, 10.4324/9781315736013.ch3

Table 3.1 Social Life Feelings Data in Response Pattern by Frequency Form

Item Response Pattern					*Observed Frequency*
0	0	0	0	0	156
0	0	0	0	1	26
0	0	0	1	0	14
0	0	0	1	1	9
0	0	1	0	0	127
0	0	1	0	1	26
0	0	1	1	0	66
0	0	1	1	1	16
0	1	0	0	0	174
0	1	0	0	1	35
0	1	0	1	0	36
0	1	0	1	1	13
0	1	1	0	0	208
0	1	1	0	1	65
0	1	1	1	0	195
0	1	1	1	1	129
1	0	0	0	0	8
1	0	0	0	1	2
1	0	0	1	0	1
1	0	0	1	1	3
1	0	1	0	0	4
1	0	1	0	1	4
1	0	1	1	0	18
1	0	1	1	1	9
1	1	0	0	0	8
1	1	0	0	1	2
1	1	0	1	0	5
1	1	0	1	1	3
1	1	1	0	0	19
1	1	1	0	1	10
1	1	1	1	0	31
1	1	1	1	1	68
					Total = 1,490

approximating the E-step integrals. Here we use a set of 49 quadrature points equally spaced between −6 and +6. Next we must also choose a set of initial values for the item parameters. For the sake of variety, we let the initial values of the item intercepts be $\alpha_1 = -1.5$, $\alpha_2 = -1$, $\alpha_3 = 0$, $\alpha_4 = 1$, $\alpha_5 = 1.5$, and let all initial slopes be equal to 1.0. We are now ready to begin our first E-step.

Figure 3.1 contains a set of three plots showing the relationship between the prior distribution, the likelihood function for response pattern (0,1,1,0,0) evaluated at the initial values of the item parameters, and the implied posterior distribution—formed by multiplying the likelihood and the prior, point by point over the quadrature points, and then normalized to sum to one. The prior and the posterior distributions are shown as discrete probability point masses over the quadrature points. The ordinates of the normalized prior distribution have been multiplied by the observed sample size (1,490), and those of the posterior distribution have been multiplied by the observed frequency associated with the response pattern (208).

For each item, depending on the response (0 or 1), the posterior probabilities are accumulated as per Equation (3.24). For instance, item 1's response is zero, which means

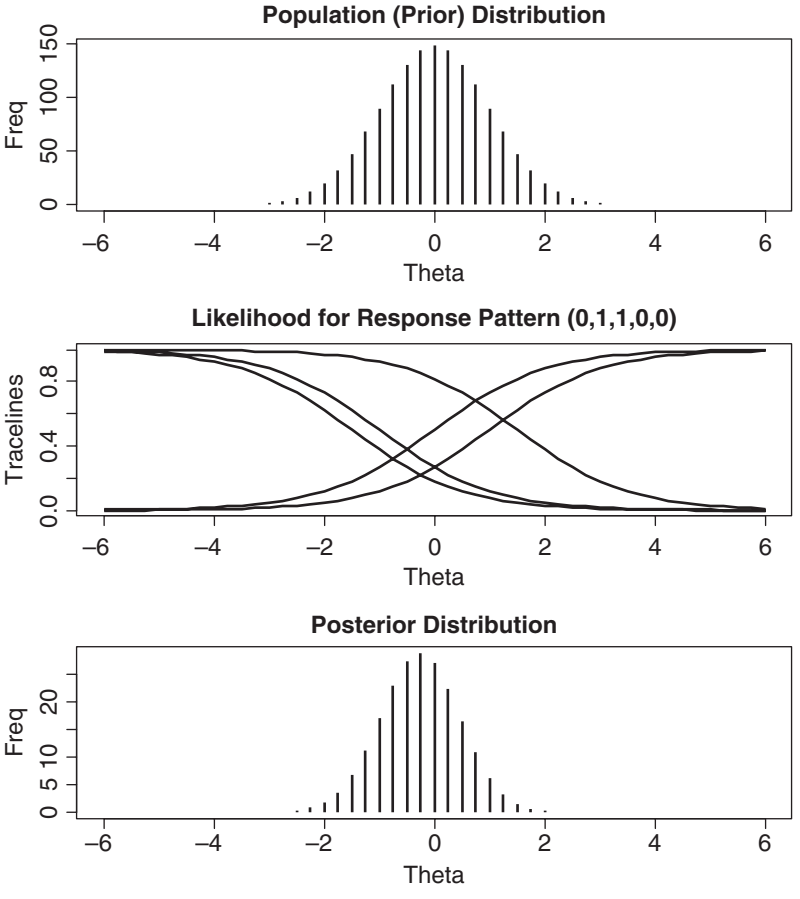


Figure 3.1 Multi-panel plot showing the relationship among the prior (population) distribution, the likelihood (products of tracelines shown) for response pattern (0,1,1,0,0), and the posterior.

Downloaded By: 10.3.98.166 At: 03:56 15 Jun 2018; For: 9781315736013, chapter3, 10.4324/9781315736013.ch3

Downloaded By: 10.3.98.166 At: 03:56 15 Jun 2018; For: 9781315736013, chapter3, 10.4324/9781315736013.ch3

that the current set of posterior probabilities must be added into the $\bar{r}_{1q}$ values over the Q quadrature points. Similarly, because item 2's response is one, for that item, the posterior probabilities are added into the r_{2q} values for all q. Regardless of the response, the posterior probabilities are added into n_{iq} for all items and quadrature points.

For each response pattern, there is a set of corresponding three-panel plots that generate the posterior probabilities over the same set of quadrature points. These posterior probabilities are accumulated into the item-specific r_{iq} and n_{iq} values, depending on the item response. At the end of the E-step, the weights n_{iq} and (artificial) response frequencies r_{iq} are submitted to the M-step for logit analyses.

Figure 3.2 presents the current and updated tracelines for item 1 after one cycle of E- and M-step. The current tracelines (dashed curves) are at their initial values of $\alpha_1 = -1.5$ and $\beta_1 = 1$. The ordinates of the solid dots are equal to r_{1q} / n_{1q}, representing the conditional expected probability of the endorsement response for item 1 at each of the quadrature points. The size of each of the solid dots is proportional to the conditional expected number of respondents at each of the corresponding quadrature points. The updated tracelines (solid curves) correspond to $\alpha_1 = -2.12$ and $\beta_1 = 1.11$. It is obvious that the updated tracelines are much closer approximations of the "data" generated by the E-step conditional expectations. Other items can be handled similarly. Thus iterating the E- and M-steps leads to a sequence of item parameter estimates that eventually converges to the MLE. At the maximum, the following item parameters are obtained: $\alpha_1 = -2.35$, $\alpha_2 = 0.80$, $\alpha_3 = 0.99$, $\alpha_4 = -0.67$, $\alpha_5 = 1.10$, $\beta_1 = 1.20$, $\beta_2 = 0.71$, $\beta_3 = 1.53$, $\beta_4 = 2.55$, $\beta_5 = 0.92$.

Let us now turn to the application of the MH-RM algorithm. The MH-RM algorithm also requires the characterization of the posterior distribution of θ, but it uses it differently than Bock-Aitkin EM: The Metropolis-Hastings sampler is used to generate dependent draws from this posterior, given provisional item parameter values and the

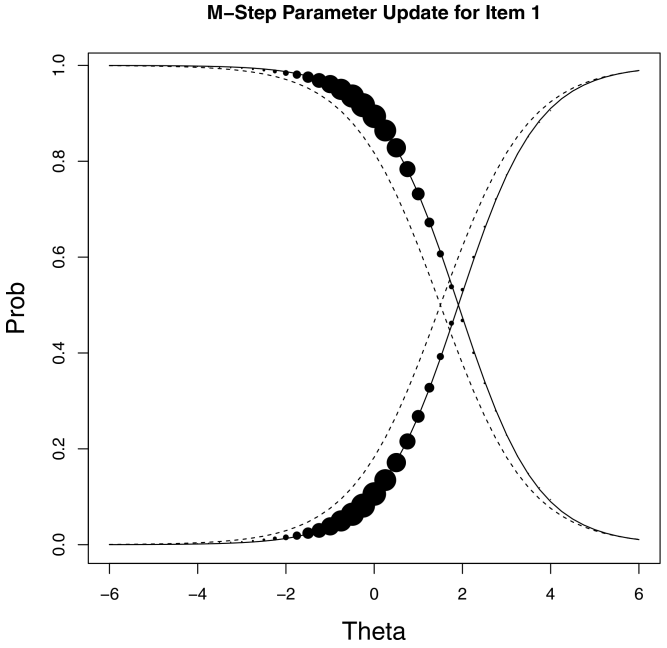


Figure 3.2 Current and updated tracelines for item 1 after one cycle of E- and M-step.

Downloaded By: 10.3.98.166 At: 03:56 15 Jun 2018; For: 9781315736013, chapter3, 10.4324/9781315736013.ch3

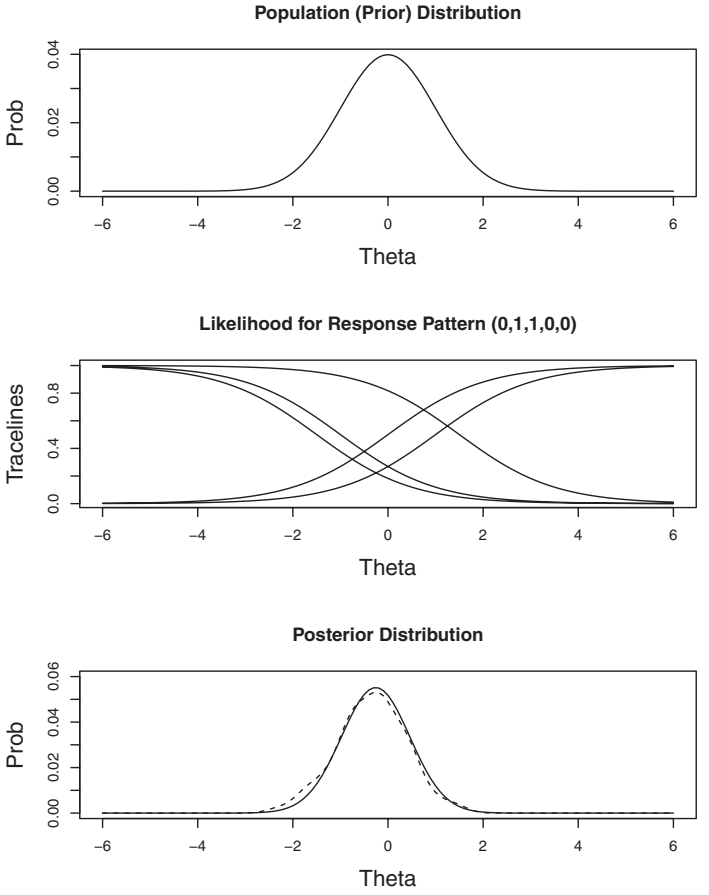


Figure 3.3 Multi-panel plot showing the relationship among the prior (population) density, the likelihood (products of tracelines shown) for response pattern (0,1,1,0,0), and the posterior density approximated in two ways. The solid posterior curve is found by numerically evaluating the normalized posterior ordinates over a range of θ values. The dashed posterior curve is found by plotting the empirically estimated density of the posterior from the random draws produced by a Metropolis-Hastings sampler for θ.

samples are used in complete data estimation with the Robbins-Monro method. Figure 3.3 plots the relationship among the prior density (standard normal), the likelihood function for response pattern (0,1,1,0,0) evaluated at the initial values of the item parameters ($\alpha_1 = -1.5, \alpha_2 = -1, \alpha_3 = 0, \alpha_4 = 1, \alpha_5 = 1.5, \beta_1 = \ldots \beta_5 = 1.0$), and the implied normalized posterior distribution. The prior and the posterior are represented as smooth solid curves.

There are 208 individuals associated with this response pattern. For each individual, we iterate the Metropolis-Hastings sampler 10 times and take the last draw as the posterior sample. We then empirically estimate a density function from the 208 posterior samples. The estimated density is shown as the dashed curve superimposed on the true implied posterior density. The two are obviously quite close, indicating the Metropolis-Hastings method can generate adequate posterior samples. For our sampler, the tuning constant (proposal dispersion σ) is equal to 2.0. The starting value of θ is equal to the standardized total score associated with response pattern (0,1,1,0,0). The total score for this response pattern is 2.0. The sample average total score over all response patterns is 2.17, and

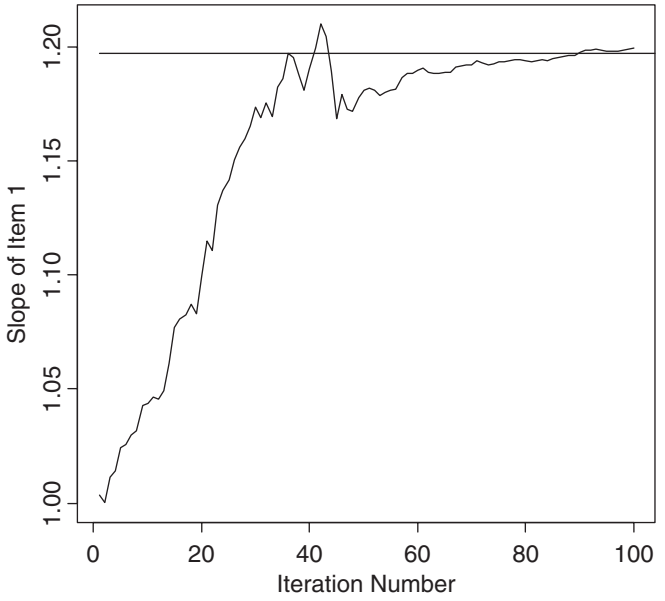


Figure 3.4 The iteration history of the slope parameter estimate for item 1 from MH-RM. The solid line is the MLE.

the sample standard deviation of the total score is 1.32, so the standardized total score is –0.13.

Together with the observed item responses, the posterior samples for all 1,490 individuals form the complete data set, with the posterior draws serving the role of predictor values. Complete data derivatives are evaluated and the item parameters are updated according to Equation (3.31) with the Robbins-Monro method. Figure 3.4 shows a sequence of parameter estimates from 100 iterations of the MH-RM algorithm for the slope parameter of item 1. The solid line is the MLE of that parameter from the Bock-Aitkin EM algorithm ($\beta_1 = 1.20$). The MH-RM estimates contain random error initially, but as the number of cycles increases, the Robbins-Monro method filters out the error to achieve convergence near the MLE.

Discussion and Conclusion

The key emphasis of our discussion of IRT and IRT parameter estimation is on a missing data formulation: The unobserved latent variable θ amounts to missing data. Had the missing data been observed, IRT parameter estimation would be standard logit analysis. Motivated by this missing data formulation, we described estimation algorithms that augment the observed data by replacing the missing data either deterministically by their posterior expected values or stochastically by multiple imputations from the posterior predictive distribution of θ. The former approach (Bock-Aitkin EM) requires numerical integration with quadrature. The latter approach (MH-RM) requires the combination of elements of Markov chain Monte Carlo (Metropolis-Hastings sampler) with stochastic approximation (Robbins-Monro method). In both approaches, it is revealed that an insight due to Fisher (1925) provided the key equation that connects the complete data and observed data models. We illustrated the estimation algorithms with an empirical data set.

Downloaded By: 10.3.98.166 At: 03:56 15 Jun 2018; For: 9781315736013, chapter3, 10.4324/9781315736013.ch3

Downloaded By: 10.3.98.166 At: 03:56 15 Jun 2018; For: 9781315736013, chapter3, 10.4324/9781315736013.ch3

This presentation has been restricted to parameter estimation for unidimensional IRT models for dichotomous responses, to keep the focus on the essential ideas. The procedures described here straightforwardly generalize to either multidimensional IRT models, or IRT models for polytomous responses, such as those used in the PROMIS® measures (Reeve et al., 2007), or both. We have alluded to the generalization to multidimensional IRT; that simply adds multidimensional quadrature grids, or vector-valued random draws, to the procedures described in the previous sections. Parameter estimation for IRT models for polytomous responses requires that the computations described in this chapter for each of the two dichotomous responses be carried out for each of the several polytomous responses, and that the values of partial derivatives be calculated for each parameter of the model. The necessary partial derivatives for most commonly used IRT models are available from a variety of sources, and are brought together in the book-length treatment of this topic by Baker and Kim (2004).

Author's Note: Li Cai is supported by grants from the Institute of Education Sciences (R305D140046 and R305B080016) and National Institute on Drug Abuse (R01DA026943 and R01DA030466). David Thissen has been supported by a PROMIS® cooperative agreement from the National Institutes of Health (NIH) Common Fund Initiative (U01AR052181). The views expressed here belong to the authors and do not reflect the views or policies of the funding agencies or grantees.

References

Baker, F.B., & Kim, S.-H. (2004). *Item response theory: Parameter estimation techniques*. New York: Dekker.

Bartholomew, D.J. (1998). Scaling unobservable constructs in social science. *Journal of the Royal Statistical Society – Series C, 47*, 1–13.

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

Bock, R.D., & Lieberman, M. (1970). Fitting a response model for *n* dichotomously scored items. *Psychometrika, 35*, 179–197.

Bolt, D. (2005). Limited and full information estimation of item response theory models. In A. Maydeu-Olivares & J.J. McArdle (Eds.). *Contemporary psychometrics* (pp. 27–71). Mahwah, NJ: Earlbaum.

Booth, J.G., & Hobert, J.P. (1999). Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *Journal of the Royal Statistical Society—Series B, 61*, 265–285.

Cai, L. (2006). *Full-information item factor analysis by Markov chain Monte Carlo stochastic approximation*. Unpublished master's thesis, Department of Statistics, University of North Carolina at Chapel Hill.

Cai, L. (2008). *A Metropolis-Hastings Robbins-Monro algorithm for maximum likelihood nonlinear latent structure analysis with a comprehensive measurement model*. Unpublished doctoral dissertation, Department of Psychology, University of North Carolina at Chapel Hill.

Cai, L. (2010a). High-dimensional exploratory item factor analysis by a Metropolis-Hastings Robbins-Monro algorithm. *Psychometrika, 75*, 33–57.

Cai, L. (2010b). Metropolis-Hastings Robbins-Monro algorithm for confirmatory item factor analysis. *Journal of Educational and Behavioral Statistics, 35*, 307–355.

Dempster, A.P., Laird, N.M., & Rubin, D.B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society—Series B, 39*, 1–38.

Edwards, M.C. (2010). A Markov chain Monte Carlo approach to confirmatory item factor analysis. *Psychometrika, 75*, 474–497.

Fisher, R.A. (1925). Theory of statistical estimation. *Proceedings of the Cambridge Philosophical Society, 22*, 700–725.

Hastings, W.K. (1970). Monte Carlo simulation methods using Markov chains and their applications. *Biometrika, 57*, 97–109.

Maydeu-Olivares, A. (2013). Goodness-of-fit assessment of item response theory models. *Measurement: Interdisciplinary Research and Perspectives, 11,* 71–101.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., & Teller, E. (1953). Equations of state space calculations by fast computing machines. *Journal of Chemical Physics, 21*, 1087–1092.

Reeve, B.B., Hays, R. D., Bjorner, J.B., Cook K.F., Crane, P.K., Teresi, J.A., Thissen, D., Revicki, D.A., Weiss, D.J., Hambleton, R.K., Liu, H., Gershon, R., Reise, S.P., Lai, J.S., & Cella, D (2007). Psychometric evaluation and calibration of health-related quality of life items banks: Plans for the patient-reported outcome measurement information system (PROMIS). *Medical Care*, *45*, S22–31.

Rijmen, F., & Jeon, M. (2013). Fitting an item response theory model with random item effects across groups by a variational approximation method. *Annals of Operations Research, 206*, 647–662.

Robbins, H., & Monro, S. (1951). A stochastic approximation method. *Annals of Mathematical Statistics, 22*, 400–407.

Roberts, G.O., & Rosenthal, J.S. (2001). Optimal scaling for various Metropolis-Hastings algorithms. *Statistical Science, 16*, 351–367.

Schilling, S., & Bock, R.D. (2005). High-dimensional maximum marginal likelihood item factor analysis by adaptive quadrature. *Psychometrika, 70*, 533–555.

Thissen, D. (1982). Marginal maximum likelihood estimation for the one-parameter logistic model. *Psychometrika, 47*, 175–186.

Wirth, R.J., & Edwards, M.C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods, 12*, 58–79.

Wu, C.F.J. (1983). On the convergence properties of the EM algorithm. *Annals of Statistics, 11*, 95–103.

Downloaded By: 10.3.98.166 At: 03:56 15 Jun 2018; For: 9781315736013, chapter3, 10.4324/9781315736013.ch3