

Real-time Bayesian parameter estimation for item response models

Ruby Chiu-Hsing Weng¹ and D. Stephen Coad²
National Chengchi University, Taipei, Taiwan¹
Queen Mary, University of London, UK²

Abstract. Bayesian item response models have been used in modelling educational testing and Internet ratings data. Typically, the statistical analysis is carried out using Markov Chain Monte Carlo methods. However, these may not be computationally feasible when real-time data continuously arrive and online parameter estimation is needed. We develop an efficient algorithm based on a deterministic moment-matching method to adjust the parameters in real-time. The proposed online algorithm works well for two real datasets, achieving good accuracy but with considerably less computational time.

Keywords: Bayesian inference; deterministic method; moment matching; online algorithm; Woodroffe-Stein's identity.

1 Introduction

Markov Chain Monte Carlo (MCMC) methods have revolutionized statistical computing in the past two decades. The developments in MCMC algorithms and software have enabled the use of Bayesian inference for item response theory (IRT) models in psychological and educational studies. For example, Johnson and Albert (1999) described Gibbs sampling methods for Normal Ogive (or Probit) IRT models; Patz and Junker (1999) developed a Metropolis-Hastings sampling method and illustrated it using the two-parameter logistic (2PL) IRT model; Fox and Glas (2001) proposed Gibbs sampling for multilevel IRT models; Ho and Quinn (2008a) fit a Bayesian ordinal item response theory (IRT) model via MCMC techniques for Internet ratings data, where the data are typically ordinal measurements on the quality of all kinds of items such as movies, consumer products, and so on; and Martin and Quinn (2002) and Wang et al. (2013) presented the MCMC strategy for dynamic item response models. For a brief survey of developments in item response modeling from a Bayesian perspective, see Albert (2015).

An MCMC method is a sampling, or nondeterministic, approach for implementing Bayesian inference. It is simple and often works well, even for complex models, but requires considerable computation. Take the ordinal IRT model in Ho and Quinn (2008a) for illustration: this model has advantages over traditional average ratings by accounting for the rater bias and item quality, but fitting the model by MCMC methods may not be feasible for large-scale real-time data. Indeed, for such a scenario, Ho and Quinn (2008a, Section 5) commented on the possible use of efficient deterministic approximation algorithms. The present paper aims to fill this gap by proposing an online deterministic algorithm for Bayesian parameter adjustment of the ordinal IRT models.

Online methods arise in several areas including statistics, machine learning, and control literature; see, for instance, Robbins and Monro (1951), Rosenblatt (1958), and Maybeck (1982). As opposed to an offline algorithm which processes the entire data at once, an online algorithm learns about one piece of data at a time and discards it after learning. On the one hand, since online algorithms do not see the entire data in advance, they often perform worse than their offline counterparts. On the other hand, online methods require less memory and have advantages when dealing with very large real-time data. In recent years, there has been a growing interest in online algorithms due to the emergence of large-scale Internet data. We refer to Saad (1998), Shalev-Shwartz (2011), and the references therein, for an overview of online methods.

Deterministic methods are useful alternatives to MCMC ones for Bayesian inference. Deterministic approaches such as variational Bayes (Bishop 2008), expectation propagation (Minka 2001), and so on, are popular machine learning methodologies. They have been widely used by computer scientists to solve large data problems. Recently, the statistical community has devoted more attention to these methods; see, for example, Rue et al. (2009), Faes et al. (2011), Hall et al. (2011), among others. Variational Bayes methods are a family of techniques that try to give a lower bound for the marginal likelihood and provide an analytical approximation to the posterior distribution of the unobserved variables. Expectation propagation is an extension to assumed-density filtering (Maybeck (1982), Boyen and Koller (1998); also called “moment matching”) and a general deterministic method that approximately minimizes the Kullback-Leibler divergence between the exact distribution p and its approximation q :

$$\text{KL}(p|q) = \int p(x) \log \left[\frac{p(x)}{q(x)} \right] dx.$$

With the constraint that q is in the Gaussian family, the solution follows from moment matching, or expectation constraints:

$$E_p(X) = E_q(X) \quad \text{and} \quad E_p(X^2) = E_q(X^2).$$

It can be viewed as a variation of the extended Kalman filter, a popular sequential method for inference in dynamic systems that approximately updates the first two moments of the state distribution. Herbrich et al. (2007) employed the expectation propagation technique to develop TrueSkillTM, the online ranking system for Xbox Live.

Recently, Weng and Lin (2011) proposed a new deterministic moment-matching approach based on a version of Stein’s identity and exact calculation of certain integrals. When applying it for online ranking of players, the accuracy is comparable to TrueSkillTM, but the running time as well as the code are much shorter. This approach has been used in other applications. For example, Wistuba et al. (2012) modified it for move prediction in Computer Go and obtained promising experimental results; Chen et al. (2013) applied it to obtain an efficient online Bayesian ranking scheme in a crowd-sourced setting. All these are within the models of ranked data.

The present paper extends the moment matching in Weng and Lin (2011) to IRT models. We begin by observing that, similar to the models for ranked data, IRT models

often model the outcomes by normal or logistic distributions, though these can be more complicated. Next, we obtain online algorithms to adjust the parameters in certain ordinal IRT models, where the parameters may be allowed to be time-varying. Then we demonstrate the effectiveness of the proposed algorithm through two real datasets. Our experiments show that the proposed real-time estimation method works well for 100,000 ratings collected over time.

The organization of the paper is as follows. In Section 2, we introduce the motivating examples, review IRT models, and comment on parameter estimation. In Section 3, we describe the approximation method. Section 4 develops online algorithms. Section 5 presents experiments on simulated and real datasets. Section 6 concludes.

2 Online product ratings and IRT models

2.1 Motivating examples

Online consumer product ratings are growing rapidly and they play an increasingly important role in consumers' purchase decisions. For example, Chevalier and Mayzlin (2006) found a positive relationship between consumer book ratings and book sales; Liu (2006) showed that both positive and negative reviews of a movie increase its box office revenue. Here we discuss the statistical analysis of online ratings. Two online ratings datasets are considered: the ratings of news outlets from Mondo Times (<http://www.mondotimes.com/>) used in Ho and Quinn (2008a), and the ratings of movies from GroupLens Research Project at the University of Minnesota (<http://movielens.umn.edu>).

Mondo Times is an online company that disseminates information from over 33,000 media outlets, including newspapers, television stations, and so on, in 213 countries. The registered members can submit five-point ratings of the content quality of news outlets from 1=awful, 2=poor, 3=average, 4=very good, to 5=great. The dataset presented in Ho and Quinn (2008a) consists of 4,511 ratings on 1,515 products (news outlets) from 946 raters, available from their Ratings package (available at <http://cran.r-project.org/>). The average number of ratings for a product is $3.0 (= 4,511/1,515)$ and the average number rated by a rater is $4.8 (= 4,511/946)$.

The movie ratings data were collected through the MovieLens web site during the seven-month period from September 19th, 1997, to April 22nd, 1998. This dataset consists of 100,000 movie ratings from 943 users on 1,682 movies. The ratings are also on a scale of 1 to 5. The average number of ratings for a product is $59.5 (= 100,000/1,682)$, the average number rated by a rater is $106.0 (= 100,000/943)$, and each user has rated at least 20 movies.

These online product ratings data are characterized by having numerous products and raters, and each rater only rates a small proportion of the products. For the Mondo Times data, the missing rate is $1 - 4,511/(946 \times 1,515) \approx 0.997$; for the MovieLens data, it is $1 - 100,000/(943 \times 1,682) \approx 0.937$. So, the rater-product matrix of ratings is

sparse. The ratings data are typically displayed by a number of stars that represent the mean rating of a product. Ho and Quinn (2008a) pointed out some potential problems with such displays. First, it does not take into account raters' rating behavior; for instance, some may be more inclined to use high or low ratings, while some may use all the rating categories. Weighting all raters equally may bias the results. Second, with just a number of stars for the mean rating, it lacks a measure of statistical uncertainty. To address these problems, Ho and Quinn (2008a) proposed some graphical displays based on an ordinal IRT model, which can incorporate statistical uncertainty in the ratings and adjust for rater-specific factors. In the next subsection, we review some IRT models that are relevant to the present paper.

2.2 The IRT models

IRT models have been widely used in modeling dichotomously and polytomously scored data from educational tests; see van der Linden and Hambleton (2013) and De Ayala (2013).

Example 1: Basic IRT. The basic one-parameter IRT model is for analyzing dichotomously scored test data. It is based on the idea that the probability of a correct response to a test item is a function of the examinee's ability and the item's difficulty. The model has the form

$$P(Y_{ij} = 1 | \theta_i, \beta_j) = F(\theta_i - \beta_j), \quad (1)$$

where the item response variable $Y_{ij} = 0$ or 1 , corresponding to whether the response to the j th item taken by the i th person is correct or not, θ_i represents the ability parameter of the i th person, β_j represents the difficulty parameter of the j th test item, and the item response function $F(\cdot)$ is a cumulative distribution function (c.d.f.) from a continuous distribution. Model (1) becomes the Rasch model (Rasch 1961) when $F(\cdot)$ is the standard logistic c.d.f., and the Normal Ogive (or Probit) model when $F(\cdot)$ is the standard normal c.d.f.

Example 2: Ordinal IRT. Samejima (1969) introduced the graded response model to analyze ordered polytomous data. Let Y_{ij} denote the score of the i th person on item j . The model specifies the probability of the i th person responding in category c or higher on item j as

$$P(Y_{ij} \geq c | \beta_j, \theta_i, \delta_{j,c}) = F(\beta_j(\theta_i - \delta_{j,c})),$$

where θ_i is the proficiency of the i th person, β_j is the discrimination parameter for item j , $\delta_{j,c}$ is the item response parameter for item j and category c , $c = \{0, 1, \dots, C_j\}$, and $F(\cdot)$ is the c.d.f. of a logistic or a normal distribution. Muraki (1990) proposed a modified graded response model suitable for Likert-type data in which the categories are $\{1, 2, \dots, C\}$ for all items. This model resolved the item response parameter $\delta_{j,c}$ into the item location parameter α_j and the category threshold parameter d_{c-1} , where $d_0 = -\infty$; that is,

$$P(Y_{ij} \geq c | \beta_j, \theta_i, \alpha_j, d_{c-1}) = F(\beta_j(\theta_i + \alpha_j - d_{c-1})). \quad (2)$$

Ho and Quinn (2008a) proposed the following ordinal IRT model to fit online product ratings data. To begin, let $Y_{ij} \in \{1, 2, \dots, C\}$ denote the rating of product i by rater j . They assume that the observed Y_{ij} is determined by an unobserved variable Y_{ij}^* :

$$Y_{ij} = c \Leftrightarrow Y_{ij}^* \in (\gamma_{c-1}, \gamma_c], \quad (3)$$

where the γ_c are cutpoints, $\gamma_0 = -\infty$, $\gamma_C = \infty$, and Y_{ij}^* is parameterized as

$$Y_{ij}^* = \alpha_j + \beta_j \theta_i + \epsilon_{ij}, \quad \epsilon_{ij} \stackrel{iid}{\sim} N(0, 1). \quad (4)$$

In (4), α_j captures the center location of rater j 's rating, β_j is assumed positive (to identify the sign of θ_i) and it measures how well rater j discriminates between low and high quality, and θ_i represents the latent quality of product i . By (3), the probability of observing Y_{ij} in category c or higher is

$$P(Y_{ij} \geq c | \alpha_j, \beta_j, \theta_i, \gamma_{c-1}) = \Phi(\beta_j \theta_i + \alpha_j - \gamma_{c-1}). \quad (5)$$

Note that ϵ_{ij} in (4) follows either the normal or the logistic distribution. We observe that model (5) is a variation of the graded response model and that it closely resembles (2), but they differ in the parameterization of the item location and category threshold parameters.

Example 3: Dynamic IRT. Many authors have proposed the dynamic IRT model to capture the changes in a person's ability over time; for example, Embretson (1991), Martin and Quinn (2002), Wang et al. (2013), among others. The simplest case has the form:

$$\text{System equation:} \quad \theta_{i,t} = \theta_{i,t-1} + w_{i,t}, \quad (6)$$

$$\text{Observation equation:} \quad P(Y_{i,j,t} = 1 | \theta_{i,t}, \beta_j) = \frac{\exp(\theta_{i,t} - \beta_j)}{1 + \exp(\theta_{i,t} - \beta_j)}, \quad (7)$$

where $\theta_{i,t}$ is the current ability of the i th person, $\theta_{i,t-1}$ is the ability at the previous time point, $w_{i,t}$ represents the random change in ability as in all linear dynamic models, and $Y_{i,j,t} = 1$ if the i th person answered correctly on day t for test item j .

For fitting a modified graded response model to product ratings data, the temporal changes in θ_i represent the evolution of product perception and popularity. Moreover, it is sensible that the raters' inclinations (α_j, β_j) may change gradually. For example, see Koren et al. (2009) for modeling temporal effects for both products and raters in the context of movie ratings. We shall consider models with time-varying coefficients in Section 4.2.

2.3 Parameter estimation

There are maximum likelihood procedures for item parameters in IRT models; for example, joint maximum likelihood, conditional maximum likelihood, and marginal maximum likelihood. For a detailed account of these methods, see Bock and Aitkin (1981),

Molenaar (1995), and the references therein. Albert (1992) and Kim (2001) found that the IRT parameter estimates using the maximum likelihood methods and the Gibbs sampling method with non-informative priors are similar.

Although the maximum likelihood methods for IRT models could give good parameter estimates in many cases, to the best of our knowledge, these estimation procedures under large sparse data matrices have not been fully considered in the literature. In fact, there could be some potential problems. First, the maximum likelihood estimate may not be unique. For the Mondo Times data, many raters give very few ratings and many items received very few ratings. Specifically, among the 946 raters, 410 rate just one item, and 138 rate exactly two items; among 1,515 items, 777 received just one rating, and 326 received just 2 ratings. The histograms for “number of items rated by a rater” and “number of ratings an item receives” are given in Figure 1. However, for the ordinal IRT model (4) in Section 2.2, each rater is associated with an intercept parameter and a slope parameter; from (4), it is not difficult to see that these two parameters can not be uniquely determined if a rater only rates one item. Second, obtaining the standard error estimates for the model parameters involves solving the inverse of a large Hessian matrix of the log-likelihood. For the Mondo Times data, the matrix is about $3,500 \times 3,500$. Even if the inverse exists, the computation may be burdensome, especially in an online setting. Ho and Quinn (2008a) proposed some graphical displays based on an ordinal IRT model, which can incorporate statistical uncertainty in the ratings and adjust for rater-specific factors. Their MCMC approach fits well for offline data; however, it may not be viable for large-scale real-time data.

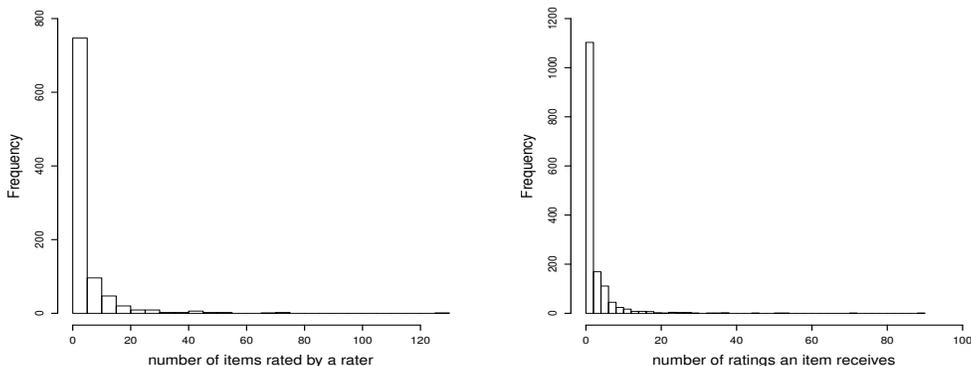


Figure 1: Histograms.

3 The approximation method

In what follows, let ϕ and Φ denote the density and distribution function of a standard normal variable, and let $\phi(x|\mu, \sigma)$ denote the density of the normal distribution with mean μ and standard deviation σ .

3.1 Moment equations

The proposed method is based on the moment equations (9) and (10) obtained from a version of Stein's identity. Stein's lemma (Stein 1981) concerns the expectation of a normally distributed random variable. The lemma is famous and of interest primarily because of its applications to the James-Stein estimator (James and Stein 1961) and to empirical Bayes methods.

In the context of deriving sequential confidence levels, Woodroffe (1989) studied integrable expansions for posterior distributions and developed a variant of Stein's identity. It concerns the expectation with respect to a distribution of the form $p(\mathbf{z}) = \phi(\mathbf{z})f(\mathbf{z})$, where f is defined on R^p . The degree of smoothness of f governs the order of the expansion. A motivation for considering such a distribution comes from large-sample theory: many suitably normalized quantities are asymptotically normally distributed and may be written in this form; therefore, by studying expectations with respect to it, one can possibly refine the normal approximation. Weng and Lin (2011) referred to this identity as Woodroffe-Stein's identity to distinguish it from Stein's lemma, and used the identity to obtain a Bayesian approximate moment-matching method.

As the identity involves complex notation, here we describe only some results necessary for the proposed online method. For a detailed account of the identity, we refer readers to Woodroffe and Coad (1997, Proposition 2) and Weng and Lin (2011, Corollary 2); see also Weng (2010, 2015) for some extensions. Let $\boldsymbol{\psi}^* = (\psi_1^*, \dots, \psi_p^*)'$ be a vector of parameters whose posterior density takes the form

$$C\phi(\boldsymbol{\psi}^*)f(\boldsymbol{\psi}^*), \quad (8)$$

where the dependence on the data is suppressed in the notation and

$$C = \left(\int \phi(\boldsymbol{\psi}^*)f(\boldsymbol{\psi}^*)d\boldsymbol{\psi}^* \right)^{-1},$$

the normalizing constant. Further suppose that f is a twice continuously differentiable function. An application of Woodroffe-Stein's identity gives the following:

$$E(\boldsymbol{\psi}^*) = E\left(\frac{\nabla f(\boldsymbol{\psi}^*)}{f(\boldsymbol{\psi}^*)}\right), \quad (9)$$

$$E(\psi_i^* \psi_q^*) = \delta_{iq} + E\left[\frac{\nabla^2 f(\boldsymbol{\psi}^*)}{f(\boldsymbol{\psi}^*)}\right]_{iq}, \quad i, q = 1, \dots, k, \quad (10)$$

provided the expectations on both sides of (9) and (10) exist, where $\delta_{iq} = 1$ if $i = q$ and 0 otherwise, and $[\cdot]_{iq}$ indicates the (i, q) entry of a matrix.

From (9) and (10), the mean and variance of ψ_i^* are

$$E(\psi_i^*) = E\left(\frac{\partial \log f(\boldsymbol{\psi}^*)}{\partial \psi_i^*}\right), \quad (11)$$

$$Var(\psi_i^*) = E((\psi_i^*)^2) - E(\psi_i^*)^2 = 1 + E\left[\frac{\partial^2 f / \partial (\psi_i^*)^2}{f}\right] - \left[E\left(\frac{\partial \log f}{\partial \psi_i^*}\right)\right]^2. \quad (12)$$

On the right-hand side of (9)-(12), we note that the terms inside the brackets do not involve the normalizing constant C , which is often intractable.

3.2 A Bayesian moment-matching scheme

In the context of inferring individual skills from group comparisons, Weng and Lin (2011) proposed a two-step method where the first step approximates team skills and the second step relates the individual skill update to the team skill update. Both steps make use of the moment equations (9) and (10). Since the present paper shall exploit and extend the first step, below we briefly review this step. To begin, consider the paired comparison case where the strength parameter of player i is denoted as ψ_i . Assume that ψ_i follows $N(\mu_i, \sigma_i^2)$, as in online rating systems such as Glicko (Glickman 1999) and TrueSkill (Herbrich et al. 2007). Next, upon observing the game outcome D (i beats j), update the skill as $N(\mu_i^{\text{new}}, (\sigma_i^2)^{\text{new}})$, where μ_i^{new} and $(\sigma_i^2)^{\text{new}}$ are the posterior mean and variance of ψ_i derived from the joint posterior distribution of (ψ_i, ψ_j) . Then the μ_i^{new} and $(\sigma_i^2)^{\text{new}}$ are viewed as the prior information for the next game, and the updating procedure repeats. The paired comparison data are often modeled as

$$P(i \text{ beats } j) = P(X_i - X_j > 0), \quad (13)$$

where the X_i are unobserved actual performance. When $X_i - X_j$ follows the normal distribution, (13) corresponds to the Thurstone-Mosteller model (Thurstone 1927), whereas, when $X_i - X_j$ follows the logistic distribution, (13) becomes the Bradley-Terry model.

Now assume that the actual performance X_i follows $N(\psi_i, \eta_i^2)$ with $\psi_i \sim N(\mu_i, \sigma_i^2)$. So, η_i represents the uncertainty in the actual performance, while σ_i represents the prior uncertainty in ψ_i . Let $\boldsymbol{\mu} = (\mu_i, \mu_j)'$, $\boldsymbol{\sigma} = (\sigma_i, \sigma_j)'$, and $\boldsymbol{\psi}^* = (\psi_i^*, \psi_j^*)'$ with

$$\psi_i^* = (\psi_i - \mu_i)/\sigma_i. \quad (14)$$

Then the posterior density of $\boldsymbol{\psi}$ given the game outcome D (i beats j) is

$$p(\boldsymbol{\psi}|D) \propto \phi(\boldsymbol{\psi})f(\boldsymbol{\psi}^*) = \phi(\boldsymbol{\psi})\Phi\left(\frac{\psi_i - \psi_j}{\bar{\eta}}\right), \quad (15)$$

where $\bar{\eta} = (\eta_i^2 + \eta_j^2)^{1/2}$. By (11), (14) and the chain rule, we have

$$E(\psi_i|D) = \mu_i + E\left(\sigma_i \frac{\partial}{\partial \psi_i^*} \log f(\boldsymbol{\psi}^*) \Big| D\right). \quad (16)$$

Weng and Lin (2011, Section 3.2) proceeded to show that one may evaluate the expectation on the right-hand side of (16) at $\boldsymbol{\psi}^* = 0$ and then correct the approximation by a scaling factor; that is, by letting

$$g(\boldsymbol{\mu}, \boldsymbol{\sigma}, \bar{\eta}) = \sigma_i \frac{\partial}{\partial \psi_i^*} \log f(\boldsymbol{\psi}^*) \Big|_{\boldsymbol{\psi}^*=0} = \sigma_i^2 \frac{\partial}{\partial \psi_i} \log f(\boldsymbol{\psi}^*) \Big|_{\boldsymbol{\psi}=\boldsymbol{\mu}}, \quad (17)$$

it is shown that

$$E(\psi_i|D) = \mu_i + g(\boldsymbol{\mu}, \boldsymbol{\sigma}, c\bar{\eta}), \quad (18)$$

where c is a scaling factor for which

$$c\bar{\eta} = (\bar{\eta}^2 + \sigma_i^2 + \sigma_j^2)^{1/2}. \quad (19)$$

The posterior variance can be expressed similarly: first, (12), (14) and the chain rule give

$$\text{Var}(\psi_i|D) = \sigma_i^2 \left[1 + E \left(\frac{\partial^2 f / \partial (\psi_i^*)^2}{f(\boldsymbol{\psi}^*)} \middle| D \right) - \left(E \left(\frac{\partial \log f(\boldsymbol{\psi}^*)}{\partial \psi_i^*} \middle| D \right) \right)^2 \right]; \quad (20)$$

then, letting

$$\begin{aligned} h(\boldsymbol{\mu}, \boldsymbol{\sigma}, \bar{\eta}) &= - \left[\left(\frac{\partial^2 f / \partial (\psi_i^*)^2}{f(\boldsymbol{\psi}^*)} \middle|_{\boldsymbol{\psi}^* = \boldsymbol{\mu}} \right) - \left(\frac{\partial \log f(\boldsymbol{\psi}^*)}{\partial \psi_i^*} \middle|_{\boldsymbol{\psi}^* = \boldsymbol{\mu}} \right)^2 \right] \\ &= - \frac{\partial^2}{\partial (\psi_i^*)^2} \log f(\boldsymbol{\psi}^*) \middle|_{\boldsymbol{\psi}^* = \boldsymbol{\mu}} = - \sigma_i^2 \frac{\partial^2}{\partial \psi_i^2} \log f(\boldsymbol{\psi}^*) \middle|_{\boldsymbol{\psi} = \boldsymbol{\mu}}, \end{aligned} \quad (21)$$

one can show that

$$\text{Var}(\psi_i|D) = \sigma_i^2 (1 - h(\boldsymbol{\mu}, \boldsymbol{\sigma}, c\bar{\eta})). \quad (22)$$

The idea of using a scaling factor to improve an estimate is not new; see, for instance, Spiegelhalter and Lauritzen (1990, Section 4.3) and MacKay (1992, Section 2.1) in the context of logistic regression.

One may incorporate the time-varying strengths into the approximation technique described above. The very idea of modeling time-varying strengths for players of games has appeared in Glickman (1999), Glickman and Stern (1998), Dangauthier et al. (2008), and the references therein. Specifically, assume that individual i 's strength has been updated from outcomes at time $t-1$ (measurement update), and the strength is distributed as $\psi_{i,t-1} \sim N(\mu_i, \sigma_i^2)$; and assume that the ability follows a Gaussian drift $\psi_{i,t} = \psi_{i,t-1} + w_{i,t}$, where $w_{i,t} \sim N(0, \nu^2)$ for some $\nu > 0$. Then integrating the distribution of $\psi_{i,t} | \psi_{i,t-1}$ with respect to the prior distribution of $\psi_{i,t-1}$ gives $\psi_{i,t} \sim N(\mu_i, \sigma_i^2 + \nu^2)$ (time update), which is considered as the prior information for the test taken at time t .

4 Online inference of IRT models for Likert-type data

For the basic one-parameter IRT models (1), if the item response function F is taken as the standard logistic or normal c.d.f., then it is not difficult to derive real-time parameter adjustment by modifying Algorithms 1 and 3 in Weng and Lin (2011).

Our primary interest here are online inference for models (2) and (5), designed for Likert-type data. We assume that each α_j follows $N(\mu_{\alpha_j}, \sigma_{\alpha_j}^2)$, each β_j follows

$N(\mu_{\beta_j}, \sigma_{\beta_j}^2)$, and each θ_i follows $N(\mu_{\theta_i}, \sigma_{\theta_i}^2)$ with all parameters mutually independent. Now define the normalized quantities

$$\alpha_j^* = \frac{\alpha_j - \mu_{\alpha_j}}{\sigma_{\alpha_j}}, \quad \beta_j^* = \frac{\beta_j - \mu_{\beta_j}}{\sigma_{\beta_j}}, \quad \theta_i^* = \frac{\theta_i - \mu_{\theta_i}}{\sigma_{\theta_i}}. \quad (23)$$

The posterior distribution of $(\alpha_j^*, \beta_j^*, \theta_i^*)$ given $y_{ij} = c$ is

$$p(\alpha_j^*, \beta_j^*, \theta_i^* | y_{ij}) \propto \phi(\alpha_j^*, \beta_j^*, \theta_i^*) f(\alpha_j^*, \beta_j^*, \theta_i^*), \quad (24)$$

where f is the likelihood based on data $y_{ij} = c$. Note that (24) is of the form (8). Therefore, one can apply (11), (12) and (23) to derive expressions similar to (18) and (22) for $\alpha_j, \beta_j, \theta_i$.

4.1 Sequential update

This subsection presents the sequential update rule when f in (24) is from (5) and comments on how similar procedures can be applied when f is as (2).

For model (5), the posterior distribution of $(\alpha_j^*, \beta_j^*, \theta_i^*)$ given $y_{ij} = c$ is (24) with

$$f(\alpha_j^*, \beta_j^*, \theta_i^*) = \Phi(\beta_j \theta_i + \alpha_j - \gamma_{c-1}) - \Phi(\beta_j \theta_i + \alpha_j - \gamma_c). \quad (25)$$

The above line resembles f in (15), yet $\bar{\eta}$ in (15) is 1 here. The proposed estimates for the posterior means and variances are

$$\tilde{\mu}_{\alpha_j} = \mu_{\alpha_j} + \left(\frac{\sigma_{\alpha_j}^2}{\nu} \right) \cdot \Omega \left(\frac{\mu_x}{\nu}, \frac{a}{\nu} \right), \quad \tilde{\sigma}_{\alpha_j}^2 = \sigma_{\alpha_j}^2 \left\{ 1 - \left(\frac{\sigma_{\alpha_j}}{\nu} \right)^2 \cdot \Delta \left(\frac{\mu_x}{\nu}, \frac{a}{\nu} \right) \right\}, \quad (26)$$

$$\tilde{\mu}_{\beta_j} = \mu_{\beta_j} + \left(\frac{\sigma_{\beta_j}^2 \mu_{\theta_i}}{\nu} \right) \cdot \Omega \left(\frac{\mu_x}{\nu}, \frac{a}{\nu} \right), \quad \tilde{\sigma}_{\beta_j}^2 = \sigma_{\beta_j}^2 \left\{ 1 - \left(\frac{\sigma_{\beta_j} \mu_{\theta_i}}{\nu} \right)^2 \cdot \Delta \left(\frac{\mu_x}{\nu}, \frac{a}{\nu} \right) \right\}, \quad (27)$$

$$\tilde{\mu}_{\theta_i} = \mu_{\theta_i} + \left(\frac{\sigma_{\theta_i}^2 \mu_{\beta_j}}{\nu} \right) \cdot \Omega \left(\frac{\mu_x}{\nu}, \frac{a}{\nu} \right), \quad \tilde{\sigma}_{\theta_i}^2 = \sigma_{\theta_i}^2 \left\{ 1 - \left(\frac{\sigma_{\theta_i} \mu_{\beta_j}}{\nu} \right)^2 \cdot \Delta \left(\frac{\mu_x}{\nu}, \frac{a}{\nu} \right) \right\}, \quad (28)$$

where

$$\mu_x = \mu_{\beta_j} \mu_{\theta_i} + \mu_{\alpha_j} - \gamma_{c-1} \quad \text{and} \quad a = \gamma_c - \gamma_{c-1}, \quad (29)$$

$$\Omega(\mu_x, a) = \frac{\phi(\mu_x) - \phi(\mu_x - a)}{\Phi(\mu_x) - \Phi(\mu_x - a)}, \quad (30)$$

$$\Delta(\mu_x, a) = \frac{\mu_x \phi(\mu_x) - (\mu_x - a) \phi(\mu_x - a)}{\Phi(\mu_x) - \Phi(\mu_x - a)} + \left(\frac{\phi(\mu_x) - \phi(\mu_x - a)}{\Phi(\mu_x) - \Phi(\mu_x - a)} \right)^2,$$

$$\nu = \sqrt{1 + \sigma_{\alpha_j}^2 + \sigma_{\beta_j}^2 \mu_{\theta_i}^2 + \sigma_{\theta_i}^2 \mu_{\beta_j}^2}. \quad (31)$$

The scalar ν plays the same role as $c\bar{\eta}$ in (19), and is derived by approximate calculation of a triple integral. The detailed derivations of the above equations are relegated to

Appendix 1. Since the variance approximations in (26)-(28) may be negative, in the algorithm below we set a small positive lower bound κ in (38)-(40) to avoid a negative variance.

The proposed algorithm is described in Algorithm 1. Clearly, from (37), if a rater r is not discriminating (so $\mu_{\beta_j} \approx 0$), then his/her rating has little effect on μ_{θ_i} ; similarly, from (36), if $\mu_{\theta_i} \approx 0$, then a rating on it has little effect on μ_{β_j} .

For the unknown cutpoints $\gamma = (\gamma_1, \dots, \gamma_{C-1})'$, we propose to estimate them through the distribution of y_{ij}^* in (4) and the relation $\{y_{ij} = c\} \Leftrightarrow \{y_{ij}^* \in (\gamma_{c-1}, \gamma_c]\}$ in (3); the details are in Section 5.2.

Algorithm 1 Online parameter update for ordinal IRT model (5)

1. Given $\kappa > 0$. Given current estimates $\mu_{\alpha_j}^{(t)}$, $\mu_{\beta_j}^{(t)}$, $\mu_{\theta_i}^{(t)}$, $\sigma_{\alpha_j}^{(t)}$, $\sigma_{\beta_j}^{(t)}$, $\sigma_{\theta_i}^{(t)}$, where $i = 1, \dots, P$ and $j = 1, \dots, R$.
2. Given the $(t+1)$ st observation $y_{ij} = c$. Calculate

$$\nu^{(t)} = \sqrt{1 + (\sigma_{\alpha_j}^{(t)})^2 + (\sigma_{\beta_j}^{(t)})^2 (\mu_{\theta_i}^{(t)})^2 + (\sigma_{\theta_i}^{(t)})^2 (\mu_{\beta_j}^{(t)})^2}, \quad (32)$$

$$\omega^{(t)} = \frac{1}{\nu^{(t)}} \Omega(x^*, a^*), \quad (33)$$

$$\delta^{(t)} = \frac{1}{(\nu^{(t)})^2} \Delta(x^*, a^*), \quad (34)$$

where $\Omega(\cdot, \cdot)$ and $\Delta(\cdot, \cdot)$ are defined in (30), and

$$a^* = \frac{\gamma_c - \gamma_{c-1}}{\nu^{(t)}} \quad \text{and} \quad x^* = \frac{\mu_{\beta_j}^{(t)} \mu_{\theta_i}^{(t)} + \mu_{\alpha_j}^{(t)} - \gamma_{c-1}}{\nu^{(t)}}.$$

3. Update parameters as follows:

$$\mu_{\alpha_j}^{(t+1)} = \mu_{\alpha_j}^{(t)} + (\sigma_{\alpha_j}^{(t)})^2 \omega^{(t)}, \quad (35)$$

$$\mu_{\beta_j}^{(t+1)} = \mu_{\beta_j}^{(t)} + (\sigma_{\beta_j}^{(t)})^2 \mu_{\theta_i}^{(t)} \omega^{(t)}, \quad (36)$$

$$\mu_{\theta_i}^{(t+1)} = \mu_{\theta_i}^{(t)} + (\sigma_{\theta_i}^{(t)})^2 \mu_{\beta_j}^{(t)} \omega^{(t)}, \quad (37)$$

$$(\sigma_{\alpha_j}^{(t+1)})^2 = (\sigma_{\alpha_j}^{(t)})^2 \max\left(1 - (\sigma_{\alpha_j}^{(t)})^2 \delta^{(t)}, \kappa\right), \quad (38)$$

$$(\sigma_{\beta_j}^{(t+1)})^2 = (\sigma_{\beta_j}^{(t)})^2 \max\left(1 - (\sigma_{\beta_j}^{(t)} \mu_{\theta_i}^{(t)})^2 \delta^{(t)}, \kappa\right), \quad (39)$$

$$(\sigma_{\theta_i}^{(t+1)})^2 = (\sigma_{\theta_i}^{(t)})^2 \max\left(1 - (\sigma_{\theta_i}^{(t)} \mu_{\beta_j}^{(t)})^2 \delta^{(t)}, \kappa\right). \quad (40)$$

For the modified graded response model (2), the posterior distribution of $(\alpha_j^*, \beta_j^*, \theta_i^*)$ given $y_{ij} = c$ is (24) with

$$f(\alpha_j^*, \beta_j^*, \theta_i^*) = \Phi(\beta_j(\theta_i + \alpha_j - d_{c-1})) - \Phi(\beta_j(\theta_i + \alpha_j - d_c)). \quad (41)$$

The parameter update can be derived similarly. To see how, first note that an observed Y_{ij} from model (2) can be determined by an unobserved variable

$$Y_{ij}^\dagger = \beta_j(\alpha_j + \theta_i) + \epsilon_{ij}, \quad \epsilon_{ij} \stackrel{iid}{\sim} \mathcal{N}(0, 1) \quad (42)$$

and the relation $Y_{ij} = c \Leftrightarrow Y_{ij}^\dagger \in (\beta_j d_{c-1}, \beta_j d_c]$; or, equivalently, $Y_{ij} = c \Leftrightarrow Y_{ij}^\dagger / \beta_j \in (d_{c-1}, d_c]$. Therefore, by approximating the distribution of Y_{ij}^\dagger / β_j , we can estimate the threshold parameters (d_1, \dots, d_{C-1}) in the same manner as for γ . The sequential update for the posterior means and variances of α_j , β_j , and θ_i can be derived analogously. We omit the details, but remark that the property that rating on an item with $\mu_{\theta_i} \approx 0$ has little effect on μ_{β_j} does not hold for model (2). The reason is that β_j in (2) interacts with not only θ_i , but also α_j and d_{c-1} .

4.2 Time-varying parameters

At the end of Section 2.2, we reviewed the dynamic IRT model. In particular, we commented that both the product perception (θ) and the raters' inclinations (α, β) may evolve over time in the product ratings scenario. For the ordinal IRT models (2) and (5), one may simply extend (6) to all the variables α , β , and θ :

$$\begin{aligned} \alpha_{j,t} &= \alpha_{j,t-1} + w_{\alpha_j,t}, \\ \beta_{j,t} &= \beta_{j,t-1} + w_{\beta_j,t}, \\ \theta_{i,t} &= \theta_{i,t-1} + w_{\theta_i,t}, \end{aligned} \quad (43)$$

where $w_{\alpha_j,t}$, $w_{\beta_j,t}$, and $w_{\theta_i,t}$ are assumed to be normally distributed with unknown variances σ^2 . The unknown σ^2 can be estimated through maximizing the marginal likelihood (also called ‘‘evidence’’) over a series of N observations. This approach has been proposed by Jazwinski (1969) in the context of the Kalman filter and extended Kalman filter, and by de Freitas et al. (2000) in the context of neural networks, among others.

For the stationary ordinal IRT model (5), the marginal likelihood of an observation y_{ij} is

$$\begin{aligned} p(y_{ij}) &= \int \int \int p(\alpha_j, \beta_j, \theta_i, y_{ij}) d\alpha_j d\beta_j d\theta_i \\ &\approx \Phi\left(\frac{\mu_{\theta_i} \mu_{\beta_j} + \mu_{\alpha_j} - \gamma_{c-1}}{\nu}\right) - \Phi\left(\frac{\mu_{\theta_i} \mu_{\beta_j} + \mu_{\alpha_j} - \gamma_c}{\nu}\right), \end{aligned}$$

where the approximation follows by (57) in Appendix 1, and $\nu^2 = 1 + \sigma_{\alpha_j}^2 + \sigma_{\beta_j}^2 \mu_{\theta_i}^2 + \sigma_{\theta_i}^2 \mu_{\beta_j}^2$. For the dynamic case, the marginal likelihood is the same except that the prior

variances for α_j , β_j , and θ_i are now $\bar{\sigma}_{\alpha_j}^2 = \sigma_{\alpha_j}^2 + \sigma^2$, $\bar{\sigma}_{\beta_j}^2 = \sigma_{\beta_j}^2 + \sigma^2$, and $\bar{\sigma}_{\theta_i}^2 = \sigma_{\theta_i}^2 + \sigma^2$, respectively; and hence

$$p(y_{ij}|\sigma) \approx \Phi\left(\frac{\mu_{\theta_i}\mu_{\beta_j} + \mu_{\alpha_j} - \gamma_{c-1}}{\bar{\nu}}\right) - \Phi\left(\frac{\mu_{\theta_i}\mu_{\beta_j} + \mu_{\alpha_j} - \gamma_c}{\bar{\nu}}\right), \quad (44)$$

where

$$\bar{\nu}^2 = 1 + \bar{\sigma}_{\alpha_j}^2 + \bar{\sigma}_{\beta_j}^2\mu_{\theta_i}^2 + \bar{\sigma}_{\theta_i}^2\mu_{\beta_j}^2. \quad (45)$$

The unknown state noise parameter σ^2 can either be chosen by users or be estimated by maximizing the sum of the log-marginal over N ratings:

$$\sum_{t=1}^N \log p(y_{ij}^{(t)}|\sigma). \quad (46)$$

Once σ^2 is determined, the online parameter update is the same as Algorithm 1 except that $\nu^{(t)}$ in (32) is now replaced by $\bar{\nu}$ in (45). The marginal likelihood for model (2) and the update of its time-varying parameters can be done similarly; we omit the details.

5 Experiments

We conduct experiments to assess the performance of Algorithm 1 on simulated data and two Internet ratings datasets.

5.1 Numerical issues

There are some numerical issues in the implementation. Note that the functions Ω and Δ in (30) play important roles in the algorithm. If both the numerator and the denominator of some term on the right-hand side of (30) approach zero, it may cause computational difficulties. To have a closer look at these functions, we let

$$\Lambda(x, a) = \frac{x\phi(x) - (x-a)\phi(x-a)}{\Phi(x) - \Phi(x-a)}$$

and rewrite $\Delta(x, a)$ as

$$\Delta(x, a) = \Lambda(x, a) + (\Omega(x, a))^2.$$

Figure 2 presents plots of $\Omega(x, a)$, $\Lambda(x, a)$, and $\Delta(x, a)$ with $a = 5$ and x ranges from -50 to 50. Due to numerical difficulties, the y -values are not available when x is less than about -35 or is more than about 8. Even so, these plots suggest that, for a fixed value of a , $\Lambda(x, a)$ and $\Delta(x, a)$ are symmetric around a certain x value, and the right half of Ω can be obtained by a 180-degree rotation of the left half. Lemma 1 below gives some properties of these functions. The proofs are given in Appendix 2.

Lemma 1 *We have the following results: for each fixed a ,*

(a) $\Lambda(x, a)$ is symmetric about $x = a/2$, that is, $\Lambda\left(\frac{a}{2} + z, a\right) = \Lambda\left(\frac{a}{2} - z, a\right)$ for all z .

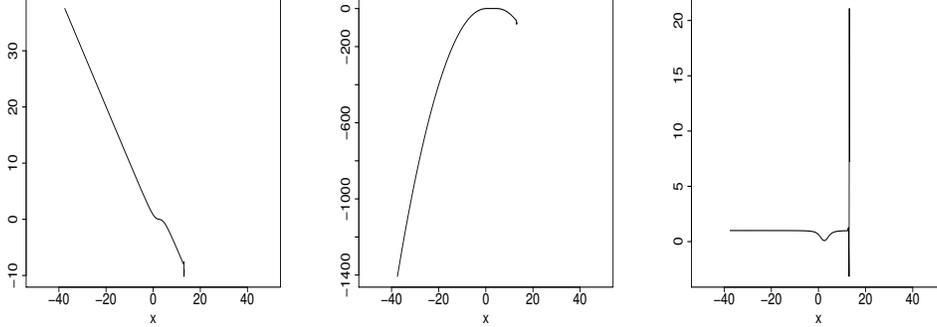


Figure 2: Left: $\Omega(x, a)$; Center: $\Lambda(x, a)$; Right: $\Delta(x, a)$; $a = 5$

- (b) $\Omega(\frac{a}{2} + z, a) = -\Omega(\frac{a}{2} - z, a)$ for all z .
(c) $\Delta(x, a)$ is symmetric about $x = a/2$.
(d) $\Omega(x, a) \approx -x$ when $|x|$ is large enough.

Therefore, in our implementation, we employ the symmetric properties and give a bound on $\Omega(x, a)$ when $|x|$ is too large to avoid numerical difficulties.

5.2 Simulation

We assume that each α_j follows $N(1, 1)$, each θ_i follows $N(0, 1)$, and each β_j follows $N(1, 20)$, with all parameters mutually independent. This corresponds to setting the initial parameter values in Algorithm 1 as $\mu_{\alpha_j}^{(0)} = 1$, $\mu_{\beta_j}^{(0)} = 1$, $\mu_{\theta_i}^{(0)} = 0$, $\sigma_{\alpha_j}^{(0)} = 1$, $\sigma_{\beta_j}^{(0)} = \sqrt{20}$, $\sigma_{\theta_i}^{(0)} = 1$ for $i = 1, \dots, I$ and $j = 1, \dots, J$. Note that these priors are exactly the same as in Ho and Quinn (2008a), except that their prior for β_j is $N(-5, 20)$ truncated to be positive, while our β_j prior has a positive mean but without truncation. Our reason for not restricting β_j is given below. The purpose of constraining $\beta_j \in \mathbb{R}^+$ for all r is to identify the sign of θ_i ; however, this may result in misleadingly assigning β_j to be near zero for some raters. In fact, it is appropriate not to restrict the parameter β_j because the sign identifiability problem induced can be resolved by constraining a particular θ_i to be positive; for example, see Ho and Quinn (2008b).

For the γ_i values, recall that $\gamma_0 = -\infty$ and $\gamma_5 = \infty$ in (3). Ho and Quinn (2008a) proposed to constrain $\gamma_1 = 0$ and set improper uniform priors for the unknown cutpoints γ_i , $i = 2, 3, 4$. We propose to use relation (3) and set them by the following steps. First, calculate the observed proportions $\#\{y_{ij} = c\}/N$ for $c = 1, \dots, 5$, where N is a pre-specified number of ratings. Next, find the z -scores corresponding to these proportions. Finally, convert the z -scores to the y^* scale approximately.

		$y_{ij} = 1$	$y_{ij} = 2$	$y_{ij} = 3$	$y_{ij} = 4$	$y_{ij} = 5$
$E(\alpha_j)$	(26)	0.731	0.897	0.979	1.072	1.253
$E(\alpha_j)$	numerical integ.	0.666	0.623	0.901	1.292	1.344
$Var(\alpha_j)$	(26)	0.967	0.957	0.957	0.957	0.968
$Var(\alpha_j)$	numerical integ.	1.023	0.936	0.797	0.869	1.021

Table 1: Approximations by (26) and by numerical integration.

Permutation	1	2	3	4	5	6	7	8	9	10
MSE	0.16	0.03	0.01	0.01	0.1	0.04	0.2	0.04	0.02	0.05
Permutation	11	12	13	14	15	16	17	18	19	20
MSE	0.03	0.09	0.02	0.05	0.03	0.07	0.02	0.03	0.01	0.04

Table 2: MSE of our approximation for each of 20 random permutations.

Now we specifically describe how we set the cutpoints for the Mondo data; the treatment for other data is similar. To begin, we randomly selected 1,000 ratings and found the relative frequencies for $c = 1, \dots, 5$ to be about 0.22, 0.13, 0.19, 0.20, 0.26. The z -values corresponding to these areas are $z = (-0.7, -0.3, 0.1, 0.6)$. Note that $\beta_j \theta_i$ follows the normal product distribution, which is symmetric about zero; see Glen et al. (2004) and the references therein for a detailed description of this distribution. Therefore, roughly speaking, the unconditional distribution of Y^* is approximately $N(1, 23)$, where the variance is derived from the independence assumptions among parameters. Then, converting the z -values to the y^* scale gives $(\gamma_1, \gamma_2, \gamma_3, \gamma_4)$ as

$$1 + \sqrt{23} \times z = (-2.36, -0.44, 1.48, 3.88). \quad (47)$$

These cutpoints will be used in Section 5.3.

For evaluating accuracy, we first compute the approximations (26) based on a single observation y_{ij} . Given the parameter values as the initial settings, Table 1 presents our approximations and the estimates using numerical integration. Note that the initial μ_{α_j} and σ_{α_j} are both 1. In general, our approximations made smaller adjustments (so the estimates are closer to the initial values 1). Clearly, when the observed rating is higher, the posterior mean of α_j tends to be larger; and the variance is bigger when the rating is high or low. Overall, the proposed approximation is quite accurate.

Next, we evaluate the performance of our algorithm on simulated data generated from the ordinal IRT model specified in (3) and (4). We take $I = 50$, $J = 200$, and set the cutpoints to be $(\gamma_1, \gamma_2, \gamma_3, \gamma_4) = (-6, -2, 2, 6)$. Then we follow the prior distribution in this section to generate α_j , β_j and θ_i , to set initial parameter values for our algorithm, and to estimate the cutpoints by the empirical frequencies of the initial 500 ratings. Here we restrict θ_1 to be positive so as to identify the signs of the model parameters. A total of 10,000 ratings are generated. We conduct online estimation on 20 random permutations of the dataset. Table 2 reports the mean squared error (MSE) between the estimate θ^* and the true θ for each of the 20 random permutations; that is, $MSE = \sum_i (\theta_i^* - \theta_i)^2 / I$. In general, our approximations are satisfactory.

5.3 Internet ratings data

We consider two Internet ratings datasets: Mondo Times and MovieLens; see Section 2.1 for descriptions of the datasets. The starplots here are proposed by Ho and Quinn (2008a). They are based on the posterior predictive probability that a randomly selected rater will provide a rating c for product i , defined as

$$\tau_{ic} = \frac{1}{J} \sum_{j=1}^J p(y_{ij} = c | Y^{\text{obs}}). \quad (48)$$

In Ho and Quinn (2008a), the probability $p(y_{ij} | Y^{\text{obs}})$ in (48) is approximated by

$$\frac{1}{M} \sum_{m=1}^M \left(\Phi(\beta_j^{(m)} \theta_i^{(m)} + \alpha_j^{(m)} - \gamma_{c-1}^{(m)}) - \Phi(\beta_j^{(m)} \theta_i^{(m)} + \alpha_j^{(m)} - \gamma_c^{(m)}) \right), \quad (49)$$

where $\{\alpha_j^{(m)}, \beta_j^{(m)}, \theta_i^{(m)}, \gamma^{(m)}\}_{m=1}^M$ are M Monte Carlo samples. For our proposed online method, this probability is approximated by

$$p(y_{ij} = c | Y^{\text{obs}}) \approx \Phi(\mu_{\beta_j} \mu_{\theta_i} + \mu_{\alpha_j} - \gamma_{c-1}) - \Phi(\mu_{\beta_j} \mu_{\theta_i} + \mu_{\alpha_j} - \gamma_c), \quad (50)$$

where μ_{α_j} , μ_{β_j} , and μ_{θ_i} are the current estimates of the posterior means of α_j , β_j , and θ_i .

Mondo Times data

The Mondo Times data are taken from Ho and Quinn (2008a), where they removed raters who rate fewer than five products and removed products that are only rated by these raters. They end up with 3,249 ratings on 1,344 products from 232 raters. Hereafter, we shall call these the “sub-Mondo” data. Since there is no information about the timestamp of the ratings, we apply our algorithm to two randomly shuffled ratings.

To compare starplots using (49) and (50), we obtain MCMC samples from the package `Ratings` (Ho and Quinn 2008a) with initial $\gamma = (0, 1.5, 3, 4.5)$, and then estimate τ_{pc} using (49) and (50), respectively. The resulting starplots are in Figure 3. The two plots are fairly close, indicating that (50) can produce a good approximation.

Next, we compare the starplots by the offline MCMC method and the proposed online method for the sub-Mondo data, both with fixed γ , as in (47). As the package `Ratings` is designed for random γ , we have modified `Ratings` (available upon request) in order to obtain MCMC estimates with fixed γ . The results are in the left and center plots of Figure 4. We observe that, compared with the starplot by MCMC with random γ (Figure 3), the results of MCMC with fixed γ are much closer to our online method. The right plot in Figure 4 is by our method for the whole-Mondo data. Overall, the three starplots are in good agreement.

We also compare the estimates of θ . Since the MCMC samples reveal that the posterior densities of μ_{θ_i} for the twelve news outlets are roughly bell-shaped (not shown

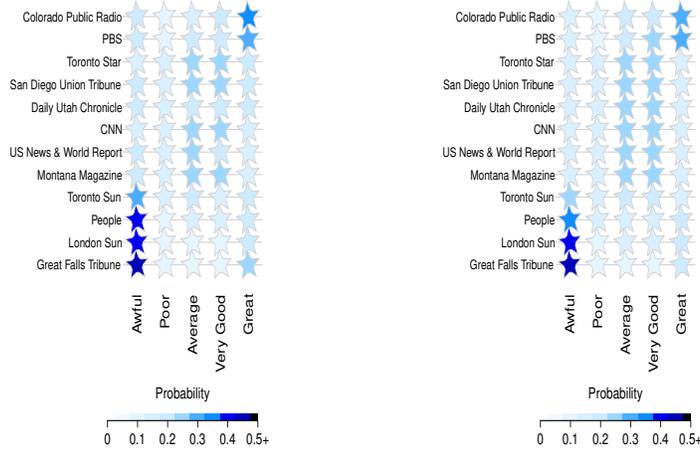


Figure 3: Starplots for Sub-Mondo data using MCMC samples. Left: by approximation (49). Right: by approximation (50).

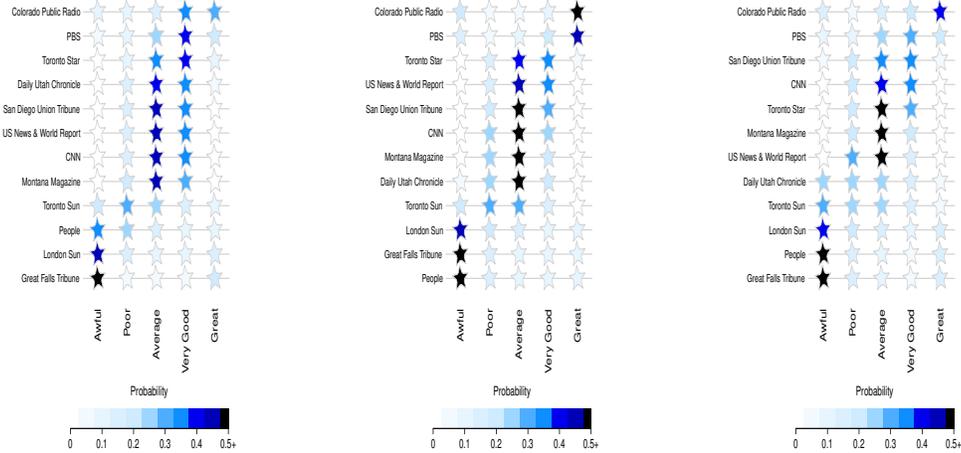


Figure 4: Starplots for news outlets with fixed γ , as in (47). Left: MCMC by modified Ratings for Sub-Mondo. Center: Sub-Mondo by online method. Right: Whole-Mondo by online method.

News outlets	MCMC1	MCMC2	online	
	random γ	fixed γ	random ordering 1	random ordering 2
US News & World Report	-0.17 (0.47)	0.13 (0.33)	0.33 (0.26)	0.42 (0.39)
Toronto Sun	-0.98 (0.37)	-1.54 (0.26)	-1.03 (0.21)	-1.36 (0.33)
Toronto Star	0.42 (0.38)	0.81 (0.28)	0.45 (0.15)	0.24 (0.29)
San Diego Union Tribune	0.1 (0.49)	0.15 (0.35)	-0.09 (0.34)	-0.01 (0.15)
People	-1.77 (0.59)	-2.22 (0.45)	-2.06 (0.69)	-2.76 (0.59)
PBS	1.22 (0.39)	1.49 (0.24)	1.30 (0.22)	1.21 (0.32)
Montana Magazine	-0.22 (0.36)	-0.37 (0.32)	-0.02 (0.15)	-0.4 (0.43)
London Sun	-2.13 (0.57)	-2.91 (0.44)	-2.13 (0.43)	-1.88 (0.35)
Great Falls Tribune	-2.8 (0.76)	-5.12 (0.7)	-1.71 (0.31)	-1.93 (0.34)
Daily Utah Chronicle	0.13 (0.8)	0.44 (0.56)	-0.28 (0.26)	-0.1 (0.28)
Colorado Public Radio	1.45 (0.6)	2.02 (0.47)	2.29 (0.7)	1.61 (0.42)
CNN	0.08 (0.19)	0.11 (0.13)	-0.34 (0.08)	-0.1 (0.1)

Table 3: Posterior means and standard deviations (in parentheses) of θ for twelve outlets.

here), Table 3 reports only the posterior means and standard deviations. In Table 3, the column MCMC1 is by package `Ratings` with default initial $\gamma = (\gamma_1, \dots, \gamma_4) = (0, 1, 3, 5)$, where $\gamma_1 = 0$ is fixed, while $\gamma_2, \gamma_3, \gamma_4$ are to be estimated; MCMC2 is obtained from the modified package in which the γ values are fixed, as in (47). The two online columns are results from two random orderings of the data, using the same fixed γ . The two MCMC columns reveal that the posterior means with random γ are somewhat different from those with fixed γ , and that the standard deviations are generally smaller if γ is fixed. Moreover, for each news outlet, the 95% posterior intervals from the four columns mostly overlap, except for the Great Falls Tribune with MCMC2. Overall, our online method performs pretty well.

MovieLens data

The MovieLens data contain the timestamp (in unix seconds since 1/1/1970 UTC) for each rating. So, we sort the ratings according to the timestamp, and study both the stationary and the dynamic model (Section 4.2). We set initial values $\sigma_{\alpha_j}^{(0)} = \sigma_{\beta_r}^{(0)} = \sigma_{\theta_p}^{(0)} = 0.1$. For the time-varying case (43), we set the state noise parameter $\sigma^2 = 0.002$, obtained by maximizing the approximate marginal likelihood (44) over 5,000 ratings. The top 10 movies are presented in Figure 5. The left panel is for the stationary case, using Algorithm 1; and the right panel is for the time-varying case, which is the same as Algorithm 1 except that $\nu^{(t)}$ in (32) is now replaced by $\bar{\nu}$ in (45).

Remarks

The positive lower bound κ in (38)-(40) is set to be 0.0001 to avoid a negative variance. Our experiments show that the results for the two real datasets are not sensitive to the κ value.

For the computational time, to produce a starplot in Figure 3 for the sub-Mondo

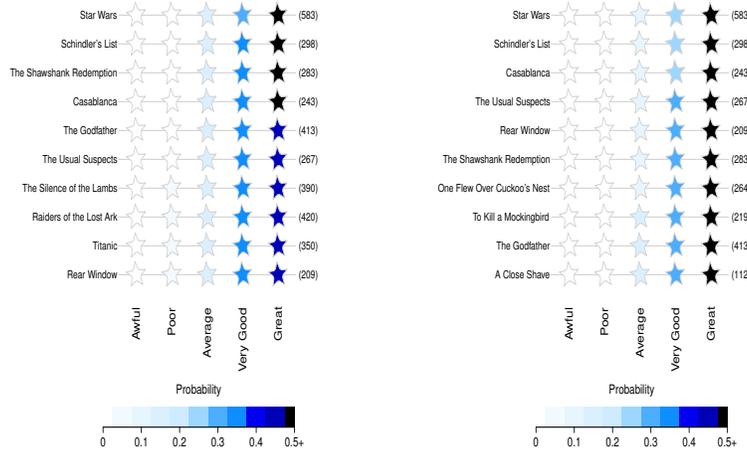


Figure 5: Starplots for MovieLens data. Left: original model; Right: model with time-varying parameters.

data, it takes about five minutes using the `Ratings` package, but just three seconds using the online algorithm written in pure R code. We also generate a subset of the MovieLens data, with about 68,000 ratings, for which the online algorithm takes about 20 seconds, while the MCMC method via `Ratings` takes about 30 minutes. The offline MCMC method may take more time if real-time parameter adjustment is required. The computations for the online method here are available at <http://www3.nccu.edu.tw/~chweng/BaIRT.zip>

6 Concluding Remarks

Online methods are necessary when large data continue to arrive and real-time parameter adjustment is needed. We have shown how to develop Bayesian online parameter estimation for IRT models with Likert-type data. The experiments on two real datasets are satisfactory.

We have also compared our real-time estimation method with the offline MCMC methods via the package `Ratings` and its modification for fixed γ . Though sacrificing some accuracy, in general, our proposed method achieves a good performance, but with considerably less computational time. Thus, for situations where faster approximate methods are desirable, our proposed method can be a useful alternative to offline methods. That said, we have to point out some limitations of our method. First, with only mean and variance updates, our method can not provide estimates of quantities of interest, which are easily obtainable by MCMC methods. Moreover, our use of fixed cutpoints γ based on initial observed proportions $\#\{y_{ij} = c\}/N$ (see Section 5.2) may

be inappropriate. We may alleviate this weakness by updating the observed proportions sequentially, and then adjust the γ values either sequentially or after collecting a certain number of observations. The adaptive γ can better reflect the evolution of the ratings distribution. In an unreported experiment, we adjusted γ whenever collecting 500 ratings for the sub-Mondo data (Section 5.3). The results are close to the unadjusted case, possibly because the ratings are randomly permuted; furthermore, the computation time does not increase much because the observed ratings frequencies can be easily accumulated.

Some questions deserve further study. First, some theoretical analysis on the discrepancies between the offline method and the proposed online one would be worthwhile. This can be challenging because the rater-product matrix of ratings is very sparse and the number of parameters would increase with the data. Second, it is desirable to extend the proposed method to other IRT models, such as the Rating Scale model (Andrich 1978) and the Partial Credit model (Masters 1982). These models are direct extensions of the one-parameter IRT model in Example 1, and sufficient statistics exist for the model parameters. One may also consider a bias term for the product in model (5). In principle, similar derivations could be obtained straightforwardly.

1 Derivations of (26)-(28)

Recall from the beginning of Section 3 that ϕ and Φ denote the density and distribution function of $N(0, 1)$, and let $\phi(x|\mu, \sigma)$ denote the density of $N(\mu, \sigma^2)$. Now we take the following three steps.

Step 1. Let f be as in (25). Obtain the following derivatives of f :

$$\frac{\partial \log f}{\partial \alpha_j^*} = \sigma_{\alpha_j} \cdot \Omega(x, a), \quad \frac{\partial \log f}{\partial \beta_j^*} = \sigma_{\beta_j} \theta_i \cdot \Omega(x, a), \quad \frac{\partial \log f}{\partial \theta_i^*} = \sigma_{\theta_i} \beta_j \cdot \Omega(x, a), \quad (51)$$

$$\frac{\partial^2 \log f}{\partial (\alpha_j^*)^2} = \sigma_{\alpha_j}^2 \cdot \Delta(x, a), \quad \frac{\partial^2 \log f}{\partial (\beta_j^*)^2} = (\sigma_{\beta_j} \theta_i)^2 \cdot \Delta(x, a), \quad \frac{\partial^2 \log f}{\partial (\theta_i^*)^2} = (\sigma_{\theta_i} \beta_j)^2 \cdot \Delta(x, a),$$

where

$$x = \beta_j \theta_i + \alpha_j - \gamma_{c-1} \quad \text{and} \quad a = \gamma_c - \gamma_{c-1}. \quad (52)$$

Step 2. Plug the derivatives of f into the right-hand sides of (16) and (20).

Step 3. Evaluate the expectations on the right-hand sides of (16) and (20) to obtain (26)-(28). Below we give details of the derivation of $\tilde{\mu}_{\alpha_j}$ in (26). To begin, we need the following results: for $c > 0$,

$$\int_{-\infty}^{\infty} \phi(\theta|\mu, \sigma) \phi(b|a\theta, c) d\theta = \phi(b|a\mu, \sqrt{a^2\sigma^2 + c^2}) = \frac{1}{\sqrt{2\pi(a^2\sigma^2 + c^2)}} e^{-\frac{(a\mu-b)^2}{2(a^2\sigma^2 + c^2)}}, \quad (53)$$

$$\int_{-\infty}^{\infty} \phi(\theta|\mu, \sigma) \Phi\left(\frac{b-a\theta}{c}\right) d\theta = \Phi\left(\frac{b-a\mu}{\sqrt{a^2\sigma^2 + c^2}}\right). \quad (54)$$

These equations are obtained by writing the function $\Phi(\cdot)$ as an integral and interchanging the order of the integrals. By (16) and (51), we have

$$E(\alpha_j|y_{ij}) = \mu_{\alpha_j} + \sigma_{\alpha_j}^2 E(\Omega(x, a)|y_{ij}),$$

where x and a are defined in (52), and

$$\begin{aligned} E(\Omega(x, a)|y_{ij}) &= \int \int \int \left(\frac{\phi(x) - \phi(x-a)}{\Phi(x) - \Phi(x-a)} \right) p(\alpha_j, \beta_j, \theta_i|y_{ij}) d\alpha_j d\beta_j d\theta_i \\ &= \frac{\int \int \int [\phi(x) - \phi(x-a)] \phi(\alpha_j|\mu_{\alpha_j}, \sigma_{\alpha_j}) \phi(\beta_j|\mu_{\beta_j}, \sigma_{\beta_j}) \phi(\theta_i|\mu_{\theta_i}, \sigma_{\theta_i}) d\alpha_j d\beta_j d\theta_i}{\int \int \int [\Phi(x) - \Phi(x-a)] \phi(\alpha_j|\mu_{\alpha_j}, \sigma_{\alpha_j}) \phi(\beta_j|\mu_{\beta_j}, \sigma_{\beta_j}) \phi(\theta_i|\mu_{\theta_i}, \sigma_{\theta_i}) d\alpha_j d\beta_j d\theta_i} \\ &= \frac{I_1 - I_2}{II_1 - II_2}. \end{aligned}$$

For I_1 , write $\phi(x) = \phi(\beta_j \theta_i + \alpha_j - \gamma_{c-1}) = \phi(\gamma_{c-1} - \beta_j \theta_i | \alpha_j, 1)$, apply (53) twice and use the relation $\phi(\theta|\mu, \sigma) = \phi(\mu|\theta, \sigma)$ to obtain

$$\begin{aligned} I_1 &= \int \int \int \phi(\gamma_{c-1} - \beta_j \theta_i | \alpha_j, 1) \phi(\alpha_j | \mu_{\alpha_j}, \sigma_{\alpha_j}) \phi(\beta_j | \mu_{\beta_j}, \sigma_{\beta_j}) \phi(\theta_i | \mu_{\theta_i}, \sigma_{\theta_i}) d\alpha_j d\beta_j d\theta_i \\ &= \int \int \phi(\gamma_{c-1} - \beta_j \theta_i | \mu_{\alpha_j}, \sqrt{1 + \sigma_{\alpha_j}^2}) \phi(\beta_j | \mu_{\beta_j}, \sigma_{\beta_j}) \phi(\theta_i | \mu_{\theta_i}, \sigma_{\theta_i}) d\beta_j d\theta_i \\ &= \int \int \phi(\gamma_{c-1} - \mu_{\alpha_j} | \beta_j \theta_i, \sqrt{1 + \sigma_{\alpha_j}^2}) \phi(\beta_j | \mu_{\beta_j}, \sigma_{\beta_j}) \phi(\theta_i | \mu_{\theta_i}, \sigma_{\theta_i}) d\beta_j d\theta_i \\ &= \int \phi(\gamma_{c-1} - \mu_{\alpha_j} | \mu_{\beta_j} \theta_i, \sqrt{1 + \sigma_{\alpha_j}^2 + \sigma_{\beta_j}^2 \theta_i^2}) \phi(\theta_i | \mu_{\theta_i}, \sigma_{\theta_i}) d\theta_i. \end{aligned}$$

For the last line, we can not apply (53) to evaluate the integral because θ_i appears in the variance of the ϕ function. Now assume that σ_{θ_i} is small, so that θ_i is close to μ_{θ_i} with a high probability. In our experiments, σ_{θ_i} is smaller than 1. So, we approximate θ_i in the variance by μ_{θ_i} to obtain (55), and then apply (53) to obtain (56):

$$I_1 \approx \int \phi(\gamma_{c-1} - \mu_{\alpha_j} | \mu_{\beta_j} \theta_i, \sqrt{1 + \sigma_{\alpha_j}^2 + \sigma_{\beta_j}^2 \mu_{\theta_i}^2}) \phi(\theta_i | \mu_{\theta_i}, \sigma_{\theta_i}) d\theta_i \quad (55)$$

$$\begin{aligned} &= \phi(\gamma_{c-1} - \mu_{\alpha_j} | \mu_{\beta_j} \mu_{\theta_i}, \sqrt{1 + \sigma_{\alpha_j}^2 + \sigma_{\beta_j}^2 \mu_{\theta_i}^2 + \sigma_{\theta_i}^2 \mu_{\beta_j}^2}) \\ &= \frac{1}{\nu} \phi\left(\frac{\mu_{\beta_j} \mu_{\theta_i} + \mu_{\alpha_j} - \gamma_{c-1}}{\nu}\right), \end{aligned} \quad (56)$$

where $\nu = (1 + \sigma_{\alpha_j}^2 + \sigma_{\beta_j}^2 \mu_{\theta_i}^2 + \sigma_{\theta_i}^2 \mu_{\beta_j}^2)^{1/2}$. The expression for I_2 is the same as I_1 except that γ_{c-1} is replaced by γ_c . The calculations of II_1 and II_2 are similar, but will use (54) rather than (53); we omit the details and only present the results:

$$II_1 \approx \Phi\left(\frac{\mu_{\theta_i} \mu_{\beta_j} + \mu_{\alpha_j} - \gamma_{c-1}}{\nu}\right), \quad II_2 \approx \Phi\left(\frac{\mu_{\theta_i} \mu_{\beta_j} + \mu_{\alpha_j} - \gamma_c}{\nu}\right). \quad (57)$$

Therefore, the posterior mean of α_j can be approximated as

$$\tilde{\mu}_{\alpha_j} \approx \mu_{\alpha_j} + \frac{\sigma_{\alpha_j}^2}{\nu} \frac{\phi\left(\frac{\mu_{\theta_i} \mu_{\beta_j} + \mu_{\alpha_j} - \gamma_{c-1}}{\nu}\right) - \phi\left(\frac{\mu_{\theta_i} \mu_{\beta_j} + \mu_{\alpha_j} - \gamma_c}{\nu}\right)}{\Phi\left(\frac{\mu_{\theta_i} \mu_{\beta_j} + \mu_{\alpha_j} - \gamma_{c-1}}{\nu}\right) - \Phi\left(\frac{\mu_{\theta_i} \mu_{\beta_j} + \mu_{\alpha_j} - \gamma_c}{\nu}\right)}.$$

This establishes the first equation in (26). To derive $\tilde{\sigma}_{\alpha_j}^2$ in (26), we need (20) and some equations analogous to (53) and (54). The formulas in (27) and (28) can be obtained similarly and we omit the derivations.

2 Proof of Lemma 1

Proof. We only prove (a). The proof of (b) is similar and is omitted; (c) is a simple consequence of (a) and (b). Part (d) can be derived by L'Hôpital's rule. For (a), we have

$$\begin{aligned}
 \Lambda\left(\frac{a}{2} + z, a\right) &= \frac{\left(\frac{a}{2} + z\right)\phi\left(\frac{a}{2} + z\right) - \left(z - \frac{a}{2}\right)\phi\left(z - \frac{a}{2}\right)}{\Phi\left(\frac{a}{2} + z\right) - \Phi\left(z - \frac{a}{2}\right)} \\
 &= \frac{\left(\frac{a}{2} + z\right)\phi\left(-\frac{a}{2} - z\right) - \left(z - \frac{a}{2}\right)\phi\left(\frac{a}{2} - z\right)}{1 - \Phi\left(-\frac{a}{2} - z\right) - 1 + \Phi\left(\frac{a}{2} - z\right)} \\
 &= \frac{\left(\frac{a}{2} - z\right)\phi\left(\frac{a}{2} - z\right) - \left(-\frac{a}{2} - z\right)\phi\left(-\frac{a}{2} - z\right)}{\Phi\left(\frac{a}{2} - z\right) - \Phi\left(-\frac{a}{2} - z\right)} \\
 &= \Lambda\left(\frac{a}{2} - z, a\right),
 \end{aligned}$$

where the second equality follows from $\phi(x) = \phi(-x)$ and $\Phi(x) = 1 - \Phi(-x)$. ■

References

- Albert, J. (2015). “Introduction to Bayesian item response modeling.” *International Journal of Quantitative Research in Education*, 2(3-4): 178–193.
- Albert, J. H. (1992). “Bayesian estimation of normal ogive item response curves using Gibbs sampling.” *Journal of Educational and Behavioral Statistics*, 17(3): 251–269.
- Andrich, D. (1978). “A rating formulation for ordered response categories.” *Psychometrika*, 43: 561–573.
- Bishop, C. M. (2008). “A New Framework for Machine Learning.” In *Lecture Notes in Computer Science LNCS 5050*, 1–24. Springer.
- Bock, R. D. and Aitkin, M. (1981). “Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm.” *Psychometrika*, 46: 443–459.
- Boyer, X. and Koller, D. (1998). “Tractable inference for complex stochastic processes.” In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, 33–42. Morgan Kaufmann Publishers Inc.
- Chen, X., Bennett, P. N., Collins-Thompson, K., and Horvitz, E. (2013). “Pairwise ranking aggregation in a crowdsourced setting.” In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, 193–202. ACM.

- Chevalier, J. A. and Mayzlin, D. (2006). “The effect of word of mouth on sales: Online book reviews.” *Journal of Marketing Research*, 43(3): 345–354.
- Dangauthier, P., Herbrich, R., Minka, T., and Graepel, T. (2008). “TrueSkill Through Time: Revisiting the History of Chess.” In *Advances in Neural Information Processing Systems 20*. Cambridge, MA: MIT Press.
- De Ayala, R. J. (2013). *The Theory and Practice of Item Response Theory*. Guilford Publications.
- de Freitas, J. F., Niranjan, M., et al. (2000). “Hierarchical Bayesian models for regularization in sequential learning.” *Neural Computation*, 12(4): 933–953.
- Embretson, S. E. (1991). “A multidimensional latent trait model for measuring learning and change.” *Psychometrika*, 56: 495–515.
- Faes, C., Ormerod, J. T., and Wand, M. P. (2011). “Variational Bayesian inference for parametric and nonparametric regression with missing data.” *Journal of the American Statistical Association*, 106: 959–971.
- Fox, J.-P. and Glas, C. A. W. (2001). “Bayesian estimation of a multilevel IRT model using Gibbs sampling.” *Psychometrika*, 66: 269–286.
- Glen, A. G., Leemis, L. M., and Drew, J. H. (2004). “Computing the distribution of the product of two continuous random variables.” *Computational Statistics and Data Analysis*, 44(3): 451–464.
- Glickman, M. E. (1999). “Parameter estimation in large dynamic paired comparison experiments.” *Applied Statistics*, 48(3): 377–394.
- Glickman, M. E. and Stern, H. S. (1998). “A state-space model for National Football League scores.” *Journal of the American Statistical Association*, 93: 25–35.
- Hall, P., Pham, T., Wand, M. P., and Wang, S. S. J. (2011). “Asymptotic normality and valid inference for Gaussian variational approximation.” *The Annals of Statistics*, 39(5): 2502–2532.
- Herbrich, R., Minka, T., and Graepel, T. (2007). “TrueSkillTM: A Bayesian Skill Rating System.” In Schölkopf, B., Platt, J., and Hoffman, T. (eds.), *Advances in Neural Information Processing Systems 19*, 569–576. Cambridge, MA: MIT Press.
- Ho, D. E. and Quinn, K. M. (2008a). “Improving the presentation and interpretation of online ratings data with model-based figures.” *The American Statistician*, 62(4): 279–288.
- (2008b). “Measuring explicit political positions of media.” *Quarterly Journal of Political Science*, 3: 353–377.
- James, W. and Stein, C. (1961). “Estimation with quadratic loss.” In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, volume I, 361–379. Univ. California Press, Berkeley.

- Jazwinski, A. H. (1969). “Adaptive filtering.” *Automatica*, 5(4): 475–485.
- Johnson, V. E. and Albert, J. H. (1999). *Ordinal Data Modeling*. New York: Springer.
- Kim, S.-H. (2001). “An evaluation of a Markov chain Monte Carlo method for the Rasch model.” *Applied Psychological Measurement*, 25(2): 163–176.
- Koren, Y., Bell, R., and Volinsky, C. (2009). “Matrix factorization techniques for recommender systems.” *IEEE Computer*, 42(8): 30–37.
- Liu, Y. (2006). “Word of mouth for movies: Its dynamics and impact on box office revenue.” *Journal of Marketing*, 70(3): 74–89.
- MacKay, D. J. (1992). “The evidence framework applied to classification networks.” *Neural Computation*, 4(5): 720–736.
- Martin, A. D. and Quinn, K. M. (2002). “Dynamic ideal point estimation via Markov chain Monte Carlo for the U.S. supreme court, 1953-1999.” *Political Analysis*, 10: 134–152.
- Masters, G. N. (1982). “A Rasch model for partial credit scoring.” *Psychometrika*, 47(2): 149–174.
- Maybeck, P. S. (1982). *Stochastic Models, Estimation, and Control*. Academic Press.
- Minka, T. (2001). “A family of algorithms for approximate Bayesian inference.” Ph.D. thesis, MIT.
- Molenaar, I. W. (1995). “Estimation of item parameters.” In Fischer, G. and Molenaar, I. (eds.), *Rasch Models: Foundations, Recent Developments, and Applications*, 39–51. Springer Verlag.
- Muraki, E. (1990). “Fitting a polytomous item response model to Likert-type data.” *Applied Psychological Measurement*, 14(1): 59–71.
- Patz, R. J. and Junker, B. (1999). “A straightforward approach to Markov Chain Monte Carlo methods for item response models.” *Journal of Educational and Behavioral Statistics*, 24: 146–178.
- Rasch, G. (1961). “On general laws and the meaning of measurement in psychology.” In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, volume IV, 321–333. Univ. California Press, Berkeley.
- Robbins, H. and Monro, S. (1951). “A stochastic approximation method.” *The Annals of Mathematical Statistics*, 22(3): 400–407.
- Rosenblatt, F. (1958). “The perceptron: A probabilistic model for information storage and organization in the brain.” *Psychological Reviews*, 7: 551–585.
- Rue, H., Martino, S., and Chopin, N. (2009). “Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations.” *Journal of the Royal Statistical Society Series B*, 71: 319–392.

- Saad, D. (1998). (editor) *On-line Learning in Neural Networks*. Cambridge University Press.
- Samejima, F. (1969). “Estimation of latent ability using a response pattern of graded scores.” *Psychometrika monograph supplement*, 71(4): 1–100.
- Shalev-Shwartz, S. (2011). “Online learning and online convex optimization.” *Foundations and Trends[®] in Machine Learning*, 4(2): 107–194.
- Spiegelhalter, D. J. and Lauritzen, S. L. (1990). “Sequential updating of conditional probabilities on directed graphical structures.” *Networks*, 20(5): 579–605.
- Stein, C. (1981). “Estimation of the mean of a multivariate normal distribution.” *The Annals of Statistics*, 9: 1135–1151.
- Thurstone, L. L. (1927). “A law of comparative judgement.” *Psychological Reviews*, 34: 273–286.
- van der Linden, W. J. and Hambleton, R. K. (2013). *Handbook of Modern Item Response Theory*. Springer Science and Business Media.
- Wang, X., Berger, J. O., and Burdick, D. S. (2013). “Bayesian analysis of dynamic item response models in educational testing.” *The Annals of Applied Statistics*, 7(1): 126–153.
- Weng, R. C. (2010). “A Bayesian Edgeworth expansion by Stein’s Identity.” *Bayesian Analysis*, 5(4): 741–764.
- (2015). “Expansions for multivariate densities.” *Journal of Statistical Planning and Inference*, 167: 174–181.
- Weng, R. C. and Lin, C.-J. (2011). “A Bayesian approximation method for online ranking.” *Journal of Machine Learning Research*, 12: 267–300.
- Wistuba, M., Schaefers, L., and Platzner, M. (2012). “Comparison of Bayesian Move Prediction Systems for Computer Go.” In *2012 IEEE Conference on Computational Intelligence and Games (CIG)*, 91–99. IEEE.
- Woodroffe, M. (1989). “Very weak expansions for sequentially designed experiments: linear models.” *The Annals of Statistics*, 17: 1087–1102.
- Woodroffe, M. and Coad, D. S. (1997). “Corrected confidence sets for sequentially designed experiments.” *Statistica Sinica*, 7: 53–74.

Acknowledgments

The authors would like to thank two referees and an Associate Editor for their comments on the paper, which helped to improve its quality. The first author was partially supported by the Ministry of Science and Technology of Taiwan.