Bayesian Psychometrics

Xiang Liu, Matthew S. Johnson, and Zhuangzhuang Han

Educational Testing Service

Abstract

The application of Bayesian methods in psychometrics is getting popular. It offers a set powerful and flexible statistical tools for analyzing psychometric data. In this entry, we provide a brief introduction of essential concepts of Bayesian data analysis - model formulation, posterior computation, and model fit evaluation. These concepts are demonstrated by using a simple 1-parameter normal ogive IRT model as an example.

Keywords: Bayesian, psychometrics, IRT, MCMC, posterior predictive model checks

Bayesian Psychometrics

Introduction

Bayesian statistical methods have become popular in the field of psychometrics. Thanks to the development of general purpose Bayesian inference software (e.g., *Stan*; Gelman et al., 2015), researchers are now able to quickly develop and fit sophisticated and complex psychometric models that once required significant computational expertise to estimate. Levy (2009) gives a nice overview and survey of the applications of Bayesian estimation for psychometric models. Some examples of development in this area may include Bayesian estimation of a broad class of cognitive diagnostic models (Culpepper, 2015; Liu & Johnson, 2019), the extended marginal Rasch model (Maris et al., 2015), a multilevel item response theory (IRT) model (Fox & Glas, 2001), and an unfolding response model (Johnson & Junker, 2003). In addition, Bayesian methods also provide a class of approaches alternative to classical or frequentist statistics (e.g. maximum likelihood, confidence intervals, p-values). In certain cases, Bayesian approaches can alleviate some difficulties classical approaches may have. Examples of the use of Bayesian methods for psychometric modeling include estimating uncertain elements of the Q-matrix in a CDM model (DeCarlo, 2012), nonparametric ordered latent class modeling (Liu, 2019), robust IRT outlier detection (Öztürk & Karabatsos, 2017), and evaluating model fit (Sinharay, 2015; Sinharay, 2005; Sinharay & Almond, 2007). These mentioned examples of Bayesian methods in psychometrics are far from exhaustive. The main purpose of this entry is to give readers a brief introduction to some of the most essential concepts and steps involved in a Bayesian analysis of psychometric models.

Bayesian data analysis

Let $\boldsymbol{X}$ denote the observed random variables (i.e. the data), and $\boldsymbol{\theta}$ denote the vector of unobserved random variables (i.e. the parameters). In Bayesian statistics, the parameters are treated as random and are given prior distributions. This is in contrast to the classical likelihood methods where parameters are treated as fixed but unknown

parameters. We are typically interested in using the data to estimate and make inferences about the vector of unknown parameters. A Bayesian analysis involves three essential steps:

1. Model specification. In this step, we set up a full probability model that describes the joint probability distribution for all observed and unobserved random variables, i.e. $p(\boldsymbol{X}, \boldsymbol{\theta})$. According to the definition of conditional probability, this joint probability distribution can be factorized as $p(\boldsymbol{X}, \boldsymbol{\theta}) = p(\boldsymbol{X}|\boldsymbol{\theta})p(\boldsymbol{\theta})$. The marginal density, $p(\boldsymbol{\theta})$, is often referred to as the *prior distribution* which expresses any prior belief we may have on the parameters with uncertainty, and the conditional density, $p(\boldsymbol{X}|\boldsymbol{\theta})$, is called the *sampling distribution* (or the generating distribution) which describes the data generation process. Setting up these models generally requires some knowledge of the underlying problems with, sometimes, considerations for mathematical convenience.

2. Posterior inference. In this step, we calculate and interpret the *posterior distribution* from which we make inferences about the parameters. The posterior distribution is the conditional distribution of the unobserved random variables (i.e. parameters) which we are typically interested in, given the observed data, i.e. $p(\boldsymbol{\theta}|\boldsymbol{X}) = p(\boldsymbol{\theta}, \boldsymbol{X})/p(\boldsymbol{X}) = p(\boldsymbol{X}|\boldsymbol{\theta})p(\boldsymbol{\theta})/p(\boldsymbol{X})$. This is known as Bayes' rule. As a function of parameters $\boldsymbol{\theta}$, $p(\boldsymbol{X}|\boldsymbol{\theta})$ is also referred to as the likelihood function through which the data $\boldsymbol{X}$ influences the posterior distribution of the parameters. To obtain the density of the marginal distribution of the data, we need to evaluate $p(\boldsymbol{X}) = \int p(\boldsymbol{X}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}$. Except for certain extremely simple models, this (potentially high dimensional) integral cannot be evaluated analytically. Consequently, the posterior is generally not available in closed forms. Instead, the posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{X})$ is usually approximated using sampling methods such as the Markov chain Monte Carlo (MCMC; Gelman et al., 2013; Neal, 1998).

3. Model fit evaluation. Once a model has been set up and the posterior distribution

has been computed, it is important to assess the fit of the model. Inferences drawn from a model that does not fit well could be very misleading. In a Bayesian framework, the assessment of model fit is typically performed through the posterior predictive model checking (PPMC; Gelman et al., 2013; Rubin, 1984). If the model fits reasonably well, the observed data or the derived statistics of our practical interest should look plausible under the posterior predictive distribution. The identified aspects of model misfit could help improve or expand the model. Sometimes, identifying misfit itself could be of practical interest, for example, to identify aberrant response patterns (e.g. Sinharay, 2015).

These steps are true to all Bayesian modeling applications. Bayesian psychometrics is an example of such applications to psychometric modeling. In the following sections, we demonstrate these essential steps of the Bayesian data analysis with a little more details by using a simple 1-parameter normal ogive IRT model as an example.

Model specification

Let $X_{ij} = x_{ij}$ be a dichotomous random variable of the item response from the i^{th} student to the j^{th} item, $i = 1, 2, \ldots, N$ and $j = 1, 2, \ldots, J$. And $x_{ij} = 1$ if the response is correct, $x_{ij} = 0$ otherwise. Under the 1-parameter normal ogive (1PNO) IRT model, the probability of a correct response is given by

$$P(X_{ij} = 1|\theta_i, a, b_j) = \Phi(a(\theta_i - b_j)), \tag{1}$$

where $\Phi(\cdot)$ denotes the cumulative probability density function (CDF) of the standard normal distribution, θ_i is the person parameter, b_j is the item difficulty parameter, and a is the common discrimination parameter shared among J items. The person parameter is assumed to follow a standard normal distribution, i.e. $\theta_i \sim N(0, 1)$. Because a is a constant across items, the item characteristic curves (ICC) never cross each other and only differ by locations (see Figure 1).

The 1PNO as in Equation 1 can be expressed as a Bayesian hierarchical model. For

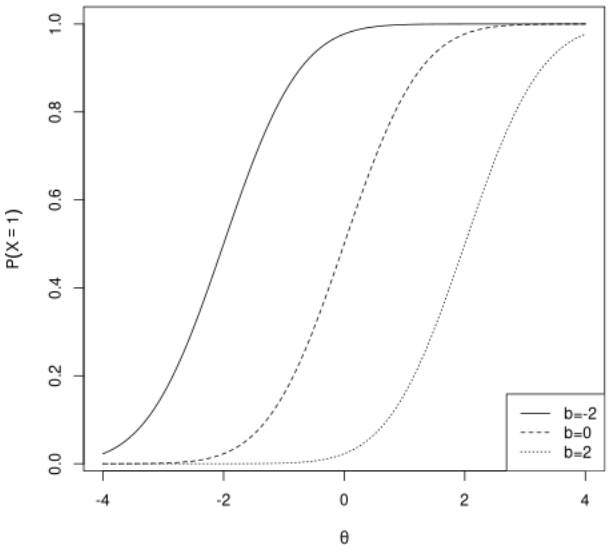

Figure 1

ICC of the 1PNO IRT model

computational efficiency, we reparameterize the model so that the model is in its slope-intercept form,

$$P(X_{ij} = 1|\theta_i, \psi_j) = \Phi(\theta_i + \psi_j), \tag{2}$$

and $\theta_i \sim N(0, a^2)$, where a^2 is the variance of the normal distribution.

A prior distribution describes any prior belief we may have (or the lack of it) on the parameter with uncertainty. For instance, in the current example, there may be reasonable prior information on the item parameter ψ_j from similar items previously administered to a comparable population of students. In this case, we may choose a normal prior with mean μ_{ψ_j} informed by prior information and variance $\sigma^2_{\psi_j}$ expressing the level of uncertainty, i.e. $\psi_j \sim N(\mu_{\psi_j}, \sigma^2_{\psi_j})$. A prior with larger variance (or a flatter prior, see Figure 2) is less informative and appropriate when there is greater uncertainty about a parameter, thus having less impact on the posterior inference of the parameter. However, prior information may not always be available. In this case, instead of choosing a fixed prior with a large variance, an alternative approach is to use a hierarchical prior. For each item parameter

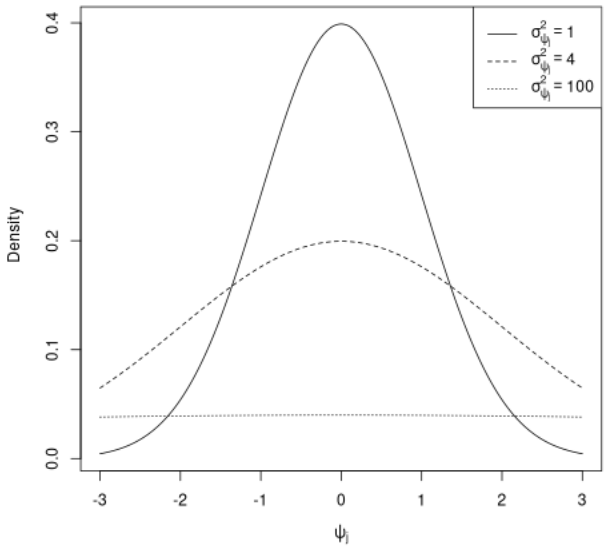

Figure 2

Density of prior distribution of ψ_j

ψ_j, we assign a same Gaussian prior, $\psi_j \sim N(\mu_\psi, \sigma^2_\psi)$. Hyper-priors are assigned to the parameters of this prior distribution, i.e. $\mu_\psi \sim N(0, 10^2)$ and $\sigma^2_\psi \sim Ga^{-1}(10^{-4}, 10^{-4})$, where $Ga^{-1}(\cdot)$ denotes the inverse-gamma distribution (see Gelman et al., 2013, p.43). The chosen inverse-gamma distribution is noninformative if most of the posterior density is away from 0 as the prior density is flat in the tail. The choice of this set of priors is also mathematically convenient. The normal prior and inverse-gamma prior are conditionally conjugate for the model. A prior is said to be conjugate for the likelihood function if the resulting posterior distribution is in the same probability distribution family. In this case, the conditional posteriors for the mean parameter and the variance parameter normal and inverse-gamma. We choose a similar noninformative inverse-gamma distribution for a^2, i.e. $a^2 \sim Ga^{-1}(10^{-4}, 10^{-4})$. We can then write the 1PNO IRT model has a Bayesian

hierarchical model,

$$
\begin{aligned}
x_{ij}|\theta_i, \psi_j &\sim Bernoulli(\Phi(\theta_i + \psi_j)) \\
\psi_j|\mu_\psi, \sigma^2_\psi &\sim N(\mu_\psi, \sigma^2_\psi) \\
\mu_\psi &\sim N(0, 10^2) \\
\sigma^2_\psi &\sim Ga^{-1}(10^{-4}, 10^{-4}) \\
\theta_i|a^2 &\sim N(0, a^2) \\
a^2 &\sim Ga^{-1}(10^{-4}, 10^{-4}).
\end{aligned} \tag{3}
$$

Posterior inference

For any practical applications, we are interested in inferring model parameters given the data, in other words, computing the posterior distribution. The *joint* posterior distribution of the parameters, $p(\boldsymbol{\theta}, \boldsymbol{\psi}, \mu_\psi, \sigma^2_\psi, a^2|\boldsymbol{x})$, cannot be analytically computed. As a result, we use MCMC to sample from this unknown distribution. The simplest MCMC algorithm is perhaps the Gibbs sampling. A Gibbs sampler (Geman & Geman, 1984) works by iteratively sampling from the full conditional distributions of the parameters. Generally, this is feasible if the full conditional distributions can be computed in closed form and sampled from easily. However, this may not be true for some models. In these cases, other MCMC algorithms are required, e.g. the Metropolis-Hastings algorithm (Hastings, 1970). The literature on MCMC algorithms is vast, and the development of new MCMC methods is still an extremely active research area. Readers may refer to Gelman et al. (2013) and Neal (1998) for some detailed introduction and review on this topic.

For our example of the 1PNO model, a direct Gibbs sampling is possible. However, we would need to augment latent variables. For every item response x_{ij}, there is a latent variable y_{ij} associated. The relationship between them is deterministic such that

$$
x_{ij} = \begin{cases} 1 \text{ if } y_{ij} > 0, \\ 0 \text{ otherwise,} \end{cases}
$$

where $y_{ij} \sim N(\theta_i + \psi_j, 1)$. It is straightforward to verify that, under this specification, the

probability of a correct response remains $\int_{-\infty}^{\theta_i+\psi_j} \phi(y)dy = \Phi(\theta_i + \psi_j)$.

To derive the densities full conditional distributions, we use the definition of conditional probabilities and the chain rule. For example,

$$p(y_{ij}|x_{ij}, \theta_i, \psi_j) = \frac{p(x_{ij}|y_{ij})p(y_{ij}|\theta_i, \psi_j)}{\sum_{y_{ij}} p(x_{ij}|y_{ij})p(y_{ij}|\theta_i, \psi_j)}. \tag{4}$$

Notice that the denominator of Equation 4 is a constant (with respect to y_{ij}) that normalizes the full conditional density. Therefore, we can conveniently write

$$\begin{aligned} p(y_{ij}|x_{ij}, \theta_i, \psi_j) &= Cp(x_{ij}|y_{ij})p(y_{ij}|\theta_i, \psi_j) \\ &\propto p(x_{ij}|y_{ij})p(y_{ij}|\theta_i, \psi_j). \end{aligned}$$

Given that $p(x_{ij}|y_{ij})$ is a piecewise point density with support on either $x_{ij} = 1$ or $x_{ij} = 0$ depending on y_{ij}, the full conditional distribution of y_{ij} is a truncated normal distribution, i.e.

$$y_{ij}|x_{ij}, \theta_i, \psi_j \sim \begin{cases} N(\theta_i + \psi_j, 1)T(0,) \text{ if } x_{ij} = 1, \\ N(\theta_i + \psi_j, 1)T(, 0) \text{ if } x_{ij} = 0. \end{cases}$$

$T(0,)$ denotes that the distribution is truncated with a lower bound of 0; while $T(, 0)$ denotes a truncation with a upper bound of 0.

The full conditional distributions of other parameters can be derived similarly. Since the priors are chosen so that they are *conditionally* conjugate, the full conditional distributions of the parameters are within the same distribution families as their priors. The standard conjugacy results apply (for a reference, see Gelman et al., 2013). Here, we state the results without detailed derivations.

Now we can summarize the Gibbs sampler. After assigning initial values to parameters, at the M^{th} iteration,

1. for each i and j, update y_{ij} by drawing from $N(\theta_i + \psi_j, 1)T(0,)$ if $x_{ij} = 1$; else draw from $N(\theta_i + \psi_j, 1)T(, 0)$;

2. for each i, update θ_i by drawing from $N\left(\frac{1}{1/a^2+J}(\sum_j (y_{ij} - \psi_j)), (1/a^2 + J)^{-1}\right)$;

3. for each j, update ψ_j by drawing from $N\left(\frac{1}{1/\sigma_\psi^2+N}(\mu_\psi/\sigma_\psi^2+\sum_i(y_{ij}-\theta_i)), \frac{1}{1/\sigma_\psi^2+N}\right)$;
4. update a^2 by drawing from $Ga^{-1}(10^{-4}+\frac{N}{2}, 10^{-4}+\frac{\sum_i \theta_i^2}{2})$;
5. update μ_ψ by drawing from $N\left(\frac{1}{1/10^2+J/\sigma_\psi^2}(\frac{\sum_j \psi_j}{\sigma_\psi^2}), (1/10^2+J/\sigma_\psi^2)^{-1}\right)$;
6. update σ_ψ^2 by drawing from $Ga^{-1}\left(10^{-4}+J/2, 10^{-4}+\frac{\sum_j(\psi_j-\mu_\psi)^2}{2}\right)$.

Keep updating long enough, the draws are eventually random samples from the joint posterior distribution. Even though the theory of Markov chains guarantees the eventual convergence to the desired target distribution which is the joint posterior (see Neal, 1998, for a review), in practice, we can only draw finite number of samples. It is important to assess the convergence so that we are reasonably confident using the draws to form valid posterior inference. There are many methods for monitoring the convergence of MCMC (Cowles & Carlin, 1996). For example, the Gelman-Rubin statistic (Gelman & Rubin, 1992) is one of the most popular methods and implemented in many Bayesian computation software.

Once the samples are drawn, to make inferences about parameters, we need to summarize the posterior distribution. Typically, we use mean, mode, and median to describe the central tendencies of the posterior distribution and serve as a point estimate. Among them, posterior mean is probably the most common. In addition the point estimate, the variance or standard deviation can be used to describe the dispersion of the posterior distribution. We may also be interested in forming posterior intervals which is referred to as the credible intervals. One way to construct a $100(1-\alpha)\%$ credible interval is to form the equal-tailed interval, i.e. $I_\alpha = [\theta_{\alpha/2}, \theta_{1-\alpha/2}]$, where θ_z is the z-quantile of the posterior. However, this interval could be unnecessarily long and not representative of the posterior if the posterior distribution is asymmetric or even multi-modal. A better approach is to construct the Highest Posterior Density (HPD) intervals/sets. A $100(1-\alpha)\%$ HPD set is defined as $I_\alpha = \{\theta : p(\theta|\boldsymbol{x}) \geq k\}$ where

$k = max\{k : \int_{\theta:p(\theta|\boldsymbol{x})\geq k} p(\theta|\boldsymbol{x})d\theta = 1-\alpha\}$. If the posterior is approximately uni-modal, then the HPD set is guaranteed to be the HPD interval.

A real data example

We use the LSAT dataset provided by the *ltm R* package (Rizopoulos, 2015) for demonstration. The dataset was first described in Bock and Lieberman (1970). It contains dichotomous item responses from 1000 students to 5 items that are designed to measure a single latent trait. We fit the Bayesian 1PNO model with the Gibbs sampler described in the previous section. The posterior samples are then transformed back to the more common parameterization as in Equation 1. The posterior summary are in Table 1. The

Table 1

Posterior summary

	Mean	SD	lower	upper
b_1	-3.626	0.337	-4.299	-3.049
b_2	-1.405	0.157	-1.716	-1.128
b_3	-0.349	0.107	-0.569	-0.146
b_4	-1.823	0.188	-2.202	-1.500
b_5	-2.857	0.270	-3.405	-2.368
a	0.432	0.040	0.349	0.504

first two columns are posterior means and posterior standard deviations. The last two columns are the lower limit and the upper limit of the 95% HPD intervals.

Posterior predictive model checking

A statistical model is essentially a set of assumptions. Before any inferences are drawn from the model, it is important to assess the fit of the model or, in other words, to

check the degree to which the model fails to explain the data that are of our practical interest. This is particularly important to psychometrics where we are interested in inferring or measuring some theoretical attributes (e.g. intelligence) from observable quantities (e.g. responses to test items). A test item may not function as intended, and the subject may not respond to items as assumed by the model. Before making any conclusions about the attribute of interest, it is critical to evaluate the assumptions. Model checking procedures can help identify unreasonable assumptions which may be potentially mitigated in some way later.

The basic intuition of the PPMC method is that we can check the observed data against the replicated data predicted by the model using some statistics. The distribution of the replicated data is also referred to as the posterior predictive distribution, i.e.

$$p(\boldsymbol{x}^{rep}|\boldsymbol{x}) = \int p(\boldsymbol{x}^{rep}|\boldsymbol{\zeta})p(\boldsymbol{\zeta}|\boldsymbol{x})d\boldsymbol{\zeta}, \tag{5}$$

where $\boldsymbol{\zeta}$ is a vector of all model parameters and $p(\boldsymbol{\zeta}|\boldsymbol{x})$ is the posterior distribution. Except for some simple models, finding the posterior predictive distribution requires sampling in general.

A statistic is a function of the data. Different statistics may describe different aspects of the data. Thus, it is important to choose appropriate statistics that reflect aspects of the model that are of interest to us. In this chapter, we follow the suggestions given by Sinharay (2005) and demonstrate the PPMC method using some most common and important statistics.

Observed sum scores

In many tests, ranking students is one of the most important objective. A model that does not predict the number of correct responses given by students well is very unlikely could rank students fairly. For the 5 dichotomous items in the example, there are 6 possible score categories. We observe a number of students within each score category. So the observed distribution of sum scores are $\boldsymbol{NC} = (NC_0, NC_1, \ldots, NC_5)$. Using the

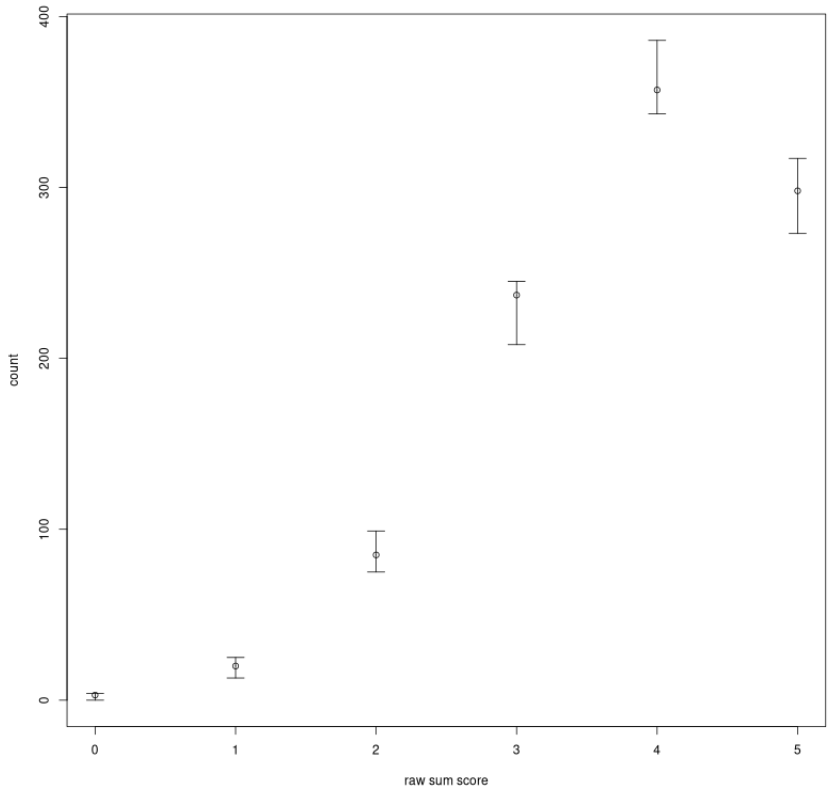

Figure 3

The vertical bars represent the 80% *HPD intervals of the predicted frequency distributions at each sum score, and the circles are the observed frequency of each sum score.*

posterior samples produced by the Gibbs sampler, we can generate replicated datasets and compute the posterior predictive distribution of the sum scores. The result is in Figure 3 The observed sum scores all fall within the 80% HPD intervals of the posterior predictive distributions. For a given set of observed data and a chosen model, the choice of the level of HPD intervals has an effect on their lengths. The smaller percentages are associated with shorter intervals, therefore representing a more strict criteria for evaluating model fit. The 80% level is the default in the *Stan* software which is widely used by applied researchers. In our case, The model does seem to be able to explain the distribution of number of correct responses well.

Polychoric correlation coefficient

An important assumption of the 1PNO IRT model is that discrimination parameters across all items are equal. To check this assumption we use the polychoric correlation coefficient. A polychoric correlation coefficient describes the correlation between the normal variables derived from two observed ordinal variables - in our case, the total number of correct responses of a student and that student's item response to the j^{th} item. If the model is reasonable, then we would expect to see they polychoric correlations for all J items are similar or the variance of the polychoric correlations is close to 0. The results in Figure 4 shows the variance of the observed polychoric correlations are very close to 0

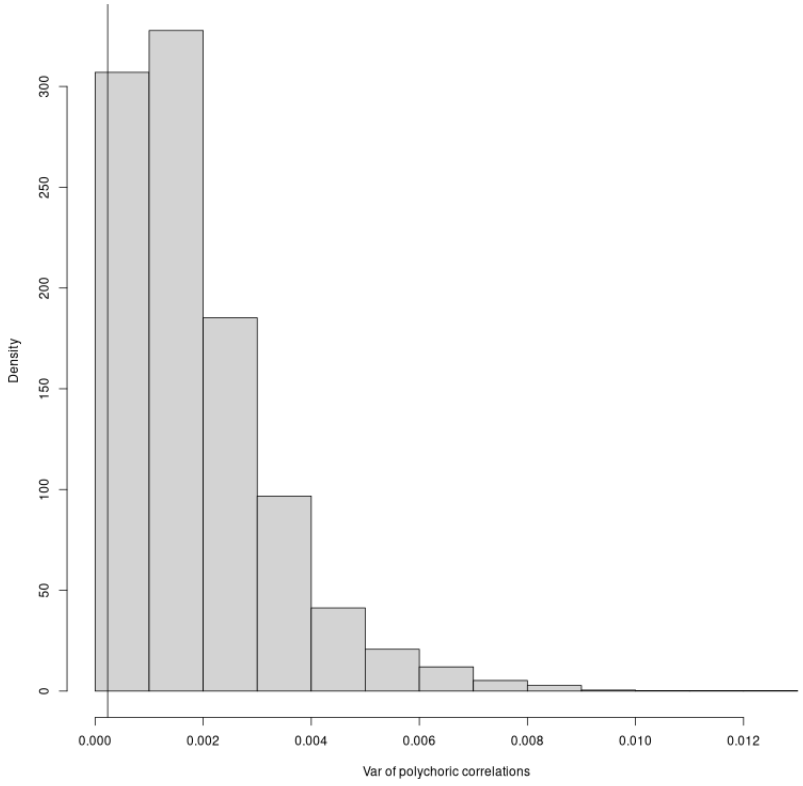

Figure 4

The histogram describes the posterior predictive distribution of the variance of the polychoric correlations of the 5 items. The vertical line is the observed variance of the polychoric correlations. The observed is in a high posterior density region.

and within a high density area of the posterior predictive distribution of the same statistic. Consequently, there is no evidence of this aspect of misfit of the model.

Odds ratio

The 1PNO IRT model assumes that item responses are independent of each other given latent abilities $\boldsymbol{\theta}$. This is also commonly referred to as the local independence (LI) assumption. If LI holds, the model should predict the odds ratio of any item pairs,

$$OR = \frac{n_{00}n_{11}}{n_{01}n_{10}}, \tag{6}$$

where $n_{kk'}$ denotes the number of students has a score k on the first item and a score k' on the second item. For 5 items, there are $\binom{5}{2} = 10$ pairs. For demonstration purpose, we plot

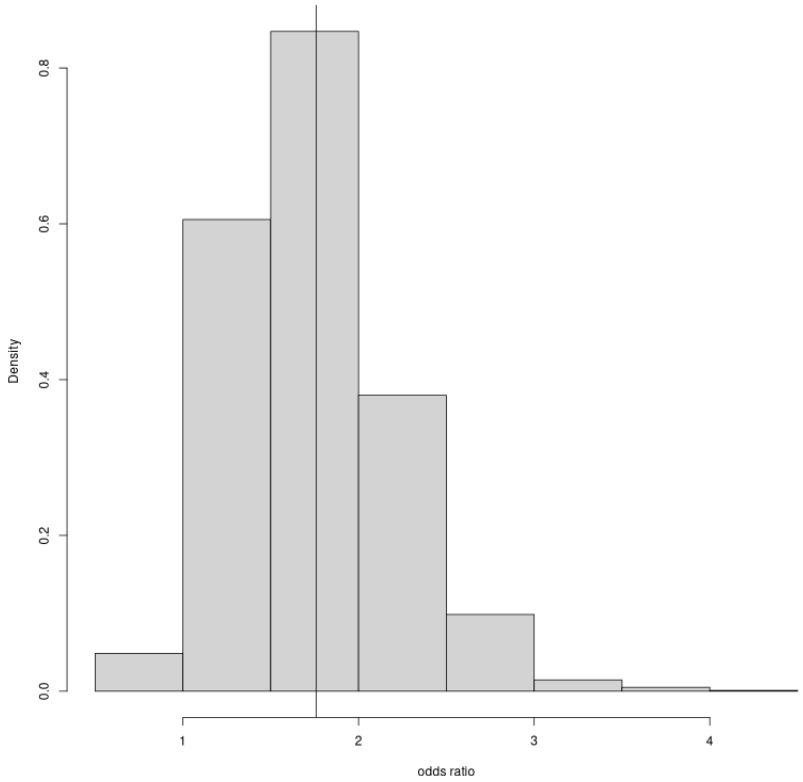

Figure 5

The histogram describes the posterior predictive distribution of the odds ratio of an item pair. The vertical line marks the observed odds ratio of that item pair.

the results of one item pair in Figure 5. The observed OR is within a high density region of the posterior predictive distribution. The examined statistic does not show evidence of violation of the LI assumption for this pair of items.

Conclusion

Thanks to the development of general purpose Bayesian inference software such as the Stan, applied researchers now has access to a set of powerful tools to develop and estimate complex and sophisticated Bayesian models. While the software can help automate the Bayesian computation, it should be emphasized that all three steps outlined earlier in the entry are important and integral parts of the Bayesian modeling. Especially, the importance of evaluating model fit should not be overlooked. A model is a set of assumptions. Drawing inferences based on assumptions without checking them can be dangerous. Furthermore, the Bayesian modeling is really a continuous process where researchers iterate through the three steps. After the aspects of model-data misfits are identified in model fit evaluations, the model can be expanded and improved. In this entry, we offered a high level introduction to the basics of Bayesian psychometrics. For a more detailed treatment of these introductory topics, readers may refer to Gelman et al. (2013) and Levy and Mislevy (2016).

References

Bock, D. R., & Lieberman, M. (1970). Fitting a response model for n dichotomously scored items. *Psychometrika*, *35*(2), 179–197. https://doi.org/10.1007/BF02291262

Cowles, M. K., & Carlin, B. P. (1996). Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review. *Journal of the American Statistical Association*. https://doi.org/10.1080/01621459.1996.10476956

Culpepper, S. A. (2015). Bayesian Estimation of the DINA Model With Gibbs Sampling. *Journal of Educational and Behavioral Statistics*, *40*(5), 454–476. https://doi.org/10.3102/1076998615595403

DeCarlo, L. T. (2012). Recognizing Uncertainty in the Q-Matrix via a Bayesian Extension of the DINA Model. *Applied Psychological Measurement*, *36*(6), 447–468. https://doi.org/10.1177/0146621612449069

Fox, J. P., & Glas, C. A. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika*, *66*(2). https://doi.org/10.1007/BF02294839

Gelman, A., Lee, D., & Guo, J. (2015). Stan: A Probabilistic Programming Language for Bayesian Inference and Optimization. *Journal of Educational and Behavioral Statistics*, *40*(5), 530–543.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian Data Analysis*. Chapman; Hall/CRC.

Gelman, A., & Rubin, D. B. (1992). Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science*. https://doi.org/10.1214/ss/1177011136

Geman, S., & Geman, D. (1984). Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *PAMI-6*(6), 721–741. https://doi.org/10.1109/TPAMI.1984.4767596

Hastings, W. K. (1970). Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika*, *57*(1), 97–109. http://www.jstor.org/stable/2334940

Johnson, M. S., & Junker, B. W. (2003). Using Data Augmentation and Markov Chain Monte Carlo for the Estimation of Unfolding Response Models. *Journal of Educational and Behavioral Statistics*, *28*(3), 195–230. https://doi.org/10.3102/10769986028003195

Levy, R. (2009). The Rise of Markov Chain Monte Carlo Estimation for Psychometric Modeling. *Journal of Probability and Statistics*, *2009*, 1–18. https://doi.org/10.1155/2009/537139

Levy, R., & Mislevy, R. J. (2016). *Bayesian Psychometric Modeling* (1st edition). Chapman; Hall/CRC.

Liu, X. (2019). *Three Contributions to Latent Variable Modeling* (Doctoral Dissertation). Columbia University.

Liu, X., & Johnson, M. S. (2019). Estimating CDMs Using MCMC. In M. von Davier & Y.-S. Lee (Eds.), *Handbook of Diagnostic Classification Models* (pp. 629–646). Springer.

Maris, G., Bechger, T., & San Martin, E. (2015). A Gibbs Sampler for the (Extended) Marginal Rasch Model. *Psychometrika*, *80*(4), 859–79. https://doi.org/10.1007/s11336-015-9479-4

Neal, R. M. (1998). Probabilistic Inference Using Markov Chain Monte Carlo Methods. *Technical Report*, *1*, 1–144. https://doi.org/10.1021/np100920q

Öztürk, N. K., & Karabatsos, G. (2017). A Bayesian Robust IRT Outlier-Detection Model. *Applied Psychological Measurement*, *41*(3), 195–208. https://doi.org/10.1177/0146621616679394

Rizopoulos, D. (2015). Ltm : An R Package for Latent Variable Modeling and Item Response Theory Analyses. *Journal of Statistical Software*, *17*(5), 1–25. https://doi.org/10.18637/jss.v017.i05

Rubin, D. B. (1984). Bayesianly Justifiable and Relevant Frequency Calculations for the Applied Statistician. *The Annals of Statistics*, *12*(4), 1151–1172. https://doi.org/10.1214/aos/1176346785

Sinharay, S. (2015). Assessment of Person Fit for Mixed-Format Tests. *Journal of Educational and Behavioral Statistics*, *40*(4), 343–365. https://doi.org/10.3102/1076998615589128

Sinharay, S. (2005). Assessing Fit of Unidimensional Item Response Theory Models Using a Bayesian Approach. *Journal of Educational Measurement*, *42*(4), 375–394. https://doi.org/10.1111/j.1745-3984.2005.00021.x

Sinharay, S., & Almond, R. G. (2007). Assessing Fit of Cognitive Diagnostic Models A Case Study. *Educational and Psychological Measurement*, *67*(2), 239–257. https://doi.org/10.1177/0013164406292025