Evaluating solutions to the label-switching issue when estimating latent variable models with the NUTS algorithm

Nathan DePuy, Jonathan Templin (Sponsor)

Background

Bayesian estimation is a useful technique for estimating item response theory (IRT) models. Compared to frequentist methods, Bayesian estimation methods incorporate prior information into the estimation process as well as quantification of uncertainty in the true value approximations of modeled parameters.

Markov Chain Monte Carlo (MCMC) methods provide a framework for the efficient sampling of parameter values in Bayesian estimation. In particular, the No-U-Turn Sampler (NUTS) serves as an efficient algorithm (e.g., Hoffman & Gelman, 2014) to sample parameters in IRT models. Bayesian estimation using the NUTS sampling algorithm can lead to convergence issues, however, even when sampling from basic item response models.

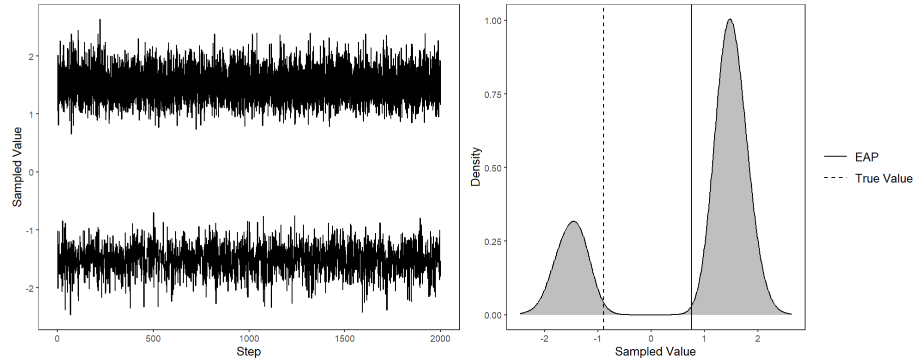
One problem that is encountered regularly in using NUTS for IRT model estimation is that of *label-switching*–the permutation of meaning of some of the parameters (e.g., Qiu & Yuan, 2023). Label-switching in NUTS occurs as a result of a set of MCMC chains converging at differing modes in posterior densities. As a result, convergence statistics look poor and timeseries plots of parameter sampling for the permuted parameters are reflected across zero. Figure 1 shows this phenomenon following the estimation of a basic IRT model with simulated data. This study investigates methods for limiting the label-switching issue when using NUTS for IRT-based models.

In IRT models, there are two modes of the posterior distribution, coming from the form of the item response function. Take, for instance, the two-parameter logistic model (from discrimination/difficulty to slope/intercept form):

$$\operatorname{logit}\left[P\left(Y_{pi}=1 \mid \theta_p\right)\right] = a_i\left(\theta_p - b_i\right) = -a_i b_i + a_i\theta_p = \mu_i + \lambda_i\theta_p \tag{1}$$

where λ_i is a measure of item i's discrimination and θ_p is a measure of the underlying latent trait. The label-switching issue comes from the equivalence of the model when the discrimination/loading parameter and the latent variable are both positive or both negative:

$$(-\lambda_i) \times (-\theta_p) = (\lambda_i) \times (\theta_p) \tag{2}$$



(a) timeseries for 4 chains

(b) posterior density

Figure 1: Visual examples of bimodality in sampled item discrimination values

Since two solutions are equally possible, MCMC chains can be drawn into multiple modes in the geometry of the resulting posterior distribution and thus, appear to not converge to the same posterior distribution.

In the Bayesian IRT literature, constraints are typically placed on the item discrimination (λ_i) parameters, such that only positively signed values are sampled (e.g., König, Spoden, & Frey, 2022; Gelman & Hill, 2006; Curtis, 2010). In turn, a strict monotonicity assumption is imposed on the relationship between the predicted log-odds of the probability of item endorsement, which, in complex modeling contexts (or in cases where some observed variables are weak indicators of the latent variable), may not be a valid assumption. Thus, a solution that allows for negatively-signed item discrimination values is needed when using the NUTS algorithm to sample true item discrimination values.

Methods

A simulation study was conducted to determine which combination of starting (parameter initialization) values or model specification may help to avoid label switching. There were three crossed experimental factors: (1) the model for simulated data, (2) a set of choices for parameter initialization values, and (3) specification of model constraints or parameterizations.

Simulated Data Models

Two data generating models were used:

1. Unidimensional 2PL model (as from 1)

2. A bifactor model with one general trait and two specific traits

Simulated response data was generated using known true parameter values, given:

$$\begin{aligned} \theta_p &\sim \mathcal{N}(0,1) \\ \lambda_i &\sim \mathcal{U}(-3,3) \\ \tau_i &\sim \mathcal{U}(-3,3) \end{aligned} \tag{3}$$

For the bifactor model, item discrimination parameters (λ_i) were generated for loadings onto a general factor (θ_G) as well as for loadings onto two dimensions of the general factor ($\theta_{g_1}, \theta_{g_2}$).

Starting Values

1. item parameters initialized using augmented *expected a posteriori* obtained with the *automatic differentiation variational inference* (ADVI) algorithm (Kucukelbir, Tran, Ranganath, Gelman, & Blei, 2016), latent trait measurements randomly initialized from $\mathcal{U}(-6,6)$

2. item parameters and latent trait measurements initialized on random draws from $\mathcal{U}(-6,6)$

3. latent trait measurements initialized on standardized sum scores, item parameters initialized on random draws from $\mathcal{U}(-6,6)$

Model Parameterization

1. Empirical normal prior placed on sampled item discrimination parameters, such that:

$$\lambda_i \sim \mathcal{N}(\mu_\lambda, \sigma^2_\lambda), \quad \mu_\lambda > 0 \tag{4}$$

2. Empirical truncated normal prior placed on sampled item discrimination values with a lower bound parameter α, such that:

$$\lambda_i \sim \text{TruncNorm}(\alpha, \mu_\lambda, \sigma^2_\lambda), \quad \lambda_i \geq \alpha \tag{5}$$

where μ_λ and σ^2_λ are freely estimated parameters in both approaches.

Each combination of approaches was then evaluated on the rate of chain convergence using $\hat{R}$ statistics, a diagnostic value calculated from the variance of sampled parameter values within- and between-chains after a period of warm-up/burn-in iterations (Vehtari, Gelman, Simpson, Carpenter, & Bürkner, 2021). In cases where label switching occurs, $\hat{R}$ values are expected to exceed $\hat{R}$ values of 1.05 and was used to evaluate the effectiveness of solution combinations.

The *cmdstanr* package (Gabry, Češnovar, Johnson, & Bronder, 2024), a command-line interface for the probabilistic programming language *Stan* (Stan Development Team, 2024) for use in *R*, was used for sampling.

The results from this study may be used when estimating item response models where the monotonicity assumption imposed on the relationship between the latent trait and the log-odds of item endorsement is relaxed. Further, the results of the study demonstrate the performative success across several methods. Finally, the reported rate of model convergence informs the success of the proposed solutions so that they may be used in applied item response modeling research.

References

Curtis, S. M. (2010). Bugs code for item response theory. *Journal of Statistical Software, Code Snippets*, *36*(1), 1–34. Retrieved from `https://www.jstatsoft.org/index.php/jss/article/view/v036c01` doi: 10.18637/jss.v036.c01

Gabry, J., Češnovar, R., Johnson, A., & Bronder, S. (2024). cmdstanr: R interface to 'cmdstan' [Computer software manual]. Retrieved from `https://mc-stan.org/cmdstanr/` (R package version 0.8.1, https://discourse.mc-stan.org)

Gelman, A., & Hill, J. (2006, 11). Data analysis using regression and multilevel/hierarchical models. In (Vol. 3). doi: 10.1017/CBO9780511790942

Hoffman, M. D., & Gelman, A. (2014). The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research*, *15*(1), 1593–1623.

Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A., & Blei, D. M. (2016). *Automatic differentiation variational inference.* Retrieved from `https://arxiv.org/abs/1603.00788`

König, C., Spoden, C., & Frey, A. (2022). Robustness of the performance of the optimized hierarchical two-parameter logistic irt model for small-sample item calibration. *Behavior Research Methods*. doi: 10.3758/s13428-022-02000-5

Qiu, M., & Yuan, K. (2023). *Label switching in latent class analysis: Accuracy of classification, parameter estimates, and confidence intervals.* doi: 10.1080/10705511.2023.2213842

Stan Development Team. (2024). Stan modeling language users guide and reference manual [Computer software manual]. Retrieved from `https://mc-stan.org/` (Version 2.35.0)

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Bürkner, P.-C. (2021). Rank-normalization, folding, and localization: An improved $\widehat{R}$ for assessing convergence of mcmc. , *16*. doi: 10.1214/20-BA1221