Bootstrap-Calibrated Interval Estimates for Scale Scores in Item Response Theory

Yang Liu\textsuperscript{1} \quad Ji Seung Yang\textsuperscript{2}

\textsuperscript{1}University of California, Merced
\textsuperscript{2}University of Maryland, College Park

July 15, 2016

IMPS 2016, Asheville, NC
IRT scoring in two stages

1. Calibration: Parameters of the scoring model are estimated from a calibration sample*
   - Maximum likelihood (ML) estimation

2. Scoring: Point and interval estimates of scores are calculated from the estimated posterior of the latent trait given the response pattern to be scored**
   - Point estimates: Expected/maximum a posteriori (EAP/MAP) scores
   - Interval estimates: Wald-type or quantile-based intervals

* Other limited- and full-information estimation methods are not considered here.
** Other scoring methods, such as ML, are not considered here. Our discussions can be extended to posterior-based scoring based on summed-scores.
Two sources of error in IRT scores

1 Measurement error
   - Test items are imperfect indicators of the latent variable
   - Even when the true model parameters are used, the variability of the posterior of the latent trait is still non-zero
   - Test length ↑, measurement error ↓

2 Sampling/calibration error
   - Estimated model parameters ≠ true parameters
   - Calibration sample size ↑, sampling error ↓
In practice...

- We often resort to the two-stage plug-in method
  1. Estimate IRT model parameters via ML
  2. Evaluate the posterior at the estimated model parameters
- The plug-in method ignores the sampling error (Cheng & Yuan, 2010; Yang, Hansen, & Cai, 2012), which is problematic when
  - The calibration sample size is small
  - The scoring model is complex
  - The true item parameters are extreme


Objectives

1. Obtain a better estimate of the true posterior using bootstrap calibration (BC)
2. Compare BC and plug-in interval estimates via Monte Carlo simulations
3. Propose a BC-based method that properly characterizes both sampling and measurement error
True and estimated posteriors

- Compare cumulative distribution functions (cdfs)

True posterior

Area = ?

Estimated posterior

Area = $\alpha$

$\alpha$th quantile
An alternative interpretation

- Predict plausible values generated from the true posterior

True posterior

Estimated posterior

Proportion of plausible values covered by the interval

Area = $\alpha$

100$\alpha$% one-sided interval estimate
One-sided plug-in prediction intervals (PIs)

- \( H_\xi(\theta|y^*) \): The posterior cdf
  - \( \xi \): Model parameters
    - \( \xi_0 \): True values
    - \( \hat{\xi} \): ML estimates
  - \( \theta \): The latent variable*
  - \( y^* \): The response pattern to be scored

- 100\( \alpha \)% one-sided plug-in PI: \( \left[ -\infty, H_{\hat{\xi}}^{-1}(\alpha|y^*) \right] \)
  - \( H_{\hat{\xi}}^{-1}(\alpha|y^*) \): The \( \alpha \)th quantile of the plug-in posterior cdf

*Only unidimensional IRT models are considered here for notational simplicity.
One-sided plug-in PIs

- Using our notation

True posterior

Plug-in posterior

\[ H_{\xi_0}(H_{\xi}^{-1}(\alpha|y^*)) \]

Area = \( \alpha \)

\([-\infty, H_{\xi}^{-1}(\alpha|y^*)]\)

\[ H_{\xi}^{-1}(\alpha|y^*) \]

IRT score interval

Bootstrap calibration
Predictive coverage

- Define the *coverage probability of plausible values* for one-sided plug-in PIs at nominal level $\alpha$

\[ C_{\xi_0}(\alpha) = E_{\xi_0}^Y \left[ H_{\xi_0} \left( H^{-1}_{\hat{\xi}}(\alpha|y^*) \big| y^* \right) \right] \]

- $E_{\xi_0}^Y$: Expectation over repeated sampling of the calibration data $Y$

- $C_{\xi_0}(\alpha)$ can be estimated by Monte Carlo simulations

- Since $\hat{\xi}$ is consistent for $\xi_0$, $C_{\xi_0}(\alpha) \rightarrow \alpha$ as the sample size $n \rightarrow \infty$
Calibrating the plug-in PI

- Under mild regularity conditions (Beran, 1990),

\[ C_{\xi_0}(\alpha) = \alpha + O\left(\frac{1}{n}\right) \]

- The order of the bias term can be improved via calibration

- At nominal level \( C_{\xi_0}^{-1}(\alpha) \), the one-sided plug-in PI has coverage \( \alpha \), because \( C_{\xi_0} \left( C_{\xi_0}^{-1}(\alpha) \right) = \alpha \)
  - But \( \xi_0 \) is unknown in practice

- Calibration: Use nominal level \( C_{\hat{\xi}}^{-1}(\alpha) \)
  - The bias term is \( o(1/n) \)

Calibration via parametric bootstrap

- Monte Carlo approximations of $C_{\hat{\xi}}(\alpha)$ and $C_{\hat{\xi}}^{-1}(\alpha)$ can be obtained in a fashion similar to the approximation of $C_{\xi_0}(\alpha)$
- The only exception is that the calibration data should be generated using estimated parameters $\hat{\xi}$, i.e., parametric bootstrap
- We use 500 bootstrap samples in our simulations*

*The ML estimation of model parameters may fail to converge in some bootstrap samples; those occasions were simply excluded from the calculation.
3PL model, constant guessing, \( n = 500 \)

- Compare Plug-in and BC one-sided PIs
- Nominal level \( \alpha = 0.025, 0.05, \ldots, 0.975 \)
- True item parameters: 36 binary items

<table>
<thead>
<tr>
<th></th>
<th>Intercept</th>
<th>1.8</th>
<th>0.9</th>
<th>−2.0</th>
<th>−1.1</th>
<th>2.4</th>
<th>1.4</th>
<th>( \cdots )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Slope</td>
<td>0.5</td>
<td>0.7</td>
<td>1.0</td>
<td>1.3</td>
<td>1.7</td>
<td>2.3</td>
<td>( \cdots )</td>
</tr>
<tr>
<td></td>
<td>Guessing</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
<td>( \cdots )</td>
</tr>
</tbody>
</table>

- Response patterns were generated at \( \theta_0 = -2, -1.5, \ldots, 2 \)
- 500 replications

- R package \texttt{mirt} (Chalmers, 2012) was used for estimation

3PL model, constant guessing, $n = 500$

**Graph Description:**
- The graph plots the coverage of plausible values against the nominal level $\alpha$ for different values of $\theta_0$ ranging from $-2$ to $2$.
- Two lines are shown: one for BC (normal) and one for Plug-in (normal).
- The coverage values are indicated on the y-axis, ranging from $-0.03$ to $0.03$.
- The x-axis represents the nominal level $\alpha$, with values ranging from $0.1$ to $0.9$.

**Key Observations:**
- As $\theta_0$ increases from $-2$ to $2$, the coverage remains relatively stable.
- The BC (normal) line generally shows a slight increase in coverage compared to the Plug-in (normal) line.

**Conclusion:**
The graph illustrates the effectiveness of using a constant guessing model in the 3PL framework, showing stable coverage across different $\theta_0$ values and nominal levels $\alpha$.
Choice of latent variable density (prior)

- So far, $\mathcal{N}(0,1)$ has been used as the prior in the posterior calculation, which is often done in practice.
- Since $\mathcal{N}(0,1)$ shrinks the posterior towards 0, the corresponding PIs often do not account for measurement error properly when the test is short.
- The coverage of the true latent variable score $\theta_0$ can be very different from the nominal level.
3PL model, constant guessing, \( n = 500 \)

Coverage of \( \theta_0 - \alpha \)

\( \theta_0 = -2 \) \quad \theta_0 = -1.5 \quad \theta_0 = -1 \quad \theta_0 = -0.5 \quad \theta_0 = 0 \quad \theta_0 = 0.5 \quad \theta_0 = 1 \quad \theta_0 = 1.5 \quad \theta_0 = 2 \)

Nominal level \( \alpha \)

IRT score interval | Coverage of true scale scores
--- | ---
0.1 0.5 0.9 | 0.1 0.5 0.9
0.1 0.5 0.9 | 0.1 0.5 0.9
0.1 0.5 0.9 | 0.1 0.5 0.9
0.1 0.5 0.9 | 0.1 0.5 0.9
0.1 0.5 0.9 | 0.1 0.5 0.9
0.1 0.5 0.9 | 0.1 0.5 0.9
0.1 0.5 0.9 | 0.1 0.5 0.9
0.1 0.5 0.9 | 0.1 0.5 0.9

BC (normal) | Plug-in (normal)
Jeffreys’ prior

- Jeffreys’ prior $\pi_{\xi_0}(\theta) \propto \sqrt{i_{\xi_0}(\theta)}$
  - $i_{\xi_0}(\theta)$: Test information function evaluated at $\xi_0$

- For continuous data and in the absence of nuisance parameters, Jeffreys’ prior is **first-order probability matching**, i.e., the discrepancy between posterior and coverage probabilities is of order $1/n$ (Welch & Peers, 1963)

- Although Jeffreys’ prior is not exactly first-order matching for discrete data, we conjecture that it can improve the coverage of $\theta_0$

---

**3PL model, constant guessing, \( n = 500 \)**

- Jeffreys’ prior was used instead of \( \mathcal{N}(0, 1) \)
- Nominal level \( \alpha = 0.025, 0.05, \ldots, 0.975 \)
- True item parameters: 36 binary items

<table>
<thead>
<tr>
<th></th>
<th>Intercept</th>
<th>1.8</th>
<th>0.9</th>
<th>-2.0</th>
<th>-1.1</th>
<th>2.4</th>
<th>1.4</th>
<th>( \cdots )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Slope</td>
<td>0.5</td>
<td>0.7</td>
<td>1.0</td>
<td>1.3</td>
<td>1.7</td>
<td>2.3</td>
<td>( \cdots )</td>
<td></td>
</tr>
<tr>
<td>Guessing</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
<td>( \cdots )</td>
<td></td>
</tr>
</tbody>
</table>

- Response patterns were generated at \( \theta_0 = -2, -1.5, \ldots, 2 \)
- 500 replications
3PL model, constant guessing, \( n = 500 \)

Coverage of \( \theta_0 - \alpha \)

\( \theta_0 = -2 \)  \( \theta_0 = -1.5 \)  \( \theta_0 = -1 \)  \( \theta_0 = -0.5 \)  \( \theta_0 = 0 \)  \( \theta_0 = 0.5 \)  \( \theta_0 = 1 \)  \( \theta_0 = 1.5 \)  \( \theta_0 = 2 \)

BC (Jeffreys)  Plug-in (Jeffreys)  Plug-in (normal)

IRT score interval  Simulation: Part 2
3PL model, constant guessing, \( n = 500 \)

Coverage of \( \theta_0 - \alpha \)

- \( \theta_0 = -2 \)
- \( \theta_0 = -1.5 \)
- \( \theta_0 = -1 \)
- \( \theta_0 = -0.5 \)
- \( \theta_0 = 0 \)
- \( \theta_0 = 0.5 \)
- \( \theta_0 = 1 \)
- \( \theta_0 = 1.5 \)
- \( \theta_0 = 2 \)

Nominal level \( \alpha \)

IRT score interval

Simulation: Part 2
3PL model, constant guessing, \( n = 500 \)

![Graph showing coverage of plausible values for different \( \theta_0 \) values.](image)

- BC (Jeffreys)
- Plug–in (Jeffreys)

IRT score interval

Simulation: Part 2
Summary and future directions

- Summary of findings
  - Bootstrap calibration yields a better estimate of the true posterior
  - Using Jeffreys’ prior leads to better coverage of the true scale score

- Future directions
  - Extensions to multidimensional IRT models
  - Using bootstrap calibration for other inferential purposes
Thanks!