Nonparametric Bayesian Methods for Item Response Theory

Steven MacEachern – The Ohio State University

Kristin Duncan – San Diego State University

Outline

• A motivating example
• Nonparametric Bayesian models for IRT
• Analysis of exam data
• Identifiability of models
• Wrap-up

Conference Honoring the Achievements of Michael Browne
Supported by grants from the National Science Foundation and the National Security Agency
A motivating example

• Introductory statistics class, final exam
  – Ohio State University, roughly 50,000 students
  – Required for a B.A. degree – a course on data analysis
  – Large lectures, parallel sections

• An evaluation system is needed. For a multiple choice exam
  – How should we score the exam?
  – How should we best design the exam?
  – What sort of accuracy can we expect in grading?

• Later, an analysis of final exam data in such a course
  – 258 students
  – 28 multiple choice questions
  – Each question graded right/wrong
The exam data

- Total scores. Note generally high scores, 5 perfect scores.

<table>
<thead>
<tr>
<th>Raw Score</th>
<th>Count</th>
<th>Raw Score</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>15</td>
<td>8</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>16</td>
<td>22</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>17</td>
<td>16</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>18</td>
<td>21</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>19</td>
<td>21</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>20</td>
<td>28</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>21</td>
<td>30</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>22</td>
<td>21</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>23</td>
<td>23</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>24</td>
<td>13</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>25</td>
<td>14</td>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td>26</td>
<td>7</td>
</tr>
<tr>
<td>12</td>
<td>3</td>
<td>27</td>
<td>9</td>
</tr>
<tr>
<td>13</td>
<td>7</td>
<td>28</td>
<td>5</td>
</tr>
</tbody>
</table>
The exam data

- Scores by question. Note range of difficulties.

<table>
<thead>
<tr>
<th>Item</th>
<th>Score</th>
<th>Percent</th>
<th>Item</th>
<th>Score</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>247</td>
<td>95.7</td>
<td>15</td>
<td>210</td>
<td>81.4</td>
</tr>
<tr>
<td>2</td>
<td>113</td>
<td>43.8</td>
<td>16</td>
<td>223</td>
<td>86.4</td>
</tr>
<tr>
<td>3</td>
<td>197</td>
<td>76.4</td>
<td>17</td>
<td>163</td>
<td>63.2</td>
</tr>
<tr>
<td>4</td>
<td>200</td>
<td>77.5</td>
<td>18</td>
<td>92</td>
<td>35.7</td>
</tr>
<tr>
<td>5</td>
<td>217</td>
<td>84.1</td>
<td>19</td>
<td>255</td>
<td>98.8</td>
</tr>
<tr>
<td>6</td>
<td>239</td>
<td>92.6</td>
<td>20</td>
<td>180</td>
<td>69.8</td>
</tr>
<tr>
<td>7</td>
<td>244</td>
<td>94.6</td>
<td>21</td>
<td>155</td>
<td>60.1</td>
</tr>
<tr>
<td>8</td>
<td>132</td>
<td>51.2</td>
<td>22</td>
<td>159</td>
<td>61.6</td>
</tr>
<tr>
<td>9</td>
<td>119</td>
<td>46.1</td>
<td>23</td>
<td>174</td>
<td>67.4</td>
</tr>
<tr>
<td>10</td>
<td>181</td>
<td>70.2</td>
<td>24</td>
<td>173</td>
<td>67.1</td>
</tr>
<tr>
<td>11</td>
<td>220</td>
<td>85.3</td>
<td>25</td>
<td>202</td>
<td>78.3</td>
</tr>
<tr>
<td>12</td>
<td>85</td>
<td>32.9</td>
<td>26</td>
<td>245</td>
<td>95.0</td>
</tr>
<tr>
<td>13</td>
<td>222</td>
<td>86.0</td>
<td>27</td>
<td>224</td>
<td>86.8</td>
</tr>
<tr>
<td>14</td>
<td>166</td>
<td>64.3</td>
<td>28</td>
<td>140</td>
<td>54.3</td>
</tr>
</tbody>
</table>
Item response theory

- Traditional forms for item characteristic curves are
  - The Rasch model
    \[ P(Y = 1|\theta) = \frac{\exp[a(\theta-b_j)]}{1+\exp[a(\theta-b_j)]} \]
  - The two-parameter logistic model
    \[ P(Y = 1|\theta) = \frac{\exp[a_j(\theta-b_j)]}{1+\exp[a_j(\theta-b_j)]} \]
  - The three-parameter logistic model
    \[ P(Y = 1|\theta) = c_j + (1-c_j)\frac{\exp[a_j(\theta-b_j)]}{1+\exp[a_j(\theta-b_j)]} \]
Nonparametric Bayesian models

• Two main types of models
  – Random effects; Flexible cdfs

• Type I: Nonparametric ability distribution
  – Abilities may not be normally distributed
  – Selected population of test-takers
  – Replace $F_\theta = N(0, 1)$ with $F_\theta \sim \text{Dir}(MF_{0,\theta})$

• Type II: Nonparametric item characteristic curves
  – ICCs may not be “logistic”
  – No single transformation of ability scale to produce set of logistic ICCs
  – Replace $F_j = \text{logistic}$ with $F_j \sim \text{Dir}(MF_{0,j})$.

• The Dirichlet process is chosen for convenience. It can be replaced by a Polya tree, a Levy process, ...
A nonparametric Bayesian version of the 2PL model

- Type II model – flexible cdfs
- Dirichlet process gives full support

\[
\begin{align*}
\theta_i \mid \mu, \sigma^2 & \sim N(\mu, \sigma^2) \\
a_j \mid \nu, \omega & \sim \text{Gamma}(\nu, \omega) \\
b_j \mid \mu, \sigma^2 & \sim N(\mu, \sigma^2) \\
F_{0j} & = \frac{\exp[a_j(\theta - b_j)]}{1 + \exp[a_j(\theta - b_j)]} \\
F_j \mid a_j, b_j, M_j & \sim \text{Dir}(M_j F_{0j}) \\
z_{ij} \mid F_j & \sim F_j \\
Y_{ij} \mid \theta_i, z_{ij} & = \begin{cases} 
1 & \text{if } z_{ij} \leq \theta_i \\
0 & \text{if } z_{ij} > \theta_i 
\end{cases}
\end{align*}
\]

- \(z_{ij}\) introduced as device for MCMC
- \(Y_{ij}\) a deterministic function of \(\theta_i\) and \(z_{ij}\)
Fitting the model

• Model fit via MCMC
  – Combination of Gibbs steps, M-H steps

• Techniques used to improve estimation
  – Full sample where possible
  – Subsample expensive functionals
  – Benchmark weights for subsamples
  – Combine estimates for identical response patterns
  – Ensure known ordering
    \[ \pi_{\text{post}}(\theta) F_j(\theta)/(1 - F_j(\theta)) \]
  – Mean, not mode for ability estimates
    \( \pi(\theta_1) \) to the right of \( \pi(\theta_2) \) does not imply that
    \( \text{mode}(\pi(\theta_1)) \) is greater than \( \text{mode}(\pi(\theta_2)) \)
Estimated item characteristic curves

- Two curves showing almost no difference between the parametric and nonparametric models
Estimated item characteristic curves

- Two curves showing a difference between the parametric and non-parametric models

- Note that the curves show similar flat spots
Summary of model misfit

- Hosmer-Lemeshow chi-square statistics
  - Groups formed on the basis of estimated probability of correct answer
  - Number of groups smaller for easier questions
- Groups formed either from fit of parametric model or from fit of nonparametric model

<table>
<thead>
<tr>
<th>Item</th>
<th>Groups</th>
<th>2PL Ordering</th>
<th>NP2</th>
<th>2PL Ordering</th>
<th>NP2</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>8</td>
<td>5.436</td>
<td>4.338</td>
<td>13.807</td>
<td>9.959</td>
</tr>
<tr>
<td>12</td>
<td>10</td>
<td>11.508</td>
<td>5.481</td>
<td>29.240</td>
<td>19.880</td>
</tr>
<tr>
<td>13</td>
<td>5</td>
<td>12.506</td>
<td>9.533</td>
<td>12.152</td>
<td>8.541</td>
</tr>
<tr>
<td>18</td>
<td>10</td>
<td>5.600</td>
<td>3.339</td>
<td>15.934</td>
<td>8.992</td>
</tr>
<tr>
<td>23</td>
<td>9</td>
<td>4.497</td>
<td>4.908</td>
<td>2.751</td>
<td>1.378</td>
</tr>
</tbody>
</table>
Estimates of abilities

- Comparisons of parametric, nonparametric, 2PL, 3PL estimates
More on abilities

- Changes in rank of ability estimates
Posterior distributions of abilities

- Blue curve is parametric model, green curve is nonparametric model
- Note relative spreads of posterior distributions
Posterior distributions of abilities

Student 199

Student 231
Recap of benefits

• Better fit to data
  – Chi-squared statistics, based on partitioning $\theta$ scale
  – Extremes (poorly fit questions under par. models) pulled in
  – Others reduced

• Fitted curves appear reasonable
  – ICC estimates are fairly smooth
  – Consistent patterns across questions
  – No single transformation makes all curves logistic

• Reduced variation in posterior for ability estimates
  – Reduction in variance matches fitted ICCs
  – Reduction in variance presumably due to ”better” model
Nested models and identifiability

- Nested models appear in many places in Statistics
  - Underlie F-test of full vs. *reduced* model
  - Likelihood ratio tests
  - Model expansion and model contraction

- However, difficult to find a formal definition
  - Wikipedia, on the likelihood ratio test:

  In statistics, a likelihood ratio test is used to compare the fit of two models, one of which is nested within the other 

  The test requires nested models, that is, models in which the more complex one can be transformed into the simpler model by imposing a set of linear constraints on the parameters.

- A little more sophistication is in order
A classical definition of nested families of models

- From a classical perspective, one might write the following definition
  
  - A family of models, $A$, is nested in a family of models, $B$, if, for every $m \in A$, we can obtain $m$ by setting a function of a collection of parameters in $B$ to some fixed value.

- The focus of the definition is on the parameter space

- Often, comparison reduces to a single overall parameter space with restrictions imposed to create the smaller family
Examples of nested models

• Regression:
  – $m$ in $B$ defined by $\beta_1, \ldots, \beta_p, \sigma^2$
    \[ Y_i \sim N(\sum x_i \beta_i, \sigma^2) \]
    with usual independence assumption
  – $m$ in $A$ defined by $\gamma_1, \ldots, \gamma_k, \tau^2$; covariates are first $k$ principle components
    \[ Y_i \sim N(\sum z_j \gamma_j, \tau^2) \]
    with usual independence assumption
  – To make the match, note that “mean space” for model in $A$ is a subspace of “mean space” for model in $B$
  – Restricting to this mean space, and setting $\sigma^2 = \tau^2$ yields identical models
Examples of nested models

• ANOVA:
  – Fixed effects models
    * Exactly as the regression setting
    * Focus is on “mean space” of the model
  – Random effects models
    * Set random effects variance to 0
    * 0 for the random effects variance implies all random effects are 0
    * This is a restriction on the values of, say, the $\alpha_i$

• Monotone polynomials IRT (Lorraine Sinnott, Longjuan Liang)
  – Restrict the polynomial to be a line
A curious example

• Family $A$ consists of a single model

\[ X_1, X_2, \ldots \sim \text{Bern}(0.5) \]

• Family $B$ is driven by a deterministic dynamical system. The system is indexed by the initial condition.

  – $\theta$, the initial condition, lies in $[0, 1]$.
  – We also call the initial condition $Y_1$.
  – The deterministic system follows the rule

\[
Y_{t+1} = \begin{cases} 
2Y_t & \text{if } 0 \leq Y_t < 0.5 \\
2 - 2Y_t & \text{if } 1/2 \leq Y_t \leq 1
\end{cases}
\]

  – This leads to $\theta = Y_1, Y_2, Y_3, \ldots$.
  – Define $X_t = \text{round}(Y_t)$, leading to $X_1, X_2, \ldots$. 

The tent map

- A picture of the tent map which shows the evolution of the deterministic system. This leads to a deterministic sequence of data.

- The question: Is $A$ nested within $B$?
Calculations

• Compute two probabilities, then compare

\[ P_A(x_1, \ldots, x_n) = \prod_{i=1}^{n} (0.5)^{x_i} (0.5)^{1-x_i} = 2^{-n} \]
\[ P_B(x_1, \ldots, x_n|\theta) = 0 \text{ if } \theta \text{ does not lead to } x_1, \ldots, x_n, \text{ and} \]
\[ P_B(x_1, \ldots, x_n|\theta) = 1 \text{ if } \theta \text{ leads to } x_1, \ldots, x_n \]

• There is no value of \( \theta \) for which

\[ P_A(x_1, \ldots, x_n) = P_B(x_1, \ldots, x_n|\theta) \]

• Since the two differ, the classical answer is that \( A \) is not nested in \( B \)
A Bayesian approach

• Since family $A$ consists of a single model, the Bayesian would retain the same probability calculation for this family.

• For family $B$, one needs a prior distribution over $\theta$

$$P_{B,\pi}(x_1, \ldots, x_n) = \int_0^1 P_B(x_1, \ldots, x_n | \theta) \pi(\theta) d\theta$$

$$= \int_{\text{region}} \pi(\theta) d\theta$$

• where region corresponds to the set of $\theta$ leading to $x_1, \ldots, x_n$

• If $\pi(\theta) = \text{Unif}(0, 1)$, then $P_{B,\pi}(x_1, \ldots, x_n) = 2^{-n}$

• With agreement between the two calculations for all $(x_1, \ldots, x_n)$, the Bayesian finds that family $A$ is contained in family $B$. 
A Bayesian definition of nested families of models

- A family of models, $A$, is nested in a family of models, $B$, if, for every prior distribution $\pi_A$ over the family $A$, there exists some prior distribution $\pi_B$ over the family $B$ such that

$$P_{A,\pi_A}((X_1, \ldots, X_n) \in C) = P_{B,\pi_B}((X_1, \ldots, X_n) \in C)$$

for all measurable $C$ and for all $n$

- This definition repackages the concept of nested models

- Instead of a comparison of points in the parameter spaces, the comparison is of distributions over the parameter spaces

- (Note: one can pass to a different notion of the parameter; that the parameter is a distribution over the conventional parameter space)
A second curious example: Finite n

- Two families of models, $n = 2$
- Family $A$ consists of a single model
  - $\theta \sim \text{Unif}(0, 1)$
  - Given $\theta$, $X_1, X_2 \sim \text{Bern} (\theta)$
- Family $B$ consists of a two point parameter space, with an arbitrary prior distribution over these two points
  - $\Theta = \{0.5 - 1/\sqrt{12}, 0.5 + 1/\sqrt{12}\}$
  - Given $\theta$, $X_1, X_2 \sim \text{Bern}(\theta)$
- For a model of (conditionally) iid Bernoulli trials, the joint distribution of $X_1, \ldots, X_n$ is characterized by the first $n$ moments
  \[
P(X_1, \ldots, X_n | \sum X_i = k) = \theta^k (1 - \theta)^{n-k} \]
  \[
P_\pi(X_1, \ldots, X_n | \sum X_i = k) = \int \theta^k (1 - \theta)^{n-k} \pi(\theta) d\theta \]
- This latter expression is a linear combination of $E[\theta], \ldots, E[\theta^n]$
Second example, continued

- Moments
  - Family A: $E[\theta] = 0.5$, $\text{Var}[\theta] = 1/12$
  - Family B: With $\pi(\theta) = 0.5$ to each of the two points in the parameter space, $E[\theta] = 0.5$, $\text{Var}[\theta] = 1/12$

- Thus, family $A$ is 2-nested in family $B$

- The parameter space and the set of potential prior distributions are both important
  - Seeming reversals of large vs. small are possible

- In realistic situations, we often impose restrictions on the prior distribution
  - $\theta \sim \text{Normal}(\mu, \Sigma)$
  - $\theta \sim \text{Beta}(\alpha, \beta)$
The dynamical system, again

- **Family A**: $X_i | \gamma \sim \text{Bern}(\gamma), \gamma \in (0, 1)$

- **Family B**: Dynamical system model

- **Family A** is $n$-nested in family $B$ for each $n$
  - Construct a prior distribution that matches each $\gamma$
  - Define a distribution over these prior distributions

- **Family B** is not $n$-nested in family $A$
Near-nesting

- Near-nesting is closely associated with nesting and n-nesting
- Modification of the definition of nesting necessitates modification of the concept of near-nesting
- The family of models $A$ is nearly nested in the family $B$ if, for each $\pi_A$ and resulting $P_{A,\pi_A}(X_1, \ldots, X_n)$, there exists $\pi_B$ such that $P_{B,\pi_B}(X_1, \ldots, X_n)$ are arbitrarily close to $P_{A,\pi_A}(X_1, \ldots, X_n)$
- Near-nesting corresponds to traditional definitions of support; Near n-nesting corresponds to traditional definitions of support for a finite collection of observables
• Result: Supp(Parametric) ∈ Supp(Type I nonparametric) ∈ Supp(Type II nonparametric)

• Proof:
  – Examine priors
  – Translate prior distribution into prior over probabilities on cells of $2^J$ table
  – Compare supports of induced prior distributions

• The Type I models are adequate if there is a single transformation of the ability scale that makes the logistic model correct

• The Type II models have full support for the $2^J$ table among all monotone, unidimensional latent trait models

• Since the Type II models have full support, we expect a modest benefit for combining Type I and Type II models
A more refined model allows for guessing

- The two-parameter logistic model

\[ P(Y = 1|\theta) = \frac{\exp[a_j(\theta-b_j)]}{1+\exp[a_j(\theta-b_j)]} \]

- The three-parameter logistic model

\[ P(Y = 1|\theta) = c_j + (1 - c_j)\frac{\exp[a_j(\theta-b_j)]}{1+\exp[a_j(\theta-b_j)]} \]

- The three-parameter logistic model

\[ P(Y = 1|\theta) = \frac{\exp[a_j(\theta-b_j)]}{1+\exp[a_j(\theta-b_j)]} + c_j\frac{1}{1+\exp[a_j(\theta-b_j)]} \]
Nonparametric versions of the models

- Again, two types of nonparametric models
  - Type I: A nonparametric distribution for abilities
  - Type II: Nonparametric ICCs
- Focus on Type II models for both forms of 3PL model
- Novel use of nonparametric model—not quite for a distribution function, as there is mass at \(-\infty\)
- Models fit via MCMC methods
- Here, runs are a bit short, not enough care taken to create good estimators
**Estimated item characteristic curves**

- Two curves showing the two versions of the nonparametric 3PL models

  ![Item 17](image1)
  ![Item 10](image2)

- The green curve is the traditional version; the brown curve is the novel version

- For one item, the curve is essentially parametric; the other curve shows the additional flat spot
Posterior distributions of abilities

- Green curve is traditional version; brown curve is novel version
- Note flat spot for low abilities
- Looking across item characteristic curves, there is little discrimination amongst low-ability students
Conclusions

- Bayesian methods are worthwhile competitors to classical methods
- Nonparametric Bayesian methods provide a means of replacing a component in a hierarchical Bayesian model
  Superior to many classical nonparametric techniques
- Especially effective for latent stages of a model
  Allow us to avoid making strong parametric assumptions about the very parts of the model that are most difficult to check
- The models are relatively easy to fit, though they require MCMC methods and some care in use
- The models place the focus on modelling—both for the structure of the model and for the prior distribution
Ongoing work

- More careful treatment of 3PL models
  - Improvements in R package are in process
  - More formal comparisons of fit of models
  - Description of appropriate test information function

- Extensions abound
  - Ordered categorical responses
  - Multivariate traits (single index; full multivariate)
  - Covariates (dependent Dirichlet processes and extensions)
  - Selection functions / selected populations (Hao Wu)