Bayesian IRT models in NIMBLE
Modelli IRT bayesiani in NIMBLE

Sally Paganin, Chris Paciorek, Perry de Valpine

Abstract IRT models relate observed data to some latent traits typically encoding item characteristics, as well as individual abilities. Often these lasts are assumed to follow a standard normal distribution, but there are situations in which such assumption may be unrealistic. A possible extension for such models uses a Dirichlet process mixture of normal distributions, which is seldom employed in real data analysis due to the lack of guidelines and software tools. We contribute to fill this gap by reviewing both parametric and semiparametric versions of such models. Using 2PL model as an example we also illustrate how these models can be easily implemented via the novel NIMBLE software.

Abstract *I modelli IRT caratterizzano la relazione tra dati osservati e variabili latenti. Quest'ultime solitamente descrivono sia caratteristiche degli item che l'abilità degli individui. Spesso si assume che tali abilità seguano la distribuzione di una normale standard, ma ci sono situazioni in cui tale assunzione non risulta appropiata. In questi casi, una possibile estensione si può ottenere utilizzando una mistura di distribuzioni normali per le abilità latenti basata sul processo di Dirichlet. Tuttavia questi modelli risultano essere poco diffusi in pratica a causa della mancanza di linee guide e software che li implentino. In questo lavoro presentiamo le versioni parametriche e non dei modelli IRT. Usando il modello 2PL come esempio, illustriamo anche come implementarli facilmente attraverso l'uso del software NIMBLE.*

Key words: 2PL, Bayesian nonparametrics, IRT, Dirichlet Process, NIMBLE.

1 Motivation

Item response theory (IRT) refers to a family of models that investigate the relationship between responses to a set of items and some latent traits, typically encoding

Department of Statistics, Department of Enviromental Science, Policy & Management
UC Berkeley e-mail: sally.paganin@berkeley.edu

individual or item characteristics. Such models are employed in different application domains, with educational measurement and psychometrics being the most popular. Models for binary responses are among the most common among IRT models, comprising the one, two or three parameter logistic models (1PL, 2PL and 3PL). These models assume that the probability of a correct answer is related to the individual's latent ability, as well as items difficulty and potentially other item characteristics.

Standard approaches rely on the assumption that latent abilities follow a standard normal distribution. This assumption is sometimes considered for computational convenience, but it may be unrealistic in many situations [9]. For example, [7] gives a comprehensive review of many psychometric datasets where the latent traits distribution does not respect the normality assumption and presents instead asymmetries, heavy-tails or multimodality.

Different proposals have been made in literature relaxing the normality assumption for the latent abilities. Arguably, the most general approach in a Bayesian framework uses a Dirichlet process [4] mixture of normal distributions as a nonparametric distributions for latent abilities. Such models are semiparametric because they retain other, parametric, assumptions of binomial mixed models. Within this approach, the semiparametric 1PL model has been the focus of more effort as well as software tool [6, **DPpackage**]. [10] investigate semi-parametric generalization of Rasch-type models from a theoretical perspective, while [5] provides results from simulation studies considering the 1PL model. An example using the 2PL model is given in [3], but there is a lack of comprehensive studies of such models as well as general tools for model estimation in real data analysis. In this work we review the semiparametric 1PL and 2PL models, and illustrate how to easily implement them in the NIMBLE software.

2 NIMBLE

NIMBLE [2] is a flexible R-based system for hierarchical modeling, which extends BUGS language used in WinBUGS, OpenBUGS, NAGS [8], providing efficient execution of algorithms via custom-generated C++ code. Besides offering new degrees of customization of MCMC algorithms, one of the latest NIMBLE features added support for MCMC inference for Bayesian nonparametric mixture models. In particular, NIMBLE provides functionality for fitting models using a Dirichlet process prior, either via the Chinese Restaurant Process (CRP) [1] or a truncated stick-breaking (SB) [11] representation of the Dirichlet process prior. These features allows Dirichlet process priors to be embedded in very general hierarchical models, supporting extensions of the approaches we illustrate here.

3 IRT models background

In this context, observed data are typically answers to exam questions or items from a set of individuals. Let y_{ij} denote the answer of an individual j to item i for $j = 1,\dots,N$ and $i = 1,\dots,I$, with $y_{ij} = 1$ when the answer is correct and 0 otherwise. Typically, different individuals are assumed to work independently, while responses from the same individuals are assumed independent conditional to the latent trait (*local independence assumption*). Hence each answer y_{ij}, conditionally to the latent parameters, is assumed to be a realization of a Bernoulli distribution, and the probability of a correct response is typically modeled via logistic regression.

In the two-parameter logistic (2PL) model, the conditional probability of a correct response is modeled as

$$\Pr(y_{ij} = 1|\eta_j,\lambda_i,\beta_i) = \frac{\exp\{\lambda_i(\eta_j - \beta_i)\}}{1+\exp\{\lambda_i(\eta_j - \beta_i)\}} \quad i = 1,\dots,I, j = 1,\dots,N, \qquad (1)$$

where η_j represents the latent ability of the j-th individual for $j = 1,\dots,N$, while β_i and λ_i encode the item characteristics for $i = 1,\dots,I$. The parameter λ_i is often referred as *discrimination*, since items with a large λ_i are better at discriminating between subjects with different abilities, while β_i is called *difficulty* because the probability of a correct response is equal to 0.5 when $\eta_j = \beta_i$. Discrimination parameters λ_i are typically assumed positive. When $\lambda_i = 1$ for $i = 1,\dots,I$ model in (1) reduces to the one-parameter logistic (1PL) model. Often, conditional log-odds in (1) are reparametrized as $\lambda_i\eta_j + \gamma_i$, with $\gamma_i = -\lambda_i \times \beta_i$. Sometimes this is reffered to as *slope-intercept* (SI)parameterization as opposed to the *IRT* parameterization in (1) traditional cosidered for interpretation.

Traditional literature assumes that $\eta_j \sim \mathcal{N}(0,1)$ for $j = 1,\dots,N$, but there are situations in which such assumption can be too restrictive. To add more flexibility, we can extend the model in (1) via a DP prior as

$$\eta_j|G \sim G \quad G \sim DP(\alpha, G_0) \qquad (2)$$

where α is the concentration parameter and G_0 the base measure. The DP process is often represented via the Chinese Restaurant Process representation, introducing a set of indicator variables z_j for $j = 1,\dots,N$ indicating the cluster assignment for the ability η_j. The prior in (2) becomes

$$(\eta_j|z_j = h) = \eta_h \quad \eta_h \sim \mathcal{N}(\mu_h, \sigma_h^2) \qquad (3)$$

with typically hyperpriors on μ_h and σ_h^2.

4 Model comparison

We compare estimation of the parametric and nonparametric estimation of the parametric 2PL models via simulation. Typically parameters of the 2PL model are not identifiable, so constraints are either included in the model or one can post-process posterior samples to meet the constraints. We consider this last option and use sum-to-zero constraints on the item parameters, i.e. $\sum_{i=i}^{I}\beta_i = 0, \sum_{i=i}^{I}\log(\lambda_i) = 0$ and estimate the 2PL under IRT and SI parameterizations.

We simulate data from two different scenarios changing the distribution generating the latent abilities. We simulate responses from $N = 2,000$ individuals to $I = 20$ binary items. Values for the discrimination parameters $\{\lambda_i^0\}_{i=1}^{20}$ are sampled from a $Unif(0.5, 1.5)$, while values for difficulty parameters $\{\beta_i^0\}_{i=1}^{20}$ are taken as equally spaced between $(-3,3)$. In particular we considered for the latent abilities η_j^0 for $j = 1,\ldots,2000$:

1. **Unimodal scenario**. Latent abilities comes from a normal distribution with mean 0 and standard deviation 1.25.
2. **Bimodal scenario**. Latent abilities comes from a mixture of two normal distribution with means $\{-2,2\}$ and common standard deviation 1.25.

We implement all the strategies in NIMBLE, choosing moderately vague priors for the item parameters, $\beta_i \sim \mathcal{N}(0,3), \gamma_i \sim \mathcal{N}(0,3), \log(\lambda_i) \sim \mathcal{N}(0.5, 0.5)$ for $i = 1,\ldots,I$. We assume normal latent abilities $\eta_j \sim \mathcal{N}(\mu_\eta, \sigma_\eta^2)$ for $j = 1,\ldots,N$, and placed a $\mathcal{N}(0,3)$ on μ_η and $Unif(0,10)$ on the standard deviation σ_η in the parametric case, while in the nonparametric setting we choose $G_0 \equiv \mathcal{N}(0,3) \times Inv-Gamma(1.01, 2.01)$. We run the MCMC for $50,000$ iterations using a 10% burn-in of $5,000$ iterations, and check traceplots for convergence.

Table 1 reports the minimum effective samples size (ESS) per second relative to the strategies, computed by dividing the ESS for the computation time. As expected there is a loss in efficiency when moving from the parametric to the semiparametric specification, given that sampling from the Dirichlet Process requires more computational effort. Flexibility comes with a price, but also with a benefit for inference when abilities are not normal. While in the unimodal scenario results match, in the bimodal there are substantial differences.

	unimodal	simulation	bimodal	simulation
Model	parametric	bnp	parametric	bnp
IRT unconstrained	0.78	0.33	2.43	1.27
SI unconstrained	1.06	0.57	0.23	0.06

Table 1 Minimum ESS/seconds for different estimations strategies of the 2PL model parameters under the two simulated scenarios.

For example, Figure 1 compares the density estimates of the posterior mean latent abilities from the parametric and semiparametric models, computed taking the posterior means of the $\{\eta_j\}_{j=1}^{N}$. It can be notice that the parametric model detect just

one mode because of the underlying normal assumption, while the semiparametric specification recover the true density structure. Better estimation of the latent abilities helps to avoid bias in inference, for example when estimating item parameters or item characteristics curves (ICC).

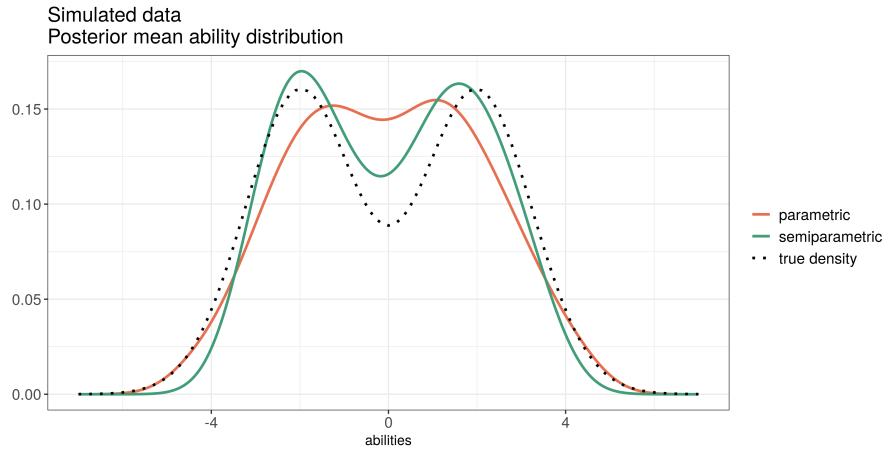

Fig. 1 Density estimates of the posterior mean latent abilities under the parametric and semiparametric 2PL models under the bimodal simulated scenario. Both models are estimated under the unconstrained IRT parameterizations, the most efficient from Table 1.

References

[1] D. Blackwell, J. B. MacQueen, et al. Ferguson distributions via pólya urn schemes. *The annals of statistics*, 1(2):353–355, 1973.

[2] P. de Valpine, D. Turek, C. J. Paciorek, C. Anderson-Bergman, D. T. Lang, and R. Bodik. Programming with models: Writing statistical algorithms for general model structures with nimble. *Journal of Computational and Graphical Statistics*, 26(2):403–413, 2017.

[3] K. A. Duncan and S. N. MacEachern. Nonparametric bayesian modelling for item response. *Statistical Modelling*, 8(1):41–66, 2008.

[4] T. S. Ferguson. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1(2):209–230, 03 1997.

[5] H. Finch and J. M. Edwards. Rasch model parameter estimation in the presence of a nonnormal latent trait using a nonparametric bayesian approach. *Educational and Psychological Measurement*, 76(4):662–684, 2016.