GPIRT: A Gaussian Process Model for Item Response Theory

JBrandon Duck-Mayr
Washington University in St. Louis
j.duck-mayr@wustl.edu

Roman Garnett
Washington University in St. Louis
garnett@wustl.edu

Jacob M. Montgomery
Washington University in St. Louis
jacob.montgomery@wustl.edu

Abstract

The goal of item response theoretic (IRT) models is to provide estimates of latent traits from binary observed indicators and at the same time to learn the item response functions (IRFs) that map from latent trait to observed response. However, in many cases observed behavior can deviate significantly from the parametric assumptions of traditional IRT models. Nonparametric IRT models overcome these challenges by relaxing assumptions about the form of the IRFs, but standard tools are unable to simultaneously estimate flexible IRFs and recover ability estimates for respondents. We propose a Bayesian nonparametric model that solves this problem by placing Gaussian process priors on the latent functions defining the IRFs. This allows us to simultaneously relax assumptions about the shape of the IRFs while preserving the ability to estimate latent traits. This in turn allows us to easily extend the model to further tasks such as active learning. GPIRT therefore provides a simple and intuitive solution to several longstanding problems in the IRT literature.

1 INTRODUCTION

Item response theory (IRT) (Rasch, 1960; Lord & Novick, 1968) is a widely used framework for estimating latent traits across many application domains, including educational testing (Lord, 1980), psychometrics (Embretson & Reise, 2000), political science (Martin & Quinn, 2002; Clinton et al., 2004), and more. Like other dimensionality reduction techniques, the goal is to take a large set of observed features and map them to a low-dimensional representation maintaining latent structure. A prototypical setting is assigning latent ability scores to students answering test questions. The observed responses are typically binary or categorical, making standard factor analysis inappropriate. Further, in addition to estimating latent scores, we also seek to simultaneously estimate the mapping between latent positions and observed outcomes. These mappings are themselves of interest to applied scholars and enable downstream tasks such as optimal test construction or active learning procedures like computerized adaptive testing (CAT) (Weiss, 1982; van der Linden & Pashley, 2010). We refer to these probabilistic mappings from latent ability to predicted outcome as *item response functions* (IRFs).

Common IRT models are parametric, typically estimating a sigmoidal mapping between the latent space and a binary outcome. Problems arise, however, when respondent behavior deviates from the assumed parametric representations. For instance, nearly all models assume a *monotonic* relationship between a respondent's position on the latent scale θ and each observed response y, which can be violated in settings such as personality measurement (e.g., Meijer & Baneke, 2004). A further problem is *non-saturation,* where the probability of observing either outcome never fully approaches zero or one (e.g., guessing on multiple-choice tests). A more subtle problem occurs when there is *asymmetry* in the IRFs. Tests for cognitive abilities require higher levels of complex thinking for success such that the shape of the IRF for respondents with low values of θ do not mirror the shape for individuals with high values (Lee & Bolt, 2018).

Numerous parametric models have been proposed to accommodate these and other irregularities. A complete inventory of IRT-like models would be prohibitively long, but notable examples include: (i) the generalized graded unfolding model (GGUM) (Roberts et al., 2000), which allows for (symmetric) non-monotonic IRFs; (ii) the three- (3PL) (Birnbaum, 1968) and four-parameter (4PL) (Barton & Lord, 1981) logistic models that relaxed saturation assumptions by including "guessing" and "carelessness" parameters; and (iii) Samejima's (2000) logis-

Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI), PMLR volume 124, 2020.

tic positive exponential family models (LPEF), which allow for asymmetric IRFs.

A related literature approaches these problems from a nonparametric framework (see Sijtsma, 1998). Mokken models (Mokken, 1971; Mokken & Lewis, 1982) relax functional form assumptions for the IRFs but require monotonicity (Sijtsma & Molenaar, 2002), which is also required for other nonparametric techniques (e.g., Poole, 2000). Ramsay (1991) proposes a model based on kernel smoothing. However, standard nonparametric methods are unsatisfactory since the IRFs and θ parameters cannot be estimated simultaneously. A standard approach for kernel-based IRT models, for instance, is to smooth over the rank ordering of respondents in terms of their raw scores on the test. Thus, while the IRFs are flexible, the smoothing occurs not over the latent parameters θ, but over "reasonable estimates" of θ constructed from y (Mazza et al., 2012, p. 4). Further, since inference is not based on the true likelihood, these models extend poorly to downstream tasks such as active learning.

In this article, we propose a Bayesian nonparametric IRT model based on Gaussian process (GP) regression (Rasmussen & Williams, 2006). As we demonstrate, our model has several clear advantages. First, similar to existing kernel smoothing methods, the Gaussian process IRT model (GPIRT) can in principal recover any smooth IRF with minimal assumptions. However, adopting the Bayesian framework facilitates building smoothed IRFs in a more coherent manner based on the actual latent parameters rather than proxies constructed from y. Second, GPIRT is a direct extension of Bayesian parametric IRT models (Albert, 1992; Albert & Chib, 1993). Indeed, many parametric Bayesian models in the literature can be viewed as a special case of our more general framework. Finally, it is simple to extend the model for downstream tasks such as item diagnostics, test construction, and computerized adaptive testing.

Below we use GPIRT to generate more accurate estimates of latent ability parameters while simultaneously capturing smooth IRFs that are non-monotonic, non-saturating, and asymmetric. We demonstrate its flexibility with real-world data, including voting records from the US Congress and responses to a personality inventory measuring narcissism, and contrast the behavior of GPIRT with parametric baselines. We also demonstrate how to use GPIRT to implement seamless active learning by maximizing mutual information to rapidly determine a new respondent's latent score from their responses to adaptively chosen items. This enables dynamic test construction in the tradition of CAT, and we will show it performs admirably on real-world data.

2 OVERVIEW OF IRT

We start by presenting the groundwork for IRT models. We then introduce the GPIRT, describe an MCMC sampling algorithm, extend the model for active learning, and contrast it with prominent nonparametric IRT approaches in the literature. We conclude with two applications. We restrict our discussion to the unidimensional case, where items are assumed to load on a single underlying latent dimension; however, our model would extend to the multidimensional setting. The multidimensional case has been explored previously in the nonparametric IRT setting (e.g., Bartolucci et al., 2017), but it is not as common in practice. We will also focus on the binary response case, which is the canonical IRT setting. However, categorical responses can be handled via standard extensions to the model below.

2.1 OBSERVATION MODEL

Assume we have n binary-response items and m respondents who have each answered some subset of items. We will encode the response of respondent j on item i as $y_{ij} \in \{-1, 1\}$, where 1 encodes a positive (or correct) response and -1 denotes an negative (or incorrect) response. This presentation is in keeping with GP classification models and deviates trivially from standard IRT presentations where we typically take $y \in \{0, 1\}$.

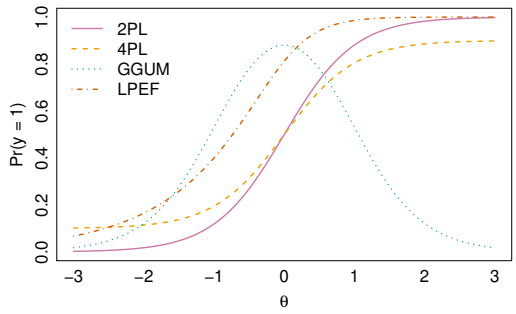
IRT assumes a simple probabilistic model for y according to (i) a latent score associated with each respondent and (ii) an item-specific mapping between latent score and the probability of observing $y_{ij} = 1$. Let Θ be some latent space (here we take $\Theta = \mathbb{R}$) in which we wish to embed respondents. We assume each respondent j has some unknown location in this space, θ_j and that that each item i has an associated latent function $f_i \colon \Theta \to \mathbb{R}$ that gives rise to responses via a sigmoidal link function. Namely, we assume the probability that respondent j answers item i positively is

$$\Pr(y_{ij} = 1 \mid \theta_j, f_i) = \sigma\big(f_i(\theta_j)\big), \tag{1}$$

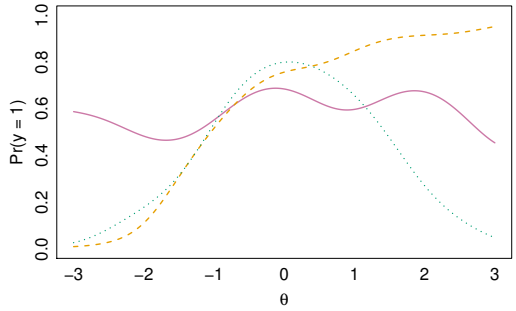
where $\sigma \colon \mathbb{R} \to (0, 1)$ is a monotonic sigmoidal "squashing" function such as the logistic or standard normal CDF (inverse probit). We will further adopt the standard assumption that multiple responses across a set of items/respondents are conditionally independent given the latent scores and latent functions. For a set of observed responses $\{y_{ij}\}$, this results in the likelihood:

$$\Pr\big(\{y_{ij}\} \mid \{f_i\}, \{\theta_j\}\big) = \prod_i \prod_j \Pr(y_{ij} \mid \theta_j, f_i), \tag{2}$$

where the product extends over the set of responses.



(a) IRFs for standard IRT models



(b) IRFs for GPIRT

Figure 1: Example IRFs for the 2PL, 4PL, GGUM, and LPEF models compared to IRFs for GP latent functions.

Interpreted as a function of θ_j, the marginal probability of a positive response to item i, Equation (1) is known as the ith *item response function* (IRF). The goal of IRT is to jointly estimate the respondents' latent locations $\{\theta_j\}$ and the IRFs from some set of training observations $\{y_{ij}\}$. This may seem daunting since we are in essence trying to simultaneously learn a latent embedding *and* the mappings $\{f_i\}$ using only observed responses. The problem is tractable if we assume that the latent functions $\{f_i\}$ are smooth over Θ. In that case, the fact that each respondent corresponds to a *single* latent location shared by all the IRFs gives us hope that we can find an embedding that places respondents with correlated responses in neighboring regions of the latent space.

2.2 PARAMETRIC IRT MODELS AND CAT

In the two-parameter logistic model (2PL) (Birnbaum, 1968), σ is taken to be the logistic function and we assume a linear parametric form for the latent functions:

$$f_i(\theta_j) = \beta_{i0} + \beta_{i1}\theta_j.$$

By convention, θ_j is referred to as the *ability* parameter for respondent j. These parametric assumptions are quite restrictive and fail in numerous settings. As a consequence, a wide array of alternative parametric forms for the $\{f_i\}$ have been proposed, including the GGUM, 3PL, 4PL, and LPEF models discussed above. Example IRFs for each are shown in Figure 1(a).

Critically, the IRFs are themselves often of interest to researchers. Sometimes, as in our analysis of the US Congress below, the shape of the IRF provides important insights about a specific item. In other cases, such as measuring personality traits, IRFs are used for test construction where the goal is to choose a set of items (i.e., a test in the sense of an examination) that will reveal as much as possible about the target population. This includes computerized adaptive testing, where items are chosen dynamically during test administration (Weiss, 1982). CAT is widely used in educational testing (van der Linden & Pashley, 2010), psychology (Waller & Reise, 1989), and survey research (Montgomery & Rossiter, Forthcoming).

A number of CAT algorithms have been proposed in the literature. Of particular relevance here are the maximum expected Kullback–Leibler (MEKL) criterion (Chang & Ying, 1996) and its Bayesian variant, the maximum posterior weighted Kullback-Leibler (MPWKL) criterion (van der Linden, 1998), which is equivalent to the mutual information between an unknown response and the latent ability score.[1] These methods seek to accelerate the convergence of our estimate of θ to the true unknown posterior by maximizing the expected information gain from each item selected from the larger inventory of potential items. MPWKL can be directly extended to our GPIRT model. It is worth noting that all of the above approaches for CAT assume a pre-calibrated parametric IRT model. To our knowledge Xu & Douglas (2006) present the only nonparametric IRT CAT algorithm.

3 GPIRT

Our proposed model, the Gaussian process IRT (GPIRT) model, extends standard models by placing Gaussian process priors on the latent functions $\{f\}$. We then perform joint inference over latent functions and ability scores. This procedure allows us to quantify and propagate our typical lack of certainty about the shape of the IRFs, and allows us to readily adopt the advances in Gaussian process modeling and inference seen over the past decade for estimation and active learning.

Although constructing and performing inference in this model will be straightforward with existing machinery, GPIRT provides a simple and intuitive solution to several longstanding problems in the IRT literature. The

[1] In the CAT literature the first appearance of MEKL, which might sound equivalent to mutual information, was defined in a somewhat idiosyncratic manner.

model makes no assumptions about the shape of IRFs beyond smoothness over the latent score θ. In principle, this allows the resulting estimated IRFs to take the shape of any smooth function (assuming suitable flexibility in the choice of covariance function) and may be non-monotonic, non-saturating, non-symmetric, or all of the above. IRFs derived from draws from a GP prior are shown in Figure 1(b). Further, adopting the Bayesian framework allows us to simultaneously estimate these flexible IRFs and the latent scores via a sampling approach we will describe below. This removes potential bias in IRF estimation from error in score estimation as well as relaxing the monotonicity assumption of Mokken-like models. Finally, the construction of the model facilitates item selection in either a manual or adaptive testing setting by adopting established ideas from Bayesian experimental design.

The most critical innovation in the GPIRT model is that rather than assuming that the functions $\{f_i\}$ belong to a specific parametric family, we place an independent Gaussian process prior distribution on each:

$$p(f_i) = GP(f_i; \mu, K), \tag{3}$$

where μ is a shared prior mean function that can be chosen according to prior beliefs, and K is a covariance function between respondents' latent traits. We could choose the covariance function as we see fit; however, for this investigation we took the pervasive squared exponential covariance function with unit length scale:

$$K(\theta, \theta') = \exp\big(-\tfrac{1}{2}(\theta - \theta')^2\big). \tag{4}$$

Note that we will also be taking a standard normal prior on the latent scores $\{\theta\}$ and thus the unit length scale is compatible with our expected range of latent scores.

For the mean function, $\mu \equiv 0$ is the most agnostic (corresponding to a prior IRF assigning 50% probability to a positive response regardless of location), but does not leverage prior knowledge about IRF profiles. Another reasonable choice is a linear prior mean $\mu(\theta) = \beta_{i0} + \beta_{i1}\theta$, which is most appropriate in a context where monotonic IRFs are expected, but where we prefer not to impose *strict* linearity. Note that with a linear mean we can recover standard two-parameter IRT models by selecting σ appropriately and taking K to be a linear covariance. GPIRT therefore provides a nonparametric generalization where each item is explained by a latent linear trend with smooth nonlinear deviations.

To complete the model, we place an independent normal prior over the respondents' latent scores $\{\theta\}$:

$$p\big(\{\theta_j\}\big) = \textstyle\prod_j \phi(\theta_j),$$

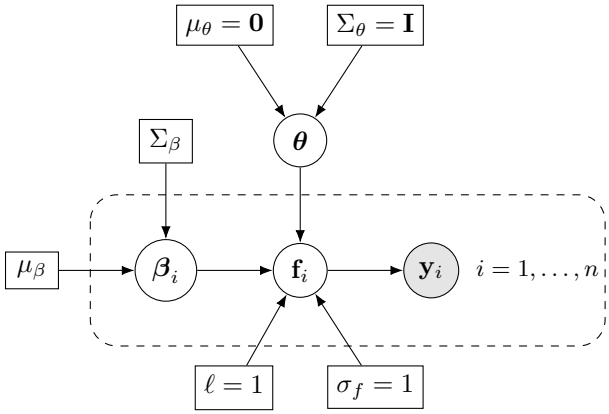


Figure 2: Graphical representation of GPIRT.

where ϕ is the standard normal PDF. We also place independent normal priors over the coefficients $\{\beta\}$ for linear or higher-order polynomial mean functions as necessary.

A graphical representation of the model is shown in Figure 2. Here, μ_θ and Σ_θ are the mean and covariance of the prior on $\boldsymbol{\theta}$. Similarly, μ_β and Σ_β are the mean and covariance of the prior on $\boldsymbol{\beta}$. Finally, σ_f and ℓ are the scale factor and length scale of K respectively.

3.1 ESTIMATION

Assume we have a set of responses $\mathcal{D} = \{y_{ij}\}$. We wish to estimate the posterior over the latent functions $\{f\}$ and latent scores $\{\theta\}$, $p\big(\{f\}, \{\theta\} \mid \mathcal{D}\big)$. Since the posterior is analytically intractable, we perform inference via Gibbs sampling.[2] We begin by initializing a Markov chain by sampling initial latent scores $\{\theta\}$ and mean function coefficients $\{\beta\}$ from their respective priors. Then, given $\{\theta\}$ and $\{\beta\}$, we sample the latent function values corresponding to the observations $\{y_{ij}\}$. Let $\mathcal{D}_j$ represent the responses from respondent j only, $\mathbf{f}_i = \{f_i(\theta_j)\}$ represent the latent function values associated with all responses to item i, and $\boldsymbol{\theta}_i$ be the latent scores of all respondents who answered item i. We sample:

$$\mathbf{f}_i \sim \mathcal{N}\big(\mu(\boldsymbol{\theta}_i; \{\beta\}), K(\boldsymbol{\theta}_i, \boldsymbol{\theta}_i)\big).$$

Finally, we extend the sampled vectors $\mathbf{f}_i$ to dense samples of each latent function. This is feasible since the latent space Θ used in IRT is universally small; in the vast majority of cases, this dimension is 1 or 2. In our implementation, we took a dense grid $\boldsymbol{\theta}^*$ spanning from -5 to 5 in increments of 0.01, which is sufficient to densely

[2] Variational inference with pseudopoints in θ-space for the items would be possible for large-scale data. This would be a relatively simple extension using existing approaches such as Hensman et al. (2015). However, the exact sampling scheme we outline is sufficient for the applications we investigate.

cover the support of the latent score prior. Now for each item i, we can sample dense vectors $\mathbf{f}_i^*$ on this set from the posteriors induced by $\boldsymbol{\theta}_i$ and $\mathbf{f}_i$, which is normal as the posterior on f_i is a Gaussian process:

$$p\big(\mathbf{f}_i^* \mid \boldsymbol{\theta}^*, \boldsymbol{\theta}, \mathbf{f}_i, \{\beta\}\big) = \mathcal{N}(\mathbf{f}_i^*; \mathbf{m}^*, \mathbf{C}^*),$$

where

$$\mathbf{m}^* = \mu(\boldsymbol{\theta}^*) + K(\boldsymbol{\theta}^*, \boldsymbol{\theta})\, K(\boldsymbol{\theta}, \boldsymbol{\theta})^{-1}\big(\mathbf{f}_i - \mu(\boldsymbol{\theta})\big);$$
$$\mathbf{C}^* = K(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*) - K(\boldsymbol{\theta}^*, \boldsymbol{\theta}) K(\boldsymbol{\theta}, \boldsymbol{\theta})^{-1} K(\boldsymbol{\theta}, \boldsymbol{\theta}^*),$$

and the dependence of μ on $\{\beta\}$ has been omitted.

With the Markov chain initialized, we proceed as follows. First we sample new latent function values at the observations for each item $\{\mathbf{f}_i\}$, using elliptical slice sampling (Murray et al., 2010). We then extend the $\{\mathbf{f}\}$ to dense samples $\{\mathbf{f}^*\}$ as described above. Next we sample each of the latent scores θ_j from the posterior induced given the latent function samples $\{\mathbf{f}^*\}$. Extending these samples onto a dense grid allows us to compute a dense approximation of the *exact* posterior of θ_j on the grid. The unnormalized posterior is:

$$p\big(\theta_j \mid \{\mathbf{f}^*\}, \mathcal{D}_j\big) \propto p(\theta_j) \textstyle\prod_i \Pr\big(y_{ij} \mid \mathbf{f}_i^*(\theta_j)\big).$$

We then use inverse transform sampling to sample a new latent score for each respondent from their posterior. Finally, we sample new values for the mean function hyperparameters $\{\beta\}$ using a Metropolis–Hastings step with a Gaussian proposal distribution. We have implemented this inference method in the R package `gpirt`. Once this chain has mixed we can then estimate IRFs if desired by pushing a chain of samples of the dense latent functions through the chosen sigmoid and averaging.

The posterior distributions in all IRT models exhibit rotational invariance where the sign on the latent parameters $\{\theta_j\}$ and the shape of the functions $\{f_i\}$ can be altered to produce an identical likelihood. In applications, generally either the directionality of the latent space is not interesting, in which case samples from either reflective mode are useful, or the directionality is known, and posterior samples stuck in a substantively inappropriate mode can be replaced or post-processed by imposing the desired orientation (Stephens, 1997).

3.2 ACTIVE LEARNING

A major benefit of our fully Bayesian model is enabling a direct scheme for adaptive testing via sequential Bayesian experimental design. We consider the following task. Suppose we have estimated IRFs from data $\mathcal{D}$, for the ith item estimating:

$$\pi_i(\theta^*) = \Pr(y_i^* = 1 \mid \theta^*, \mathcal{D})$$

by marginalizing the latent function f_i. We then seek to adaptively present a series of items to some new respondent so as to learn their latent score as quickly as possible. Here we propose a natural scheme based on maximizing the mutual information between the unknown response to each item and the unknown latent score $\theta^\star$.

Our proposed algorithm works as follows. Let $\mathcal{D}^\star$ represent a dataset augmenting our training data $\mathcal{D}$ by the (initially empty) available responses from the new respondent. We initialize our belief about $\theta^\star$ to the chosen prior used during inference: $p(\theta^\star \mid \mathcal{D}^\star) = p(\theta)$. Now we compute the mutual information between the response $y_i^\star$ to item i and $\theta^\star$ given the available data:

$$I(y_i^\star; \theta^\star \mid \mathcal{D}^\star) = h(p_i^\star) - \mathbb{E}_{\theta^\star}\big[h(\pi_i(\theta^\star)) \mid \mathcal{D}^\star\big]. \quad (5)$$

Here $h(p)$ is the binary entropy function and $p_i^\star$ is the *marginal* probability of a positive response to item i:

$$p_i^\star = \int \pi_i(\theta^*)\, p(\theta^\star \mid \mathcal{D}^\star)\, \mathrm{d}\theta^\star.$$

We present the respondent with the item maximizing the mutual information and augment $\mathcal{D}^\star$ with the response. We then approximate the updated posterior distribution on $\theta^\star$ by multiplying the current belief by the IRF for the chosen item (for a positive response) or its complement (for a negative response), e.g. if $y_i^\star = 1$ we take:

$$p(\theta^\star \mid \mathcal{D}^\star, y_i^\star) \propto p(\theta^\star \mid \mathcal{D}^\star)\, \pi_i(\theta^\star).$$

We repeat this process until a stopping condition is met. This may be a pre-chosen number of items, but could also be sufficient certainty in $\theta^\star$. This procedure is both computationally efficient and effective in practice. Although we could refit the entire GPIRT model each time we get a new response, the extra cost is not likely to be worth it if the training dataset is reasonably sized.

4 RELATED WORK

Inference in parametric IRT models is often achieved via some variant of maximum likelihood estimation (MLE), maximizing (2) as a function of $\{\beta_i\}$ and $\{\theta_j\}$. Unfortunately, maximizing the full joint likelihood has proven to be difficult and most parametric models therefore are estimated using the marginal maximum likelihood (MML) framework (Bock & Aitkin, 1981). Here, we assume that the marginal *posterior* distribution of the latent scores $\{\theta_j\}$ is known a priori. We will denote this assumed marginal distribution $q(\theta)$. In MML we estimate each respondent's contribution to the likelihood (2) by marginalizing her latent score under the assumed θ distribution:

$$\mathcal{L}_j \approx \int \prod_i \Pr(y_{ij} \mid \theta_j, f_i)\, q(\theta_j)\, \mathrm{d}\theta_j, \quad (6)$$

where $\mathcal{L}_j$ represents the component of the likelihood associated with respondent j in (2). In this procedure, the item-level parameters are estimated first, marginalizing the latent scores as in (6). $\{\theta_j\}$ is then estimated afterwards via some procedure such as calculating the expected a posteriori (EAP) estimate (Rizopoulos, 2006).

Most nonparametric methods build from Equation 6. The most relevant approaches here focus on relaxing assumptions about the link function σ (but see, e.g., Woods & Thissen, 2006). However, problems arise since it is difficult to simultaneously estimate $\{f_i\}$ and $\{\theta_j\}$. Indeed, the entire idea of the MML approach is to marginalize out $\{\theta_j\}$. Therefore, standard nonparametric IRT models estimate the IRF not based on $\{\theta_j\}$ but instead based on the observed data as a proxy.

For instance, in kernel-based IRT (Ramsay, 1991) we first transform respondents' scores (number correct) into quantiles of a specified latent trait distribution $q(\theta)$. That is, if respondent j is in the empirical s_jth percentile in raw average score across the items, we first estimate their latent score with $\hat{\theta}_j = Q^{-1}(s_j)$, where Q^{-1} is the inverse CDF for $q(\theta)$. Fixing these proxies for the latent scores, we can then use Nadaraya–Watson (Nadaraya, 1964; Watson, 1964) regression to estimate the IRFs by kernel smoothing over the training data.

Due to the inherent difficulties associated with simultaneously estimating IRFs and $\{\theta_j\}$, adopting a Bayesian framework is an attractive option (Albert, 1992). Given the likelihood in Equation (2), all that we need to complete the model is a prior on θ and (optionally) on the parameters defining $\{f_i\}$. Albert & Chib (1993) used normal priors in the context of the normal ogive model and developed a complete Gibbs sampler for the resulting model. Imai et al. (2016) estimates this same model using the expectation-maximization (EM) approach, which is the method we use in the Congress application below. Subsequent work has developed Bayesian versions of nearly all of the common parametric models.

Previous research has also been done in the area of Bayesian nonparametric IRT. Karabatsos & Sheu (2004) considered Bayesian inference under the monotone nonparametric framework of Mokken while Arenson & Karabatsos (2018) approximated monotone IRFs using a finite mixture of beta distributions. Duncan & MacEachern (2008) places a Dirichlet process (DP) prior on the $q(\theta)$ distribution while retaining the 2PL form for the latent functions and further reported a variant that instead models the IRFs as a DP as well (see also San Martín et al., 2011). However, we are aware of no existing model that adopts the GP approach we outline above.

The most closely related GP approach to our own might be Gaussian process latent variable models, or GPLVM (Lawrence, 2004). GPLVM is not particularly well-suited for IRT as there is a fundamental mismatch in the choice of likelihood, which is naturally binomial but normal in the GPLVM. A further issue is that GPLVM typically assumes that all "dimensions" (items) have a common/shared error term, which fits poorly in the measurement model domain most closely related to IRT. Moreover, the mismatch in likelihood makes response prediction less principled than the IRFs provided by GPIRT. Nonetheless, we include GPLVM in our benchmarks.

5 APPLICATIONS

We illustrate the benefits of GPIRT with two applications to real-world observational data. First we embed Members of the US House of Representatives from the 116th Congress (elected in 2018) from their roll-call records. Here we focus on finding interesting attributes of the IRFs that standard scaling cannot uncover due to their restrictive functional form assumptions. We then apply the model to a survey dataset, where respondents were given a narcissism battery. With this data we focus on improving predictive performance, and also illustrate that the model performs strongly in an active learning task.

5.1 ROLL CALL VOTING IN THE HOUSE OF REPRESENTATIVES

Members of the U.S. House of Representatives give recorded "yea" and "nay" votes on the various proposals the House considers. There is a long history in political science of using these votes to embed the legislators in a latent space, with a one-dimensional left–right ideological continuum being of interest in recent sessions. The gold standard ideological scores for members of Congress are the DW–NOMINATE scores (Poole & Rosenthal, 1997) and Bayesian IRT models (Clinton et al., 2004). Similarly to the 2PL model described above, the NOMINATE procedure assumes a specific (monotonic) functional form for IRFs.

The monotonicity assumption in 2PL and NOMINATE often holds and these models are highly predictive both in-sample and out-of-sample. In the Supplementary Material, we show that GPIRT performs equally well and sometimes better than the 2PL, although prediction is rarely a task in this setting. However, the strong parametric assumptions can obscure important dynamics in key roll-call votes, which can in turn result in embeddings that correspond poorly with other evidence about members' ideology.

Our focus in this application, therefore, is on recovering interesting aspects of IRFs and the qualitative value

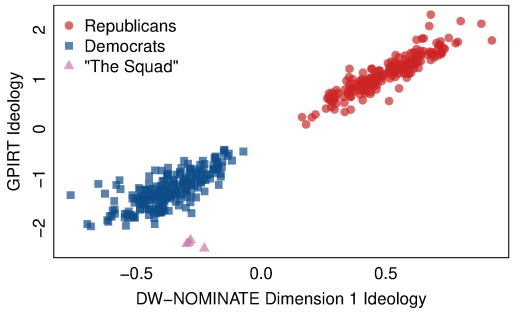


Figure 3: Ideology estimates for members of the 116th House of Representatives for the GPIRT model and DW–NOMINATE. Republicans' estimates are plotted in red circles and Democrats' estimates are plotted in blue squares, except Reps. Ocasio–Cortez, Omar, Pressley, and Tlaib (popularly known as "The Squad"), who are noted with purple triangles.

of the resulting embeddings. For example, in the 2019 session several proposals drew "nay" votes from both ends of the ideological spectrum, as Republicans voted against bills for being too liberal, and liberal Democrats voted against bills for not being liberal enough. In such a case, standard parametric assumptions result in extreme members to instead be scaled as moderates. Indeed, DW–NOMINATE rates Rep. Alexandria Ocasio–Cortez, a member of the Democratic Socialists of America, as more conservative than 86% of Democrats in the House of Representatives (Lewis et al., 2019). The procedure similarly scales the other members of "The Squad," Reps. Ilhan Omar, Ayanna Pressley, and Rashida Tlaib, as among the most conservative Democrats in the chamber. The more flexible approach of the GPIRT model does not force extreme members who vote against their party to be scaled as moderates.

We estimate GPIRT latent traits and IRFs using all roll call votes in the first session of the U.S. House of Representatives for the 116th Congress. We use only votes where the minority vote is at least 1% of the total with a quadratic mean function as we anticipate some non-monotonic IRFs. We include all members of the House who voted on at least one bill, with the exception of Rep. Justin Amash who left the Republican party.

Figure 3 compares GPIRT ideology estimates with their DW–NOMINATE scores. We can see that GPIRT and DW–NOMINATE largely agree on the relative ideological placement of members. However, GPIRT finds The Squad to be the most liberal members of the House as we would substantively expect.

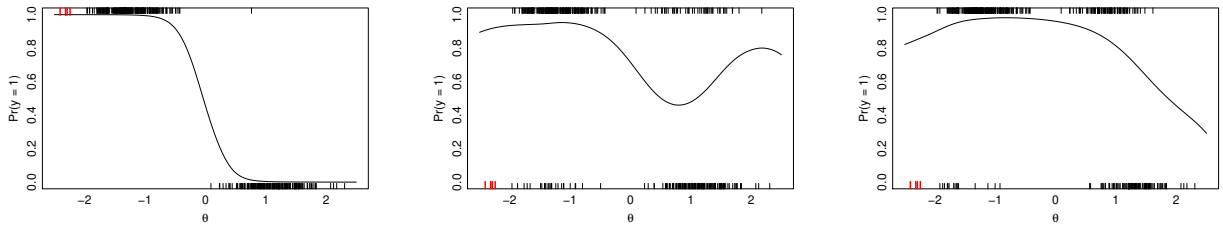
For most roll calls we observe the linear, monotonic patterns that parametric IRT-like models require (Figure 3a). However, we also observe other patterns, such as the non-monotonic IRFs depicted in Figures 4b and 4c. Here, the traditional IRT model treats this as uninformative vote with a relatively flat IRF (Figure 4e), while the GPIRT detects the more nuanced ideological structure of the vote including non-saturation, non-monotonicity, and asymmetry in θ (Figure 4b).

Sometimes the ability to account for non-monotonic voting patterns is necessary to catch an important legislative dynamic. For an example, consider the item response functions for H.J. Res. 31, a funding bill to end a partial government shutdown, depicted in Figures 4c and 4f. Here, the vast majority of Democrats supported the bill along with many moderate Republicans. Conservative Republicans opposed the bill on the grounds that it did not include funding for the border wall, while liberal Democrats such as Ocasio–Cortez, Omar, Pressley, and Tlaib argued it did not sufficiently reduce funding for border detention facilities (McPherson, 2019). The GPIRT is able to recover this non-monoticity as shown in Figure 4c, while the 2PL IRF shown in Figure 4f treats the item as largely not informative.

5.2 NARCISSISTIC PERSONALITY INVENTORY

We also apply the GPIRT model to responses to a 40 item narcissistic personality inventory (NPI) (Raskin & Terry, 1988), which measures one's "grandiose yet fragile sense of self and entitlement as well as a preoccupation with success and demands for admiration" (Ames et al., 2006, pp. 440–441). For each item on the NPI, respondents are presented with two statements and asked to select which one fits them best; an example item is "Modesty doesn't become me" (positive) vs. "I am essentially a modest person" (negative). Responses to the complete inventory were collected from the Open Source Psychometrics Project (Open-Source Psychometrics Project, 2012) ($n = 10,440$) and a convenience sample from the Qualtrics panel ($n = 2,945$). We estimate the 2PL IRT, kernel-smoothed IRT, GPLVM, and GPIRT models on a randomly selected sub-sample of 2,000 respondents. The IRFs in the GPIRT showed deviations from the standard 2PL IRT model for several items.

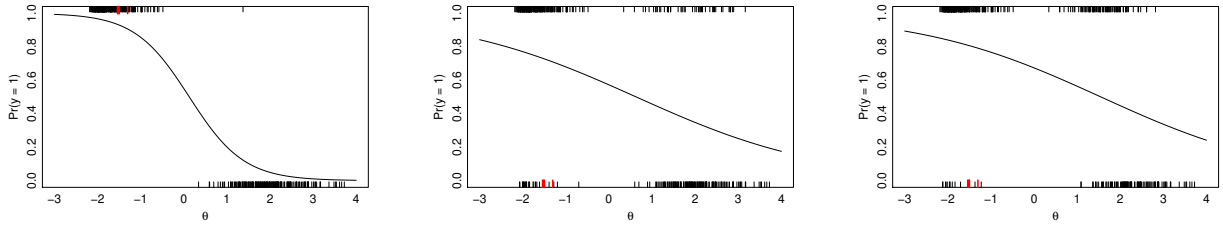
Permitting flexibility in the IRFs allows the GPIRT to perform better than the comparison models as measured by predictive performance on held-out responses. To show this, we performed the following experiment. We randomly selected 20% of the available observations and treated them as missing, then estimated the IRFs and the respondents' latent traits from the remaining data. We then used the trained model to predict the missing responses and calculated the log likelihood, predictive accuracy, and area under the receiver operating character-



(a) IRF for HR2722, the "Securing America's Federal Elections Act"

(b) IRF for an amendment to HR2500 ending the Cyprus arms embargo

(c) IRF for HJRES31, a budgetary resolution to avoid a government shutdown



(d) 2PL IRF for HR2722

(e) 2PL IRF for the Cyprus amendment

(f) 2PL IRF for HJRES31

Figure 4: Example IRFs for three roll call votes in the U.S. House of Representatives' first session of the 116th Congress. θ estimates for members who voted "Yea" are displayed in a rug at the top of the plots, while θ estimates for members who voted "Nay" are displayed in a rug at the bottom. Rug lines for members of the Squad are drawn in red.

Table 1: Fit for GPIRT, 2PL, GPLVM, and kernel-smoothed IRT models on NPI responses with 20% of the data held out. $\varepsilon \approx 2.2 \times 10^{-16}$ is machine epsilon.

Model	$\mathcal{L}/N$	t-test vs. GPIRT	AUC	Accuracy
GPIRT	-0.52		0.82	0.74
2PL	-0.73	0.20 ($p < \varepsilon$)	0.66	0.63
GPVLM	-1.39	0.87 ($p < \varepsilon$)	0.71	0.66
KS IRT	-0.62	0.10 ($p \approx 10\varepsilon$)	0.68	0.68

istic curve (AUC) for the the held out observations using the trained model; we repeated this procedure 20 times. The mean log likelihood for the held out observations is reported in Table 1.

For the NPI dataset, the GPIRT outperformed the comparison models, with a higher mean log likelihood than any model in the comparison set. A paired t-test confirms these are significant improvements in model fit, and is also confirmed by comparing the mean accuracy and AUC across the 20 simulations.

5.3 ACTIVE LEARNING

We also used these datasets to evaluate our adaptive testing procedure. For the NPI dataset, we compare our procedure to a published reduced-form NPI battery (Ames et al., 2006). The reduced-form battery contains 16 questions deemed by experts to be a suitable subset.

Table 2: RMSE for 1000 responses to 16 items

	CAT	Fixed	Random
RMSE	0.257	0.338	0.321
Improvement vs. random	20%	-5.5%	—

For 1,000 randomly selected respondents that were not included in our initial training sample, we estimated their latent traits using their actual responses to the full 40-item battery and our estimated GPIRT IRFs. For each respondent, we then estimated their latent trait using the 16-item reduced-form battery from Ames et al. (2006), 16 items chosen using our adaptive testing scheme, and 16 randomly selected items. We calculated the root mean squared error (RMSE) of these batteries' θ estimates with the full-battery estimates; the results are presented in Table 2. Notably the RMSE for the adaptive battery gives a 20% improvement over randomly selected items, whereas the reduced-form battery actually performs *worse* than random selection.

6 CONCLUSION

In this article we provided a fully Bayesian nonparametric IRT model that allows for the simultaneous estimation of ability parameters and IRFs while allowing for high levels of flexibility in the IRF shapes. We showed that this model performs better than standard parametric models in terms of estimating unusual IRFs, predictive

accuracy, and in an active learning setting.

Trivial extensions include allowing for multiple dimensions and categorical response functions. Additionally, we could avoid the independence assumption on the items and couple them together via a multi-task kernel. For example if the items were parameterized by some vector x we could take a product kernel such as $K(x, x')K(\theta, \theta')$. While a strength of the method is avoiding potentially unfounded assumptions about monotonicity, saturation, and symmetry, we could impose monotonicity via EP to impose derivative constraints (e.g. Riihimäki & Vehtari, 2010), saturation via a slightly modified likelihood, and symmetry via an appropriately modified kernel (e.g. it would suffice to take $K(|x - c|, |x' - c|)$, where K is a desired base kernel, to impose symmetry around c). Other potential extensions are to include feature vectors in the learning kernel to allow, for instance, smooth changes in ability over time or differential item functioning for subgroups.

Acknowledgements

RG was supported by the National Science Foundation (NSF) under award numbers IIS–1939677, OAC–1940224, and IIS–1845434. JM was supported by the NSF under award number SES–1558907.

References

Albert, J. H. Bayesian estimation of normal ogive item response curves using Gibbs sampling. *Journal of Educational Statistics*, 17(3):251–269, 1992.

Albert, J. H. and Chib, S. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669–679, 1993.

Ames, D. R., Rose, P., and Anderson, C. P. The NPI-16 as a short measure of narcissism. *Journal of Research in Personality*, 40(4):440–450, 2006.

Arenson, E. and Karabatsos, G. A Bayesian beta-mixture model for nonparametric IRT (BBM-IRT). *Journal of Modern Applied Statistical Methods*, 17(1):1–17, 2018.

Bartolucci, F., Farcomeni, A., and Scaccia, L. A non-parametric multidimensional latent class IRT model in a Bayesian framework. *Psychometrika*, 82(4):952–978, 2017.

Barton, M. A. and Lord, F. M. An upper asymptote for the three-parameter logistic item-response model. *ETS Research Report Series*, 1981(1):i–8, 1981.

Birnbaum, A. L. Some latent trait models and their use in inferring an examinee's ability. In Lord, F. and Novick, M. (eds.), *Statistical Theories of Mental Test Scores*. Addison-Wesley, Reading, MA, 1968.

Bock, R. D. and Aitkin, M. Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, 46(4):443–459, 1981.

Chang, H.-H. and Ying, Z. A global information approach to computerized adaptive testing. *Applied Psychological Measurement*, 20(3):213–229, 1996.

Clinton, J., Jackman, S., and Rivers, D. The statistical analysis of roll call voting: A unified approach. *American Political Science Review*, 98(2):355–370, 2004.

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

Embretson, S. E. and Reise, S. P. *Item Response Theory for Psychologists*. Lawrence Erlbaum, Mahwah, New Jersey, 2000.

Hensman, J., Matthews, A., and Ghahramani, Z. Scalable variational Gaussian process classification. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, pp. 351–360, 2015.

Imai, K., Lo, J., and Olmsted, J. Fast estimation of ideal points with massive data. *American Political Science Review*, 110(4):631–656, 2016.

Karabatsos, G. and Sheu, C.-F. Order-constrained Bayes inference for dichotomous models of unidimensional nonparametric IRT. *Applied Psychological Measurement*, 28(2):110–125, 2004.

Lawrence, N. D. Gaussian process models for visualisation of high dimensional data. In *Advances in Neural Information Processing Systems*, volume 16, pp. 239–336, 2004.

Lee, S. and Bolt, D. M. Asymmetric item characteristic curves and item complexity: Insights from simulation and real data analyses. *Psychometrika*, 83(2): 453–475, 2018.

Lewis, J. B., Poole, K., Rosenthal, H., Boche, A., Rudkin, A., and Sonnet, L. Voteview: Congressional roll-call votes database, 2019. URL `https://voteview.com/`.

Lord, F. and Novick, M. R. *Statistical Theories of Mental Test Scores*. Addison-Wesley, Reading, MA, 1968.

Lord, F. M. *Applications of Item Response Theory to Practical Testing Problems*. L. Erlbaum Associates, Hillsdale, NJ, 1980.

Martin, A. D. and Quinn, K. M. Dynamic ideal point estimation via Markov chain Monte Carlo for the U.S. Supreme Court, 1953–1999. *Political Analysis*, 10(2): 134–153, 2002.

Mazza, A., Punzo, A., and McGuire, B. KernSmoothIRT: An R package for kernel smoothing in item response theory. *arXiv preprint arXiv:1211.1183*, 2012.

McPherson, L. House passes appropriations package to avert shutdown, sends to Trump, 2019. URL `https://www.rollcall.com/news/congress/house-passes-appropriations-package-avert-shutdown-sends-trump/`.

Meijer, R. R. and Baneke, J. J. Analyzing psychopathology items: A case for nonparametric item response theory modeling. *Psychological Methods*, 9(3):354–368, 2004.

Mokken, R. J. *A Theory and Procedure of Scale Analysis*. D Gruyter, Berlin, 1971.

Mokken, R. J. and Lewis, C. A nonparametric approach to the analysis of dichotomous item responses. *Applied Psychological Measurement*, 6(4):417–430, 1982.

Montgomery, J. M. and Rossiter, E. L. So many questions, so little time: Adaptive personality inventories for survey research. *Journal of Survey Statistics and Methodology*, Forthcoming.

Murray, I., Adams, R. P., and MacKay, D. J. C. Elliptical slice sampling. *The Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*, pp. 541–548, 2010.

Nadaraya, E. A. On estimating regression. *Theory of Probability & Its Applications*, 9(1):141–142, 1964.

Open-Source Psychometrics Project. Open psychology data: Raw data from online personality tests, 2012. URL `https://openpsychometrics.org/_rawdata/`.

Poole, K. T. Nonparametric unfolding of binary choice data. *Political Analysis*, 8(3):211–237, 2000.

Poole, K. T. and Rosenthal, H. *Congress: A Political-Economic History of Roll Call Voting*. Oxford University Press, New York, 1997.

Ramsay, J. O. Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika*, 56(4):611–630, 1991.

Rasch, G. *Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests.* Nielsen & Lydiche, 1960.

Raskin, R. and Terry, H. A principal-components analysis of the narcissistic personality inventory and further evidence of its construct validity. *Journal of Personality and Social Psychology*, 54(5):890–902, 1988.

Rasmussen, C. E. and Williams, C. K. I. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

Riihimäki, J. and Vehtari, A. Gaussian processes with monotonicity information. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pp. 645–652, 2010.

Rizopoulos, D. ltm: An R package for latent variable modeling and item response theory analyses. *Journal of Statistical Software*, 17(5):1–25, 2006.

Roberts, J. S., Donoghue, J. R., and Laughlin, J. E. A general item response theory model for unfolding unidimensional polytomous responses. *Applied Psychological Measurement*, 24(1):3–32, 2000.

Samejima, F. Logistic positive exponent family of models: Virtue of asymmetric item characteristic curves. *Psychometrika*, 65(3):319–335, 2000.

San Martín, E., Jara, A., Rolin, J.-M., and Mouchart, M. On the Bayesian nonparametric generalization of IRT-type models. *Psychometrika*, 76(3):385–409, 2011.

Sijtsma, K. Methodology review: Nonparametric IRT approaches to the analysis of dichotomous item scores. *Applied Psychological Measurement*, 22(1): 3–31, 1998.

Sijtsma, K. and Molenaar, I. W. *Introduction to Nonparametric Item Response Theory*, volume 5. Sage, 2002.

Stephens, M. *Bayesian Methods for Mixtures of Normal Distributions*. PhD thesis, University of Oxford, 1997.

van der Linden, W. J. Bayesian item selection criteria for adaptive testing. *Psychometrika*, 63(2):201–216, 1998.

van der Linden, W. J. and Pashley, P. J. *Elements of Adaptive Testing*. Springer, New York, 2010.

Waller, N. G. and Reise, S. P. Computerized adaptive personality assessment: An illustration with the absorption scale. *Journal of Personality and Social Psychology*, 57(6):1051–1058, 1989.

Watson, G. S. Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, pp. 359–372, 1964.

Weiss, D. J. Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement*, 6(4):473–492, 1982.

Woods, C. M. and Thissen, D. Item response theory with estimation of the latent population distribution using spline-based densities. *Psychometrika*, 71(2): 281–301, 2006.

Xu, X. and Douglas, J. Computerized adaptive testing under nonparamteric IRT models. *Psychometrika*, 71 (1):121–137, 2006.