GD-GPIRT: A Generalized Dynamic Gaussian Process Model for Item Response Theory with Longitudinal Observations

Anonymous Author
Anonymous Institution

Abstract

Social scientists are often interested in using ordinal indicators to estimate latent traits that change over time. Frequently, this is done with item response theoretic (IRT) models that describe the relationship between those latent traits and observed indicators. We combine recent advances in Bayesian nonparametric IRT, which makes minimal assumptions on shapes of item response functions, and Gaussian process time series methods to capture dynamic structures in latent traits from longitudinal observations. We propose a generalized dynamic Gaussian process item response theory model (GD-GPIRT) as well as a Markov chain Monte Carlo sampling algorithm for estimation of both latent traits and response functions. We evaluate GD-GPIRT in simulation studies against baselines in dynamic IRT, and apply it to various substantive studies, including assessing public opinions on economy environment and congressional ideology leaning on abortion.

1 INTRODUCTION

How do the issue positions of the Congress evolve over time? Is there growing dissatisfaction with the economy after recessions? Are patients emotionally stable after psychological therapies? Answering these questions requires dynamic measures of traits or attitudes. Since self-reported ratings are known to be sensitive to individual variances and inconsistency (Wilcox et al., 1989), social scientists rely on latent trait models, where the latent variable of interest is inferred from a collection of noisy categorical indicators such as survey responses, voting outcomes or event counts.

However, analyzing how these traits change over time in practice introduces two problems. First, researchers must ensure that the inferred latent traits are comparable over time. It is widely understood that failure to do so can result in misleading or even nonsensical inferences (Bollen, 1980), and this problem is particularly difficult when we do not have repeated observations of the same items. Second, scholars must make model assumptions about the functional relationship between the latent traits and observed indicators. In practice, the assumed function form is typically fairly restrictive (e.g., generalized linear models), which can lead to biased or inefficient estimates when the real data fail to match the models' assumptions.

In this work, we propose a novel dynamic Gaussian process latent variable model based on item response theory (GD-GPIRT) for longitudinal and ordinal observations. While item response theory has seen applications in machine learning area such as predicting user preference in recommendation systems (Chen et al., 2005; Baylari and Montazer, 2009) students' answer in educational testing (Bergner et al., 2012; Cheng et al., 2019; Park et al., 2023) and evaluating different machine learning methods (Lalor et al., 2016; Martínez-Plumed et al., 2019), we focus on estimation of latent traits. We exploit recent advances in Bayesian nonparametric IRT with Gaussian process priors for flexibly modeling the response functions, and Gaussian process time series methods to capture dynamic structures in latent traits while maintaining measurement comparability. We also propose an efficient Markov chain Monte Carlo sampling algorithm, whose effectiveness is demonstrated in simulation studies. Finally, we apply GD-GPIRT to substantive studies: assessing public opinions on economy environment and congressional ideology leaning on abortion.

Our model makes two contributions to the field. First, it extends recent advances in Bayesian nonparametric models of latent traits to ordered categorical indicators. The existing models in this family are limited to continuous (Lawrence, 2003) or binary (Duck-Mayr

Preliminary work. Under review by AISTATS 2024. Do not distribute.

et al., 2020) indicators. However, in fields such as psychology or survey research, ordered-categorical responses are much more common. Other standard tools for ordered categorical responses assume strict parametric functional forms for the item response functions (IRFs) (Roberts et al., 2000; Duck-Mayr and Montgomery, 2022), strict monotonicity (Molenaar, 1997; Van der Ark, 2007), or both (Mokken, 1971). In contrast, GD-GPIRT offers a compromise, allowing flexibility in specification of prior structures to control the shapes (e.g., non-monotonic, asymmetric) of IRFs .

Second, GD-GPIRT provides a natural way to encode dynamics in the latent traits. Instead of estimating latent variables at different time periods independently, GD-GPIRT models the trajectory of each unit jointly through time using latent space models with dynamic structures. Some existing IRT-like models also allow latent traits to move over time, by assuming these trends are low-order polynomials (Poole and Rosenthal, 2001; Bailey, 2007; Proust-Lima et al., 2022) or realizations of a Wiener process (Martin and Quinn, 2002; Wang et al., 2013; Schnakenberg and Fariss, 2014; Chung et al., 2015). While the former can be far too restrictive, the latter may face a variance explosion issue in prediction due to their non-stationary nature. For example, the well-known Martin–Quinn scores for Supreme Court ideology can lead to extreme scores for justices at the ending of their careers due to this unbalanced model structure (Martin and Quinn, 2002). Alternatively, GD-GPIRT encodes the reasonable expectation that the latent trends are centered and stationary *a priori*, with hyperparameters controlling the bandwidth of variation in time and latent space.

In summary, to the best of our knowledge, GD-GPIRT is the first dynamic Bayesian non-parametric measurement model in the literature appropriate for categorical indicators. GD-GPIRT offers a method where ordinal indicators are used to estimate dynamic latent traits over time while making minimal assumptions about the IRF shapes. In addition, the GP priors on the time trends offer a balanced structure for inferring their dynamics, reducing the risk of poor identification due to scaling variance (Jackman, 2001). As shown in our experiments, GD-GPIRT estimates are superior in terms of model fit and measurement quality.

2 RELATED WORK

Item response theory (IRT) is a popular measurement framework in social science studies, such as computerized adaptive testing (Xu and Douglas, 2006), survey experiments (Muraki, 1990; Olino et al., 2012), and political ideology scaling (Poole and Rosenthal, 2000, 2001; Bafumi et al., 2005). Classic static and binary IRT methods usually rely on parametric assumptions for shapes of IRFs, including logistic relation (2PL and 3PL) for monotonic and symmetric IRFs (Molenaar, 1997; Mokken, 1971), graded unfolding structure for non-monotonicity (Roberts et al., 2000), and logistic positive exponential family for asymmetry (Samejima, 2000). Non-parametric IRT (NIRT) such as Mokken scaling (Mokken, 1971), monotone unidimensional model (Holland and Rosenbaum, 1986), and dimensionality assessment model (Stout, 1987) has emerged to address potential mismodeling in parametric IRT (Junker and Sijtsma, 2001). In addition, machine learning-based IRT (Chen et al., 2019; Cheng et al., 2019; Nguyen and Zhang, 2022) have also been developed but focusing on response prediction, and not yet been applied to dynamic IRT. Recently, Duck-Mayr et al. (2020) introduced a Bayesian non-parametric Gaussian process IRT to study voting patterns of the U.S. Congress, which relaxes the common monotonicity assumption in NIRT. There are also parametric IRT models for ordered-categorical responses, including graded response model (Samejima, 1997), graded unfolding models (Roberts and Laughlin, 1996; Roberts et al., 2000), generalized partial credit model (Muraki, 1992) and more (Agresti, 2003; Zumbo et al., 2007; Van Schuur, 2011; Bacci et al., 2014), but none have enjoyed the modeling flexibility of NIRT.

Dynamic latent variable models were combined with IRT to accommodate temporal shifts in latent traits. In political science, Poole and Rosenthal (1985) proposed an ideal-point spatial model (NOMINATE) for scaling congressional roll-call votes. Their model has been extensively used in studies of Congress, and was later extended to analyze ideological trends over multiple sessions by either assuming a simple polynomial time series model (Poole and Rosenthal, 2000) or estimating each session separately (Nokken and Poole, 2004). In legal studies, Martin and Quinn (2002) proposed a dynamic Bayesian measurement model (D-IRT) based on Bayesian random walk priors to study case dispositions of the U.S. Supreme Court, which was extended to ordinal responses by Schnakenberg and Fariss (2014) to study governmental respects of human rights. In educational research, Wang et al. (2013) applied dynamic linear models to IRT for computerized adaptive testing. There are other latent variable models for analyzing longitudinal panel data, such as the growth curve model (Rogosa et al., 1982; Curran and Hussong, 2003; Masyn et al., 2014) and autoregressive latent trajectory model (Bollen and Curran, 2004; Hamaker, 2005; Bollen and Zimmer, 2010), but these methods limit their modeling of latent traits to either strict linearity (Rogosa et al., 1982; Poole and Rosenthal, 2001; Hamaker, 2005; Bollen and Zimmer, 2010; Wang et al., 2013) or non-smooth autoregression

(Martin and Quinn, 2002; Bollen and Curran, 2004; Hamaker, 2005; Bollen and Zimmer, 2010; Schnakenberg and Fariss, 2014) with strong assumptions.

Gaussian Process latent variable model (GPLVM) is a family of algorithms that summarizes high-dimensional data into low-dimensional embeddings (Lawrence, 2003; Lawrence and Hyvärinen, 2005), and has found applications for data visualization (Jiang et al., 2012), manifold learning (Urtasun and Darrell, 2007; Titsias and Lawrence, 2010; Gao et al., 2010) and modeling dynamic systems (Wang et al., 2005; Lawrence and Moore, 2007; Damianou et al., 2011). However, directly applying GPLVMs to IRT is not appropriate because GPLVM usually marginalizes out the mappings and optimizes the latent variables (Lawrence and Hyvärinen, 2005), while IRT requires posterior inference of both. Previous attempts of improving GPLVM for static IRT include Urtasun and Darrell (2007) and Duck-Mayr et al. (2020), but significant research gap still persists in exploiting GPLVM for dynamic IRT with ordinal responses.

3 PROBLEM STATEMENT

Our statement of problem starts with stationary item response theory with ordinal responses, and then dynamic IRT in the longitudinal setting.

3.1 Stationary Item Response Theory

Consider the case of n respondents answering m different items, where response y_{ij} $(i = 1, \dots, n, j = 1, \dots, m)$ of the ith respondent and jth item belongs to an ordered category set $\mathcal{Y}_j = \{1, 2, \dots, C_j\}$ with total C_j levels. For example, all C_j will be 5 in the five-level Likert scale (Likert, 1932), where respondents may choose from "strongly (dis)agree", "(dis)agree" and "neutral". Item response theory states that the likelihood of observing y_{ij} is jointly determined by some respondent-level latent trait or ability score $x_i \in \mathcal{X}$ and item-level response function (IRF) $f_j : \mathcal{X} \to \mathcal{Y}_j$. For now we focus on the unidimensional latent space $\mathcal{X} = \mathbb{R}$, leaving higher dimensional $\mathcal{X}$ as future work. Dropping subscripts momentarily, the likelihood of observing level c is modeled as an ordered logistic with discrimination and difficulty parameter β_1 and β_0:

$$\boldsymbol{\beta} = [\beta_0, \beta_1]^T; \qquad f(x; \boldsymbol{\beta}) = \beta_1 x + \beta_0 \tag{1}$$

$$p\big(y = c \mid f, \{b_c\}\big) = \Phi(b_{c-1} - f) - \Phi(b_c - f) \tag{2}$$

where $\Phi(z)$ represents the cumulative density function of a standard normal. In addition, the latent function space is subset into C intervals, whose end points are denoted by $C + 1$ ordered threshold parameters $b_0 < b_1 < \cdots < b_C$. The interval $(b_{c-1}, b_c]$ on which the value of the function falls represents the range for the cth category. While $b_0 = -\infty$ and $b_C = \infty$ are fixed, b_1 to b_{C-1} can move freely under the ordered constraint. Intuitively, these $\{b_c\}$s control the shape of the categorical likelihood given latent function value. Note that all $\boldsymbol{\beta}$s and $\{b_c\}$s can be further indexed by j to represent their dependency on items, and determined by maximizing the joint likelihood $\prod_i \prod_j p\big(y_{ij} \mid f_j(x_i; \boldsymbol{\beta}_j), \{b_{jc}\}\big)$.

3.2 Dynamic Item Response Theory

In the longitudinal setting, respondents repeatedly answer potentially different sets of questions over multiple time periods, for instance, members of congress voting for different roll calls from session to session. Hence, dynamic structures in latent traits need to be accommodated for possible changes over repeated observations. For exposition, we assume items are different across time, as static items are special cases. We append an additional index for time t to the latent trait x_{it} and IRF f_{jt}, as well as its parameters $\boldsymbol{\beta}_{jt}$.

Some prior work simply estimates x_{it} separately for each time period, known as NOKKEN–POOLE scores in the application of ideology of Congress. For the same application, some adopt a polynomial structure for the dynamic latent traits $x_{it} \sim \text{poly}(t)$ and found that linear model is sufficient for capturing the majority of changes in the dynamic latent positions (Poole and Rosenthal, 2001; Bailey, 2007). Other works rely on Bayesian non-parametric methods for modeling non-polynomial latent traits (Martin and Quinn, 2002; Wang et al., 2013; Schnakenberg and Fariss, 2014). Broadly speaking, these non-parametric methods utilize the autoregressive (AR) model, which simplifies the inference of the whole dynamic trait trajectory to that at a single time period based on an informative prior:

$$x_{i,1} \sim \mathcal{N}(0, C_i) \tag{3}$$

$$x_{i,t} \sim \mathcal{N}(x_{i,t-1}, \sigma_t^2), \quad \forall \quad t = 2, \dots, T \tag{4}$$

where $\mathcal{N}(0, C_i)$ is the anchoring prior for the dynamic trait trajectory at time $t = 1$ and σ_t^2 is the diffusion variance. However, these AR models may encounter variance explosion as the prior variances accumulate through the diffusion terms over time, leading to possible overestimation of extremity in later periods (Martin and Quinn, 2002). While one may multiply $\sqrt{1 - \frac{\sigma_t^2}{C_i}}$ to $x_{i,t-1}$ to enforce variance balance, information of earlier observations summarized in the prior is discounted. In addition, the implicit Markov assumption impedes utilizing information from future observations in estimating current $x_{i,t}$, as future traits are not represented in the prior. Finally, AR trajectories tend to be rough and not well-suited to applica-

tions preferring smoother trends.

4 PROPOSED MODEL

In this section, we present a novel Bayesian approach for dynamic item response theory with ordinal responses. We leverage recent advances in Bayesian nonparametric IRT based on Gaussian process for inferring IRFs with flexible shapes, and propose a Gaussian process time series model for joint estimation of dynamic latent traits over time with balanced priors. As a Bayesian method, our model welcomes integration of any prior knowledge such as asymmetry and non-monotonicity to IRFs, or smoothness to the latent trait trends. We refer our model as generalized dynamic Gaussian process item response theory (GD-GPIRT).

4.1 Generalized Dynamic Gaussian Process Item Response Theory

Contrary to parametric approaches, Gaussian Process item response theory (GPIRT) makes minimal assumptions about the shape of IRFs except smoothness, so as to infer non-monotonicity or asymmetry (Duck-Mayr et al., 2020). Gaussian process (GP) is widely used for modeling distributions over functions, such that any realization of functional values has a joint Gaussian distribution. Specifically, GPIRT places a hierarchical GP prior on each latent f_{it}:

$$p(f_{jt}) \sim \mathcal{GP}(\mu_{jt}, K_x) \tag{5}$$
$$\mu_{jt}(x) = \beta_{jt1}x + \beta_{jt0} \tag{6}$$
$$K_x(x, x') = \exp\big(-\tfrac{1}{2}(x - x')^2/\ell_x^2\big) \tag{7}$$
$$\beta_{jt1} \sim \mathcal{N}(0, \sigma_{\beta_1}^2) \tag{8}$$
$$\beta_{jt0} \sim \mathcal{N}(0, \sigma_{\beta_0}^2) \tag{9}$$

where each latent function contains an item-specific linear trend $\mu_{jt}(\cdot)$ and a non-linear deviation with a prior $\mathcal{GP}(0, K_x)$. The slope and intercepts $\boldsymbol{\beta}$s in $\mu_{jt}(\cdot)$ have zero-mean normal priors with variance $\sigma_{\beta_1}^2$ and $\sigma_{\beta_0}^2$. The ℓ_x parameter is the length scale of the kernel controlling the bandwidth of correlations.

To accommodate dynamic latent traits, we also place independent GP priors on individual trait vector $\mathbf{x}_i = [x_{i1}, \ldots, x_{iT}]^T$. Here we use zero mean functions and Matérn kernel of degree 5/2 to model moderately smooth (twice differentiable) trajectories with length scale ℓ_t. Compared to the AR model discussed in Sec. 3.2, our GP model ensures measurement comparability as each entry in the dynamic latent trends will have the same marginal prior distribution as a standard normal.

$$p\big(\{\mathbf{x}_i\}\big) = \prod_i p(\mathbf{x}_i) \tag{10}$$
$$p(\mathbf{x}_i) \sim \mathcal{GP}(\mathbf{0}, \mathbf{K}_t) \tag{11}$$
$$K_t(d) = (1 + \frac{\sqrt{5}d}{\ell_t} + \frac{5d^2}{3\ell_t^2}) \exp\big(-\frac{\sqrt{5}d}{\ell_t}\big) \tag{12}$$

The last set of parameters is the threshold parameters. Following the reparameterization trick (Chu and Ghahramani, 2005), we define a set of positive padding variables $\Delta_l > 0$ and $l = 2, \ldots, C-1$ such that $b_c = b_1 + \sum_{l=2}^{c} \Delta_l$. We place independent standard normal prior on the log scale of the padding variables: $\log(\Delta_l) \sim \mathcal{N}(0, \sigma_\Delta^2)$. Note this is equivalently to placing normal priors on log scale of $b_1, b_2 - b_1, \ldots, b_{C-1} - b_{C-2}$.

4.2 Model Inference

In contrast to GPLVM models, GD-GPIRT places equal emphasis on the latent variables and mapping functions, whereas GPLVM models typically marginalize either of these components. Therefore, we propose a Markov chain Monte Carlo (MCMC) sampling procedure for both latent variables and IRFs. Specifically, our model parameters include all latent function values $\mathbf{f}_{jt} = [f_{1jt}, \ldots, f_{njt}]^T$, ability scores $\mathbf{x}_i$, slope and intercept parameters $\{\boldsymbol{\beta}_{jt}\}$ and threshold parameters $\{b_c\}$ with the following joint posterior distribution:

$$p\big(\{\mathbf{x}_i\}, \{\mathbf{f}_{jt}\}, \{b_c\}, \{\boldsymbol{\beta}_{jt}\} \mid \{y_{ijt}\}\big) \propto \underbrace{\prod_i p(\mathbf{x}_i)}_{\text{latent trait prior}} \underbrace{\prod_t \prod_j p(\mathbf{f}_{jt})}_{\text{IRF prior}} \underbrace{\prod_t \prod_j p(\beta_{jt1})}_{\text{slope prior}} \underbrace{\prod_t \prod_j p(\beta_{jt0})}_{\text{intercept prior}} \underbrace{\prod_c p(b_c)}_{\text{threshold prior}} \underbrace{\prod_t \prod_i \prod_j p\big(\{y_{ijt}\} \mid \{\mathbf{x}_i\}, \{\mathbf{f}_{jt}\}, \{b_c\}, \{\boldsymbol{\beta}_{jt}\}\big)}_{\text{likelihood}} \tag{13}$$

This joint posterior is highly multivariate, which makes direct sampling difficult. Hence, we proceed in a Gibbs sampling fashion. Let superscripts k denote the kth iteration in the MCMC sampler. The sampler is initialized by drawing $\{\mathbf{x}_i^{(0)}\}$s, $\{\boldsymbol{\beta}_{jt}^{(0)}\}$s and $\{b_c^{(0)}\}$s from their respective priors, and drawing the latent function values $\{\mathbf{f}_{jt}^{(0)}\}$s from the induced multivariate Gaussian at all n initial locations $\mathbf{x}_t^{(0)}$ at time t:

$$p\big(\mathbf{f}_{jt}^{(0)} \mid \mathbf{x}_t^{(0)}, \boldsymbol{\beta}_{jt}^{(0)}\big) \sim \mathcal{GP}\big(\mu_{jt}(\mathbf{x}_t^{(0)}; \boldsymbol{\beta}_{jt}^{(0)}), K_x(\mathbf{x}_t^{(0)}, \mathbf{x}_t^{(0)})\big) \tag{14}$$

After initialization, we alternatively sample each variable in the targeted joint distribution in Eq. (13) from

its conditioning distribution on all the other variables. First, the sampler draws new latent function values by conditioning on all $\{y_{ijt}\}$s:

$$p\big(\mathbf{f}_{jt}^{(k+1)} \mid \mathbf{x}_t^{(k)}, \{y_{ijt}\}, \boldsymbol{\beta}_{jt}^{(k)}, \{b_c^{(k)}\}\big) \propto p\big(\mathbf{f}_{jt}^{(k+1)} \mid \mathbf{x}_t^{(k)}, \boldsymbol{\beta}_{jt}^{(k)}\big) p\big(\{y_{ijt}\} \mid \mathbf{f}_{jt}^{(k+1)}, \{b_c^{(k)}\}\big) \quad (15)$$

As GP regression allows no analytical form for non-Gaussian likelihood, we exploit eclipse slice sampling (ESS) (Murray et al., 2010) for the conditional distributions. Eclipse slice sampling is a generic sampler for posterior of arbitrary target variable $\mathbf{z}$ with a Gaussian prior $\mathbf{z} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and a likelihood function $\mathcal{L}(\mathbf{z})$, and more efficient than the traditional Metropolis–Hastings stepping. ESS samples the next iteration by adaptively performing slicing sampling on the eclipse defined by current state $\mathbf{z}$ and a random draw $\boldsymbol{\nu}$ from prior $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ (see Supplement for details). The prior mean $\boldsymbol{\mu}_f$, covariance $\boldsymbol{\Sigma}_f$ and likelihood $\mathcal{L}(\mathbf{f_{jt}})$ for sampling $\mathbf{f}_{jt}^{(k+1)}$ are defined as:

$$\boldsymbol{\mu}_f = \mu_{jt}\big(\mathbf{x}_t^{(k)}; \boldsymbol{\beta}_{jt}^{(k)}\big), \quad \boldsymbol{\Sigma}_f = K_x\big(\mathbf{x}_t^{(k)}, \mathbf{x}_t^{(k)}\big) \quad (16)$$

$$\mathcal{L}(\mathbf{f}_{jt}) = \prod_i p\big(\{y_{ijt}\} \mid \mathbf{x}_t, \{f_{ijt}\}, \{b_c\}, \boldsymbol{\beta}_{jt}\big) \quad (17)$$

After obtaining $\mathbf{f}_{jt}^{(k+1)}$s for all items at all time periods, we then sample the latent trait $\mathbf{x}_i^{(k+1)}$s. However, direct sampling from the conditional distribution of $\mathbf{x}_i^{(k+1)}$ is not obvious, because the likelihood of locations is not defined by $\mathbf{f}_{jt}$s at points other than $\mathbf{x}_i^{(k+1)}$s. Hence, we introduce a set of auxiliary variables $\mathbf{f}_{jt}^*$, which are the latent functions values defined on an evenly-spaced dense grid $\mathbf{x}^*$ from -5 to 5 in one-hundred increment. Samples of $\mathbf{f}_{jt}^*$ can be obtained by applying GP posterior update rule, conditioning on current location $\mathbf{x}_t^{(k)}$s and function value $\mathbf{f}_{jt}^{(k+1)}$s:

$$p\big(\mathbf{f}^{*(k+1)}_{jt}\big) \sim \mathcal{GP}(\boldsymbol{\mu}^*, K^*) \quad (18)$$

$$\mathbf{V} = K_x\big(\mathbf{x}_t^{(k)}, \mathbf{x}_t^{(k)}\big) \quad (19)$$

$$\boldsymbol{\mu}^* = K_x\big(\mathbf{x}^*, \mathbf{x}_t^{(k)}\big)\mathbf{V}^{-1}\mathbf{f}_{jt}^{(k+1)} \quad (20)$$

$$K^* = K_x\big(\mathbf{x}^*, \mathbf{x}^*\big) - K_x\big(\mathbf{x}^*, \mathbf{x}_t^{(k)}\big)\mathbf{V}^{-1}K_x\big(\mathbf{x}_t^{(k)}, \mathbf{x}^*\big) \quad (21)$$

With these auxiliary variables $\mathbf{f}_{jt}^*$, we obtain a dense approximation of likelihood values for all latent locations besides $\mathbf{x}_t^{(k)}$. We construct mean $\boldsymbol{\mu}_x$, covariance $\boldsymbol{\Sigma}_x$ and likelihood $\mathcal{L}(\mathbf{x}_i)$ for sampling $\mathbf{x}_i^{(k+1)}$ as:

$$\boldsymbol{\mu}_x = \mathbf{0}, \quad \boldsymbol{\Sigma}_x = K_t\big(\mathbf{x}_i^{(k)}, \mathbf{x}_i^{(k)}\big) \quad (22)$$

$$\mathcal{L}(\mathbf{x}_i) = \prod_j \prod_t p\big(\{y_{ijt}\} \mid \mathbf{x}_i, \{\mathbf{f}_{jt}^*\}, \{b_c\}, \boldsymbol{\beta}_{jt}\big) \quad (23)$$

Note the latent trait location samples are rounded to the nearest rug in the dense grid $\mathbf{x}^*$. We then update the latent function values $\mathbf{f}_{jt}^{(k+1)}$ to those $\mathbf{f}^{*(k+1)}_{jt}$ defined on new $\mathbf{x}_t^{(k+1)}$. Finally, we sample new slope and intercept parameters $\{\boldsymbol{\beta}_{jt}^{(k+1)}\}$s and threshold parameters $\{b_c^{(k+1)}\}$s using ESS, based on the new latent locations $\mathbf{x}_t^{(k+1)}$ and updated function values $\mathbf{f}_{jt}^{(k+1)}$.

Our inference procedure can also be further adjusted when the latent item function $\mathbf{f}_{jt}$s come from the same set of items, meaning $\mathbf{f}_{j1} = \cdots = \mathbf{f}_{jt} = \mathbf{f}_j$ for all js. Now inference of $\mathbf{f}_j$ need to condition on all nT latent traits $\{x_{it}\}$s and corresponding observations $\{y_{ijt}\}$s, making sampling of auxiliary $\mathbf{f}^*{}_j$s computationally demanding. Hence, we exploit a sparse GP trick that selects 100 inducing locations on the dense grid whose inducing values are determined by its k-nearest neighbors. In our exploration, we found notable speed-up when $nT \approx 6,000$ but no performance lost.

5 EXPERIMENTS

We evaluate the measurement quality of GD-GPIRT in simulation studies, and then illustrate the advantages of GD-GPIRT in model fit with two real-world case studies regarding public opinions on economic environment and congressional ideology leaning on abortion issues.

5.1 Simulation Studies

Data generating process. The simulation consists of 100 synthetic respondents and 10 items over 10 time periods. We consider a variety of scenarios with binary ($C = 2$) or ordinal ($C = 5$) responses and whether the same or different set(s) of items are used across time. The latent trait vectors are drawn i.i.d from the zero-mean GP defined in Eq. (11) with $\ell_t = 5$, and the IRFs from the GP in Eq. (7) with $\ell_x = 1$ and $\sigma^2_{\beta_1} = \sigma^2_{\beta_1} = 1$. We also draw $\{b_c\}$s from Unif$[-2, 2]$, and demean, normalize and sort $\{b_c\}$s to ensure all ordinal responses have non-zero probabilities. Finally, we generate noisy responses y_{ij} from the probabilistic model defined in Eq. (2).

Baselines and metrics. We compare GD-GPIRT to several time-series baselines in dynamic IRT literature: 1) a naïve method that estimates each time period separately (N-GPIRT); 2) a linear trend method without non-linear deviation(L-GPIRT) ; and 3) the dynamic ordinal IRT (DO-IRT) model with AR trends (Schnakenberg and Fariss, 2014). We also consider three sets of metrics for evaluating measurement quality of estimated latent traits and IRFs as well as predictive fit of responses: 1) the averaged correlation and log likelihood of the estimated traits w.r.t the ground truth;

Table 1: Comparison of measurement quality and predictive fit between GD-GPIRT and baselines under various synthetic settings. Bold numbers indicate statistical significance compared to the other methods using standard paired t-tests and italicized numbers indicate the method is not statistically worse than the best method.

Binary Response ($C = 2$)								
	Same set of IRFs				Different sets of IRFs			
MEASURE	GD-GPIRT	L-GPIRT	N-GPIRT	DO-IRT	GD-GPIRT	L-GPIRT	N-GPIRT	DO-IRT
COR$(\mathbf{x}, \hat{\mathbf{x}})$	**0.949±0.003**	0.571±0.043	0.692±0.036	0.887±0.004	**0.961±0.005**	0.878±0.005	0.855±0.005	0.896±0.003
LL$(\mathbf{x}, \hat{\mathbf{x}})$	**−0.778±0.069**	−1.732±0.056	−1.232±0.037	−1.441±0.057	**−0.574±0.327**	−3.902±0.202	*−0.763±0.037*	−1.287±0.029
COR(ICC)	**0.884±0.014**	0.612±0.022	0.735±0.027	0.831±0.015	**0.915±0.004**	0.891±0.005	0.877±0.005	0.825±0.005
RMSE(ICC)	**0.116±0.003**	0.270±0.005	0.226±0.007	0.146±0.004	**0.089±0.003**	0.102±0.003	0.103±0.003	0.148±0.002
ACC$(y, \hat{y})$	*0.806±0.003*	0.774±0.002	**0.818±0.003**	0.754±0.003	**0.794±0.006**	0.674±0.009	0.708±0.010	0.751±0.007
LL$(y, \hat{y})$	*−1.381 ±0.060*	**−1.223±0.051**	*−1.353±0.054*	−1.762±0.087	*−1.235 ±0.086*	−3.376±0.524	**−1.014±0.062**	−1.919±0.171
Ordinal Response ($C = 5$)								
	Same set of IRFs				Different sets of IRFs			
MEASURE	GD-GPIRT	L-GPIRT	N-GPIRT	DO-IRT	GD-GPIRT	L-GPIRT	N-GPIRT	DO-IRT
COR$(\mathbf{x}, \hat{\mathbf{x}})$	**0.968±0.002**	0.658±0.045	0.637±0.041	0.913±0.004	**0.981±0.002**	0.867±0.008	0.927±0.004	0.920±0.002
LL$(\mathbf{x}, \hat{\mathbf{x}})$	**−0.834±0.085**	−1.730±0.050	−1.278±0.039	−1.797±0.142	*−1.379±0.864*	−11.62±1.436	**−0.452±0.055**	−1.461±0.067
COR(ICC)	**0.899±0.008**	0.649±0.022	0.677±0.028	0.833±0.015	**0.940±0.003**	0.843±0.018	*0.924±0.004*	0.805±0.004
RMSE(ICC)	**0.401±0.012**	0.914±0.018	0.846±0.023	0.440±0.012	**0.262±0.012**	0.444±0.035	*0.272±0.008*	0.449±0.006
ACC$(y, \hat{y})$	*0.507±0.004*	0.475±0.004	**0.517±0.004**	0.424±0.004	**0.455±0.011**	0.248±0.008	0.285±0.011	0.418±0.008
LL$(y, \hat{y})$	−2.977 ±0.049	**−2.596±0.042**	−2.966±0.047	−4.674±0.102	*−2.506 ±0.101*	*−2.527±0.326*	**−2.220±0.118**	−4.981±0.235

2) the correlation and RMSE of item characteristics curves (ICC), or the expected response given attribute $\text{ICC}(x; f, \mathbf{b}) = \mathbb{E}[y \mid f, \mathbf{b}, x]$; 3) the predictive accuracy ACC$(y, \hat{y})$ and log likelihood LL$(y, \hat{y})$ of responses.

Results. We split data into 80%/20% for training and testing. We repeat each simulation setting using 25 different seeds approximately 300 Intel Xeon 268 CPUs. For each run, we simulate three MCMC chains with 500 burnout steps and 500 sampling iterations, thinned every four samples. The averaged R-hat diagnostics for all variables are below 1.1 in all runs. Table 1 shows comparison of measurement quality and predictive fit between GD-GPIRT and baselines under various synthetic settings. Bold numbers indicate statistical significance compared to the other methods using standard paired t-tests and italicized numbers indicate the method is not statistically worse than the best method. Overall, our results consistently demonstrate the superiority of GD-GPIRT over the baseline methods in terms of the measurement quality of estimated traits and IRFs across all experimental conditions, while predicting no worse the (noisy) responses.

5.2 Public Opinions on Economy

The American Panel Survey (TAPS) was a long-running research project to study public opinions from all 50 states and included an extensive array of survey items asked across multiple waves.[1] Specifically, respondents were asked questions monthly from Jan. 2014 to Jan. 2018 about their opinions on the economic conditions and income allocation (either to spend or to save) of their households and the country as a whole. Since the same set of questions were repeatedly asked, we apply GD-GPIRT with sparse GP speedup to estimate people's financial status and how they react to those questions. We set $\ell_t = 12$ to capture yearly shifts in attitudes.

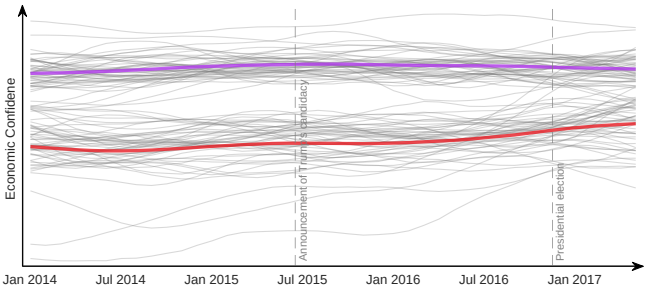


Figure 1: Illustration of public opinion trends in economic confidence. The grey lines represent individual confidence levels estimated using GD-GPIRT, while the bold blue and red lines depict the average confidence levels for Democrats and Republicans respectively. Vertical dashed lines mark key events such as the announcement of Trump's candidacy and the 2016 Presidential election. Notably, Democrats levels have remained relatively stable, while Republicans exhibited a slight increase in confidence leading up to and especially following the election of President Trump.

Figure 1 illustrates public opinion trends in economic confidence. The grey lines represent individual confidence levels estimated using GD-GPIRT, while the bold blue and red lines depict the average confidence levels among Democrats and Republicans groups. Vertical dashed lines mark key events such as the announcement of Trump's candidacy and the 2016 Presidential

[1] TAPS is conducted by Weidenbaum Center at Washington University in St. Louis.

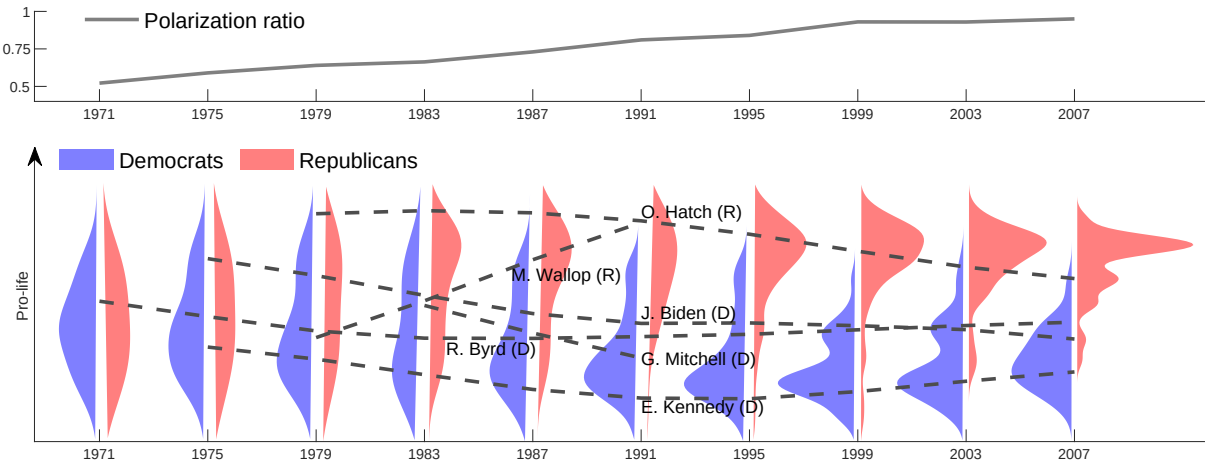


Figure 2: The upper panel illustrates the alignment between senators' estimated ideology and their party affiliations ('polarization ratio') on abortion issues using GD-GPIRT. The increase in the ratio from the 92nd Congress (0.521) to the 108th Congress (0.950) indicates the growing partisan divide on abortion. The lower panel displays GD-GPIRT scores by party spaced every two sessions, with Dem. and Rep. Senators in blue and red, respectively. Dashed lines represent the evolving ideological trajectories of selected senators.

election. Notably, the Democrats' confidence levels remained relatively stable, while Republicans exhibited a slight increase in confidence leading up to and following the election of President Trump. This is consistent with existing theories of the partisan sources of confidence. We provide estimation of selective IRFs in Supplement. In general, as economic confidence diminishes, respondents are more inclined to disagree that economic conditions are improving, and they have reduced expectations for savings.

ACC	FORECASTING HORIZON		
MODEL	1 month	6 months	12 months
OURS	**0.597±0.005**	**0.665±0.008**	**0.572±0.005**
DO-IRT	0.573±0.006	0.604±0.007	0.538±0.007
LL	FORECASTING HORIZON		
MODEL	1 month	6 month	12 month
OURS	**−1.003±0.008**	*−0.916±0.008*	**−1.074±0.013**
DO-IRT	−1.144±0.025	−0.957±0.021	−1.276±0.024

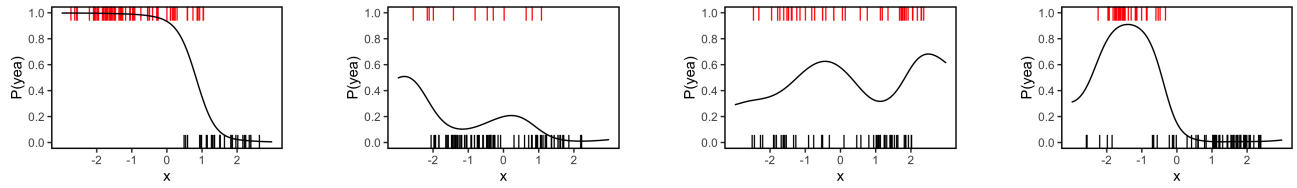
Table 2: Predictive accuracy and log likelihood of GD-GPIRT and DO-IRT in predicting future responses at various forecasting horizons. GD-GPIRT significantly outperforms DO-IRT for the majority of forecasting horizons.

In order to assess the predictive capabilities of GD-GPIRT with respect to future responses, we conducted an additional forecasting analysis focused on out-of-sample predictions of actual responses. Specifically, we first estimate confidence levels of each individual based on data spanning from 2014 to 2017, and then extrapolate their confidence levels in 2018. We then hold out 20% of distinct individuals for every question, leverage the remaining 80% of observations to estimate IRFs and predict their future responses from the extrapolated confidence levels across multiple forecasting horizons ranging from 1 to 12 months. Table 2 shows the predictive accuracy and log likelihood of GD-GPIRT and DO-IRT in forecasting future responses at various horizons. Our findings show that GD-GPIRT significantly outperforms DO-IRT for the majority of forecasting horizons, suggesting effectiveness of GD-GPIRT in modeling the trajectories of confidence levels and its superiority in measurement quality.

5.3 Ideology of Senate on Abortion

Ideology or the configuration of interconnected beliefs and attitudes (Converse, 2006), plays a central role in understanding congressional dynamics such as political polarization and partisan sorting (Poole and Rosenthal, 2001; Fiorina et al., 2008). Scaling congressional votes to ideology faces challenges in accommodating temporal changes while ensuring comparability, as politicians have demonstrated substantial shifts in their views over time (CNN, 2019, 2022). However, previous studies often simplified complex ideology trajectories with linear models or low order polynomials (Poole and Rosenthal, 2001; Bailey, 2007, 2013).

We run GD-GPIRT to estimate the U.S. Senate's ideology over multiple congressional sessions, using roll-call voting data obtained from the *voteview* database (Lewis et al., 2019). We focus on votes related to



(a) Public Health Service Act to permit freedom of access to medical facilities.

(b) Authorization of extra $5M funds for family planning project grants.

(c) Expressing the sense of Congress regarding forced abortions in China.

(d) Prohibition of abortion where fetus is determined to be viable.

Figure 3: IRFs of four selected roll call votes in the U.S. Senate on abortion issues between the 92th to 108th Congress that are standard, asymmetric, non-saturate and non-monotonic. Estimated probability of voting "yea" is plotted against ideology score x. Actual "yea" and "nay" roll-call votes are displayed as red and black vertical dashes.

abortion as identified in Montgomery (2011), resulting in 758k total votes of 235 Senators spanning over 20 years. We set $\ell_t = 6$, the maximum-a-posterior estimate from a simple GP model of NOKKEN–POOLE scores (Nokken and Poole, 2004) against time.

Our findings reveal a clear and noticeable pattern of increasing partisan polarization within the United States Congress on abortion issues. In Figure 2, the upper panel illustrates the alignment between senators' estimated ideology and their party affiliations, denoted as 'polarization ratio', on abortion issues under GD-GPIRT. The increase in the ratio from the 92nd Congress (0.521) to the 108th Congress (0.950) indicates the growing partisan divide on abortion. The lower panel displays GD-GPIRT scores by party spaced every two sessions, with Dem. and Rep. Senators in blue and red, respectively. Dashed lines represent the evolving ideological trajectories of selected senators. The prominent party divide emerging in the 1990s underscores the partisan sorting related to abortion, aligned with prior research (Brady and Schwartz, 1995; Adams, 1997). Furthermore, GD-GPIRT is capable of inferring IRFs beyond standard logistic shapes. Figure 3 shows IRFs for four selected roll call votes in the US Senate between the 92th to 108th Congress. The estimated probability of voting "yea" is plotted against the ideology score x. The actual "yea" and "nay" roll-call votes are displayed as red and black vertical dashes. From left to right, these IRFs are either standard, asymmetric, non-monotonic, or non-saturate (do not approach zero or one).

Table 3: Comparison of in-sample model fits between GD-GPIRT and baselines for binary roll-call vote data.

	$\mathcal{L}/N$	Acc	AUC
GD-GPIRT	**−0.160**	**0.930**	**0.930**
DO-IRT	−0.402	0.825	0.818
NOKKEN–POOLE	−0.557	0.733	0.730

We also assess the predictive performance of GD-GPIRT regarding actual votes. Table 3 shows the comparison of in-sample model fits between GD-GPIRT and baselines for binary roll-call vote data. On average, GD-GPIRT correctly predicts 93% of the votes, which is significantly higher than prediction from DO-IRT and NOKKEN–POOLE scores. Besides, GD-GPIRT outperforms baselines in averaged log likelihood and area under the receiver operating characteristic curve (AUC).

6 CONCLUSION

We propose GD-GPIRT, the first dynamic Bayesian non-parametric measurement model for longitudinal categorical observations, to estimate dynamic latent traits while making minimal assumptions about shapes of the response functions. We validate GD-GPIRT and the sampler in simulation studies, and demonstrate its superiority in both model fit and measurement quality against baselines. Lastly, we apply GD-GPIRT to address substantive problems, including assessing public opinions on economy environment and estimating trends in congressional ideology leaning on abortion.

We see potentials of GD-GPIRT in the advancement of IRT in several ways. Firstly, GD-GPIRT can be extended to model multi-dimensional latent traits, which is particularly relevant in fields such as political science where traits like the Big Five personality traits (Gerber et al., 2011) and the 2-d NOMINATE scores (Poole and Rosenthal, 2001) are essential. In addition, by clustering models such as mixtures of GPs with Dirichlet process prior, the conditional independence assumption among individual traits may be further relaxed when certain participating units naturally form subgroups. Finally, although the MCMC sampling method technique has proven sufficient in our experiments, one may emplore other Bayesian variational techniques for inference in GD-GPIRT.

References

Adams, G. D. (1997). Abortion: Evidence of an issue evolution. *American Journal of Political Science*, pages 718–737.

Agresti, A. (2003). *Categorical Data Analysis*. John Wiley & Sons.

Bacci, S., Bartolucci, F., and Gnaldi, M. (2014). A Class of Multidimensional Latent Class IRT Models for Ordinal Polytomous Item Responses. *Communications in Statistics-Theory and Methods*, 43(4):787–800.

Bafumi, J., Gelman, A., Park, D. K., and Kaplan, N. (2005). Practical Issues in Implementing and Understanding Bayesian Ideal Point Estimation. *Political Analysis*, 13(2):171–187.

Bailey, M. A. (2007). Comparable preference estimates across time and institutions for the court, congress, and presidency. *American Journal of Political Science*, 51(3):433–448.

Bailey, M. A. (2013). Is Today's Court the Most Conservative in Sixty Years? Challenges and Opportunities in Measuring Judicial Preferences. *The Journal of Politics*, 75(3):821–834.

Baylari, A. and Montazer, G. A. (2009). Design a personalized e-learning system based on item response theory and artificial neural network approach. *Expert Systems with Applications*, 36(4):8013–8021.

Beetham, D. (1999). *Democracy And Human Rights*, volume 249. Polity Press Cambridge.

Bergner, Y., Droschler, S., Kortemeyer, G., Rayyan, S., Seaton, D., and Pritchard, D. E. (2012). Model-Based Collaborative Filtering Analysis of Student Response Data: Machine-Learning Item Response Theory. *International Educational Data Mining Society*.

Bollen, K. A. (1980). Issues in the Comparative Measurement of Political Democracy. *American Sociological Review*, pages 370–390.

Bollen, K. A. and Curran, P. J. (2004). Autoregressive Latent Trajectory (ALT) Models A Synthesis of Two Traditions. *Sociological Methods & Research*, 32(3):336–383.

Bollen, K. A. and Zimmer, C. (2010). An Overview of the Autoregressive Latent Trajectory (ALT) Model. *Longitudinal Research with Latent Variables*, pages 153–176.

Brady, D. and Schwartz, E. P. (1995). Ideology and Interests in Congressional Voting: The Politics of Abortion in the U.S. Senate. *Public Choice*, 84(1):25–48.

Chen, C.-M., Lee, H.-M., and Chen, Y.-H. (2005). Personalized e-learning system using Item Response Theory. *Computers & Education*, 44(3):237–255.

Chen, Y., Silva Filho, T., Prudencio, R. B., Diethe, T., and Flach, P. (2019). β^3-IRT: A New Item Response Model and its Applications. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1013–1021. PMLR.

Cheng, S., Liu, Q., Chen, E., Huang, Z., Huang, Z., Chen, Y., Ma, H., and Hu, G. (2019). DIRT: Deep Learning Enhanced Item Response Theory for Cognitive Diagnosis. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 2397–2400.

Chu, W. and Ghahramani, Z. (2005). Gaussian Processes for Ordinal Regression. *Journal of Machine Learning Research*, 6(35):1019–1041.

Chung, J., Kastner, K., Dinh, L., Goel, K., Courville, A. C., and Bengio, Y. (2015). A Recurrent Latent Variable Model for Sequential Data. *Advances in Neural Information Processing Systems*, 28.

Cingranelli, D. L. and Richards, D. L. (2002). Measuring the Level, Pattern, and Sequence of Government Respect for Physical Integrity Rights. *International Studies Quarterly*, 43(2):407–417.

CNN (2019). Joe Biden described being an 'odd man out' with Democrats on abortion in 2006 interview.

CNN (2022). Biden's political and personal evolution on abortion on display after publication of draft Supreme Court opinion.

Converse, P. E. (2006). The nature of belief systems in mass publics (1964). *Critical Review*, 18(1-3):1–74.

Curran, P. J. and Hussong, A. M. (2003). The Use of Latent Trajectory Models in Psychopathology Research. *Journal of Abnormal Psychology*, 112(4):526.

Damianou, A., Titsias, M., and Lawrence, N. (2011). Variational Gaussian Process Dynamical Systems. *Advances in Neural Information Processing Systems*, 24.

Donnelly, J. (1999). Human Rights, Democracy, and Development. *Human Rights Quarterly*, 21(3):608–632.

Duck-Mayr, J., Garnett, R., and Montgomery, J. (2020). GPIRT: A Gaussian Process Model for Item Response Theory. In *Conference on Uncertainty in Artificial Intelligence*, pages 520–529. PMLR.

Duck-Mayr, J. and Montgomery, J. (2022). Ends Against the Middle: Measuring Latent Traits When Opposites Respond the Same Way for Antithetical Reasons. *Political Analysis. Conditionally Accepted*.

Fiorina, M. P., Abrams, S. A., and Pope, J. C. (2008). Polarization in the American Public: Misconceptions and Misreadings. *The Journal of Politics*, 70(2):556–560.

Gao, X., Wang, X., Tao, D., and Li, X. (2010). Supervised Gaussian Process Latent Variable Model for Dimensionality Reduction. *IEEE transactions on systems, man, and cybernetics, Part B (Cybernetics)*, 41(2):425–434.

Gerber, A. S., Huber, G. A., Doherty, D., and Dowling, C. M. (2011). The Big Five Personality Traits in the Political Arena. *Annual Review of Political Science*, 14:265–287.

Hamaker, E. L. (2005). Conditions for the Equivalence of the Autoregressive Latent Trajectory Model and a Latent Growth Curve Model With Autoregressive Disturbances. *Sociological Methods & Research*, 33(3):404–416.

Holland, P. W. and Rosenbaum, P. R. (1986). Conditional Association and Unidimensionality in Monotone Latent Variable Models. *The Annals of Statistics*, pages 1523–1543.

Jackman, S. (2001). Multidimensional Analysis of Roll Call Data via Bayesian Simulation: Identification, Estimation, Inference, and Model Checking. *Political Analysis*, 9(3):227–241.

Jiang, X., Gao, J., Wang, T., and Zheng, L. (2012). Supervised Latent Linear Gaussian Process Latent Variable Model for Dimensionality Reduction. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 42(6):1620–1632.

Junker, B. W. and Sijtsma, K. (2001). Nonparametric Item Response Theory in Action: An Overview of the Special Issue. *Applied Psychological Measurement*, 25(3):211–220.

Lalor, J. P., Wu, H., and Yu, H. (2016). Building an Evaluation Scale using Item Response Theory. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2016, page 648. NIH Public Access.

Lawrence, N. (2003). Gaussian Process Latent Variable Models for Visualisation of High Dimensional Data. *Advances in Neural Information Processing Systems*, 16.

Lawrence, N. and Hyvärinen, A. (2005). Probabilistic Non-linear Principal Component Analysis with Gaussian Process Latent Variable Models. *Journal of Machine Learning Research*, 6(11).

Lawrence, N. D. and Moore, A. J. (2007). Hierarchical Gaussian Process Latent Variable Models. In *Proceedings of the 24th International Conference on Machine Learning*, pages 481–488.

Lewis, J. B., Poole, K., Rosenthal, H., Boche, A., Rudkin, A., and Sonnet, L. (2019). Voteview: Congressional roll-call votes database. *See* `https://voteview.com/` *(accessed 27 July 2018)*.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*.

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

Martínez-Plumed, F., Prudêncio, R. B., Martínez-Usó, A., and Hernández-Orallo, J. (2019). Item response theory in AI: Analysing machine learning classifiers at the instance level. *Artificial Intelligence*, 271:18–42.

Masyn, K. E., Petras, H., and Liu, W. (2014). Growth Curve Models with Categorical Outcomes. *Encyclopedia of Criminology and Criminal Justice*, 2013.

Mokken, R. J. (1971). A Theory and Procedure of Scale Analysis with Applications in Political Research. In *A Theory and Procedure of Scale Analysis* . De Gruyter Mouton.

Molenaar, I. (1997). Nonparametric models for polytomous responses. In *Handbook of Modern Item Response theory*, page 367–380. Springer.

Montgomery, J. (2011). *An Evolutionary Theory of Democracy: Dynamic Evolutionary Models of American Party Compteititon with an Emprical Application to the Case of Abortion Policy from 1972-2010*. PhD thesis, Duke University, Durham, North Carolina.

Muraki, E. (1990). Fitting a Polytomous Item Response Model to Likert-Type Data. *Applied Psychological Measurement*, 14(1):59–71.

Muraki, E. (1992). A Generalized Partial Credit Model: Application of an EM Algorithm. *Applied Psychological Measurement*, 16(2):159–176.

Murray, I., Adams, R., and MacKay, D. (2010). Elliptical slice sampling. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 541–548. JMLR Workshop and Conference Proceedings.

Nguyen, D. and Zhang, A. Y. (2022). A Spectral Approach to Item Response Theory. *Advances in Neural Information Processing Systems*, 35:38818–38830.

Nokken, T. P. and Poole, K. T. (2004). Congressional Party Defection in American History. *Legislative Studies Quarterly*, 29(4):545–568.

Olino, T. M., Yu, L., Klein, D. N., Rohde, P., Seeley, J. R., Pilkonis, P. A., and Lewinsohn, P. M.

(2012). Measuring depression using item response theory: an examination of three measures of depressive symptomatology. *International Journal of Methods in Psychiatric Research*, 21(1):76–85.

Park, J. Y., Dedja, K., Pliakos, K., Kim, J., Joo, S., Cornillie, F., Vens, C., and Van den Noortgate, W. (2023). Comparing the prediction performance of item response theory and machine learning methods on item responses for educational assessments. *Behavior Research Methods*, 55(4):2109–2124.

Poole, K. T. and Rosenthal, H. (1985). A Spatial Model for Legislative Roll Call Analysis. *American Journal of Political Science*, 29(2):357–384.

Poole, K. T. and Rosenthal, H. (2000). *Congress: A Political-Economic History of Roll Call Voting*. Oxford University Press on Demand.

Poole, K. T. and Rosenthal, H. (2001). D-nominate after 10 years: A comparative update to congress: A political-economic history of roll-call voting. *Legislative Studies Quarterly*, 26(1):5–29.

Proust-Lima, C., Philipps, V., Perrot, B., Blanchin, M., and Sébille, V. (2022). Modeling repeated self-reported outcome data: A continuous-time longitudinal Item Response Theory model. *Methods*, 204:386–395.

Roberts, J. S., Donoghue, J. R., and Laughlin, J. E. (2000). A General Item Response Theory Model for Unfolding Unidimensional Polytomous Responses. *Applied Psychological Measurement*, 24(1):3–32.

Roberts, J. S. and Laughlin, J. E. (1996). THE GRADED UNFOLDING MODEL: A UNIDIMENSIONAL ITEM RESPONSE MODEL FOR UNFOLDING GRADED RESPONSES. *ETS Research Report Series*, 1996(1):i–60.

Rogosa, D., Brandt, D., and Zimowski, M. (1982). A growth curve approach to the measurement of change. *Psychological bulletin*, 92(3):726.

Samejima, F. (1997). Graded Response Model. In *Handbook of Modern Item Response theory*, pages 85–100. Springer.

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

Schnakenberg, K. E. and Fariss, C. J. (2014). Dynamic Patterns of Human Rights Practices. *Political Science Research and Methods*, 2(1):1–31.

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika*, 52(4):589–617.

Titsias, M. and Lawrence, N. D. (2010). Bayesian Gaussian Process Latent Variable Model. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 844–851. JMLR Workshop and Conference Proceedings.

Treier, S. and Jackman, S. (2008). Democracy as a Latent Variable. *American Journal of Political Science*, 52(1):201–217.

Urtasun, R. and Darrell, T. (2007). Discriminative Gaussian process latent variable model for classification. In *Proceedings of the 24th International Conference on Machine Learning*, pages 927–934.

Van der Ark, L. A. (2007). Mokken Scale Analysis in R. *Journal of Statistical Software*, 20:1–19.

Van Schuur, W. H. (2011). *Ordinal Item Response Theory: Mokken Scale Analysis*. Sage.

Wang, J., Hertzmann, A., and Fleet, D. J. (2005). Gaussian Process Dynamical Models. *Advances in Neural Information Processing Systems*, 18.

Wang, X., Berger, J. O., and Burdick, D. S. (2013). Bayesian analysis of dynamic item response models in educational testing. *The Annals of Applied Statistics*, 7(1):126–153.

Wilcox, C., Sigelman, L., and Cook, E. (1989). SOME LIKE IT HOT INDIVIDUAL DIFFERENCES IN RESPONSES TO GROUP FEELING THERMOMETERS. *Public Opinion Quarterly*, 53(2):246–257.

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

Zumbo, B. D., Gadermann, A. M., and Zeisser, C. (2007). Ordinal Versions of Coefficients Alpha and Theta for Likert Rating Scales. *The Journal of Modern Applied Statistical Methods*, 6(1):4.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Not Applicable]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Yes]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Yes]
   (d) Information about consent from data providers/curators. [Yes]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Yes]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]