Using Item Response Theory in a Political Setting

Rebekah Fearnhead

1 Introduction

Item Response Theory (IRT) Models, also known as Latent Trait Models, are used in many areas of cognitive and behavioural measurement. They are especially used in the construction and evaluation of educational tests, and sometimes also the scoring of them [2]. However, as well as these main applications, they can also be used in a wide variety of other areas including psychology, marketing, and politics [1, 6, 9].

IRT models are typically used with categorical data and are probabilistic models for individuals' responses to a set of items or questions. These models are based on latent factor models which classify the individuals into groups based on 'traits' where everyone with the same trait behaves in the same way [5]. IRT models are also similar to linear factor models [16], however, whilst these assume the observed variables are continuous, IRT models focus on categorical variables.

One of the most common parametric forms of an IRT model is the two-parameter logistic (2PL) model [14] which is widely used in educational settings.

2 IRT Model Assumptions

To understand the modelling framework, we can focus on one of the most popular contexts that this model is used in - educational testing. In this context we can assume that we have N individuals taking a test which consists of J items. We can then represent individual i's response to item j as a binary random variable Y_{ij} where a value of 1 indicates a correct response, and $Y_{ij} = 0$ otherwise. We can also define y_{ij} to denote a realisation of Y_{ij}.

The IRT model assumes that the $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{iJ})^\top$ are independent and aims to model the joint distribution of the random vector $\mathbf{Y}_i$. This is done by assuming one latent variable for each individual, denoted by θ_i. This can be thought of as the individual's level for the ability (latent trait) which is being measured in the test. We also assume these latent trait levels completely characterise each individual's response patterns.

This is reflected by the conditional distribution of $\mathbf{Y}_i$ given θ_i with the specification relying on two assumptions [5]. The first is that there is conditional independence of $Y_{i1}, \ldots, Y_{iJ}$ given

the latent trait level θ_i. The second is an assumption on the Item Response Function (IRF) or Item Characteristic Curve (ICC) which is defined as $g_j(\theta|\boldsymbol{\pi}_j) := P(Y_{ij} = 1|\theta_i = \theta)$ with $\boldsymbol{\pi}_j$ representing the parameters of item j.

2.1 Two-Parameter Logistic Model

The two-parameter logistic (2PL) model [14] is a common parametric form of the IRT model and is widely used. We need to define $g_j(\theta|\boldsymbol{\pi}_j)$ which for this model is:

$$g_j(\theta|\boldsymbol{\pi}_j) = \frac{\exp(d_j + a_j\theta)}{1 + exp(d_j + a_j\theta)}, \tag{1}$$

where the item parameters $\boldsymbol{\pi}_j = (a_j, d_j)$. A unidimensional IRT model such as the 2PL model assumes that Y_{ij} is equal to $g_j(\theta|\boldsymbol{\pi}_j)$ plus some noise which is often referred to as the measurement error.

In the educational setting as described above, the IRF is often assumed to be a monotonically increasing function meaning that $a_j > 0$ in the 2PL model. This can be interpreted as higher latent trait levels (ability in the subject tested) leading to a higher chance of answering an item correctly. There are however examples of other contexts that IRT models are used in which do not hold this property. The other parameter, d_j is known as the easiness parameter.

2.1.1 Assumptions on θ_i

To complete the specification of this model, assumptions need to be imposed on the θ_i. There are two different regimes that can be considered that lead to different parameter spaces [7]. The 'stochastic sampling' regime [19] treats each θ_i as a unknown parameter to be estimated from data and leads to a joint likelihood function being used [13].

On the other hand, the 'random sampling' regime assumes the θ_is are independent and identically distributed samples from a population with density f with respect to a dominating measure μ. This means that the distribution function f is estimated from the data instead of the individual values of the θ_is. This leads to a marginal likelihood function being used [3].

3 IRT Models with Covariates

Sometimes, along with the item response data, covariates of the individuals can also be collected. The p-dimensional observed covariates of an individual i can be represented by $\mathbf{x}_i$. There are different ways that these covariates can be incorporated in the IRT model depending on their affect on the latent traits and the model as a whole.

The simplest example of this is the covariates affecting the distributions of the latent traits [17], however, in some models they can also be viewed to affect the responses for some

of the items. This leads to a type of model known as a Multiple Indicators, Multiple Causes (MIMIC) model [18]. These models are used to study differential item functioning (DIF) [8] which is where items may function differently, or even measure different things, for one group of individuals compared to another.

We can start by considering the case of a single binary covariate, x_i which indicates the group membership of each individual. A MIMIC model [10] then allows the IRFs of the DIF items to depend on the group memberships while enabling each of the two groups to have different distributions for their latent traits. This model can also be expanded to having more than two groups.

If we assume that the item j is the only one out of the J items that is a DIF item, we can use 2PL model framework from equation 1 to model the IRF of item j as [23]:

$$g_j(\theta|\boldsymbol{\pi}_j) = \frac{\exp(d_j + a_j\theta + \delta_j x_i)}{1 + exp(d_j + a_j\theta + \delta_j x_i)}. \tag{2}$$

In equation 2, the parameter δ_j characterises the group effect on the IRF. The baseline (or reference) group is represented by $x_i = 0$. If $\delta_j = 0$ this means that this is not a DIF item, so members from the different groups act the same when faced with these items. On the other hand, if $\delta_j \neq 0$, this means that item j is a DIF item. The latent trait distribution can be modelled by setting $\theta_i|x_i \sim N(\beta x_i, 1)$ instead of a standard normal distribution, where β is a vector of coefficients.

In the testing setting, δ_j is usually positive for DIF items meaning that the second group performs better on these items compared to the reference group. For example, if the individuals in the group where $x_i = 1$ are believed to have cheated on a test, this will mean that they are more likely to answer the questions correctly. This model also means that for the baseline group, the IRF is represented by $a_j\theta_i + d_j$, which is the same as their form in the standard 2PL model.

In real-world analysis, however, the DIF items are unknown and therefore need to be detected based on the data provided. This can be done by casting the DIF analysis into a model selection problem.

3.1 The DIF-effect Parameter

The DIF-effect parameter, δ_j characterises how the individuals in the second group differ from those in the reference group, where $x_i = 0$, for a specific item j. As $\delta_j x_i = 0$ for the reference group for all of the items, this is able to serve as a reference point to compare the behaviours of the other group.

This DIF-effect parameter can also be expressed in terms of log-odds under the 2PL model

[23]:

$$\delta_j = \log\left(\frac{P(Y_{ij}=1|\theta_i=\theta, x_i=1)/(1-P(Y_{ij}=1|\theta_i=\theta, x_i=1))}{P(Y_{ij}=1|\theta_i=\theta, x_i=0)/(1-P(Y_{ij}=1|\theta_i=\theta, x_i=0))}\right). \tag{3}$$

This means that δ_j is the log-odds-ratio when comparing two respondents, one from the group where $x_i = 1$, and the other from the reference group, given that they have the same latent construct level.

3.2 Expanding to More Groups

This method can be adjusted to allow for there to be more than two different groups that the individuals can be classified in. Let us assume that the respondents are from $K+1$ unobserved groups, and we can represent the group membership using a latent variable, $\xi_i \in \{0, 1, \ldots, K\}$. The DIF-effect parameter can then be represented as $\delta_{j\xi_i}$, and the 2PL model in equation 2 can be changed to:

$$g_j(\theta|\boldsymbol{\pi}_j) = \frac{\exp(d_j + a_j\theta + \delta_{j\xi_i})}{1 + exp(d_j + a_j\theta + \delta_{j\xi_i})}. \tag{4}$$

For the baseline group, $\xi_i = 0$, we can then set $\delta_{j0} = 0$ for all $j = 1, \ldots, J$ so that $a_j\theta + d_j$ denotes the IRF for this group.

This means that δ_{jk} is used to characterise the individuals in group k differ to the reference group based on the item response behaviour for item j. For the latent classes that are not the reference group, this DIF parameter can be non-zero for some items where the behaviour differs between groups. The magnitude of the parameter is also allowed to differ across the latent classes because of the possibility for varying degrees of DIF effects across the different groups.

3.3 Structural Model

Using the formulation for having $K + 1$ unobserved groups, we can then specify a joint distribution for the latent variables (θ_i, ξ_i) [23]. The latent classes, ξ_i can be assumed to follow a categorical distribution.

$$\xi_i \sim \text{Categorical}(\{0, 1, \ldots, K\}, (\nu_0, \nu_1, \ldots, \nu_K)), \tag{5}$$

where $\nu_k = P(\xi_i = k)$ such that $\nu_k \geq 0$ and $\sum_{k=0}^{K} \nu_k = 1$.

We can then assume that the latent variable θ_i conditional on the latent class ξ_i follows a

normal distribution with the mean and variance specified by the class:

$$\theta_i|\xi_i = k \sim N(\mu_k, \sigma_k^2). \tag{6}$$

We also want to fix the mean and variance of the reference group to $\mu_0 = 0$ and $\sigma_0^2 = 1$ to ensure model identification [20].

3.3.1 Form of the Marginal Likelihood

In this model, both the latent classes, ξ_i and the latent variables, θ_i are unobserved. This means that the inference on the proposed model is based on the marginal likelihood with both ξ_i and θ_i marginalised out. For the 2PL model, this marginal likelihood function has a form of [23]:

$$L(\Delta) = \prod_{i=1}^{N} \sum_{k=0}^{K} \nu_k \int \left(\prod_{j=1}^{J} (\exp((a_j\theta + d_j + \delta_{jk})Y_{ij})/(1 + \exp(a_j\theta + d_j + \delta_{jk}))) \right) \phi(\theta|\mu_k, \sigma_k^2) d\theta, \tag{7}$$

where $\phi(\theta|\mu_k, \sigma_k^2)$ denotes the density function of a normal distribution with mean μ_k and variance σ_k^2. We also let the vector Δ denote all the known parameters, a_j, d_j, δ_{jk}, ν_k, μ_k and σ_k^2.

3.4 Model Selection and Estimation

When solving the DIF analysis problem, it is helpful to adopt the sparsity assumption [15] that assumes that for many of the DIF parameters, $\delta_{jk} = 0$. In many applications, the number of DIF items is low, so this is a meaningful assumption.

Under this assumption, we can use a L_1 regularised estimator to both learn the sparsity pattern of the DIF-effect parameters, and estimate the unknown model parameters. This estimator has the form:

$$\begin{aligned} \tilde{\Delta}^{(\lambda)} &= \arg_{\Delta} \min - \log L(\Delta) + \lambda \sum_{j=1}^{J} \sum_{k=1}^{K} |\delta_{jk}|, \\ \text{s.t. } \nu_k \geq 0, \;\; k = 0, 1, \ldots, K, \;\; &\text{and } \sum_{k=0}^{K} \nu_k = 1, \end{aligned} \tag{8}$$

where $L(\Delta)$ is defined in equation 7 as the marginal likelihood function, and $\lambda > 0$ is the tuning parameter.

The L_1 regularisation term $\lambda \sum_{j=1}^{J} \sum_{k=1}^{K} |\delta_{jk}|$ behaves similar to Lasso regression [22] as it tends to shrink some of the DIF-effect parameters to zero, and in the extreme case when

$\lambda \to \infty$, all the parameters will shrink to zero. When λ is chosen properly, this estimator yields estimation and selection consistency [24]. This means that the latent trait is identified correctly and the estimated DIF-effect parameters can be used to classify each item as either being a DIF or a non-DIF item.

The tuning parameter λ can be selected by using the Bayesian Information Criterion (BIC) [21] by using a grid search approach. The tuning parameter, $\hat{\lambda}$ can then be selected by finding the value of λ that minimises the BIC.

4 Simulation Study

We can look at how this method performs by simulating data to represent a scenario where the number of latent classes, is fixed and known as being 2. To do this we need to first simulate the data to represent each of the individual's behaviour and latent traits, and then we can use this to perform analysis.

4.1 Simulation Settings

We first need to set up the simulation parameters. In this simulation we set the number of individuals, N to be 1000 and the number of items, J to be 25. We also set the number of DIF items to $p = 10$, and the proportion of individuals in the outlier group to be 0.1. Finally, the class threshold which is used to identify which group the individuals are assigned for is set to 0.5.

To best see how the algorithm performs when trying to identify the DIF items and the groups that the individuals belong in, we set the number of iterations of the data generation and simulation to 100, and then the averages of the values of interest can be taken.

4.1.1 Data Generation

For generating the data for each simulation, the item parameters a_j and d_j are generated from a Uniform(0.5, 1.5) and Uniform(-2, 2) distribution respectively. In this example, we let the DIF items be the last p items, and for these δ_{j1} is distributed as a Uniform(0.5, 1.5), and for the non-DIF items this value is 0.

The individuals are separated into the two classes using a binomial distribution meaning that the respondents in the outlier class can occur anywhere in the N individuals, with a probability of 0.1. The latent ability θ for each individual in the reference group is generated from a Standard Normal distribution, and the latent ability for the outlier class is generated from $N(\mu, \sigma^2)$ where we have set $\mu = 0.5$ and $\sigma^2 = 2.25$.

This then allows the item response probabilities to be calculated which can then be used to generate the item responses using a Bernoulli distribution for each individual i and item j

using the response probabilities calculated.

4.2 Results

In this analysis, the thing that we are most interested in is how well the model performs, specifically when it comes to correctly identifying which items are the DIF items, and also correctly separating the individuals into the correct groups. As the data used was simulated and we know the true classifications of all the data, we are able to also compare how well this approach performs compared to if we were to perform the classifications while knowing the true values of the parameters.

4.2.1 ROC Curves

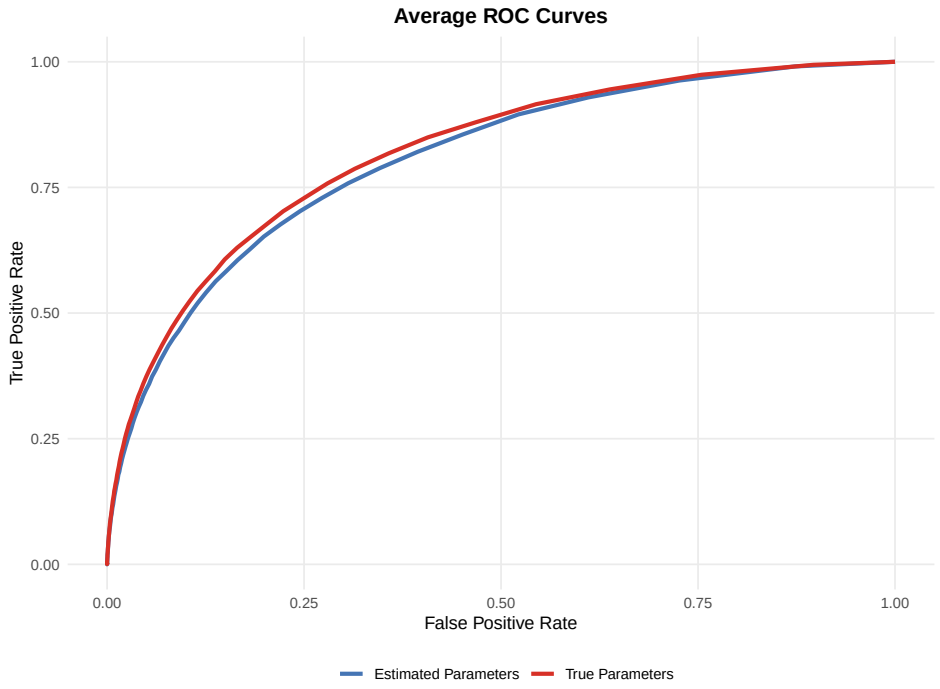

Figure 1: ROC curve for classification of individuals into groups using the True Parameters (red) and the Estimated Parameters (blue).

Figure 1 shows the ROC curves for the classification of the individuals into one of the two groups - the reference group and the outlier group. The true positive rate represents the proportion of individuals that were correctly identified to be in the outlier group and is calculated as the number of people who were predicted to be in the outlier group and are actually in the outlier group divided by the number of people who are actually belong in the

outlier group, no matter which group the model classified them in. We want this value to be high as it means that the models we are using for classification are performing well at correctly classifying the people in our group of interest, the outlier group.

On the other hand, the false positive rate represents the number of individuals who were incorrectly identified as belonging to the outlier group. This is calculated as the number of individuals who were predicted to be in the outlier group but are actually in the reference group, divided by the the number of individuals that belong in the reference group. We want this value to be small as it means that the model is not assigning individuals to the outlier group who are not part of the group.

By varying the class threshold between 0 and 1, this allows us to plot the ROC curve. A threshold close to 0 means that individuals are more likely to be assigned to the outlier group, and as the threshold value increases, the individuals become more likely to instead be assigned to the reference group with the model needing much more extreme behaviour to assign individuals to the outlier group.

The closer to the top left corner an ROC curve is, the better the classifier is performing. In Figure 1, the red curve represents the performance of the classification when the true parameters used to generate the data are used, and shows the optimal performance able to be obtained using the classification method. As the blue line representing the performance of the classification using the estimated parameters is close to the red line, this means that this method for estimating the parameters performs almost as well as if we knew the true parameters.

4.2.2 DIF Detection Performance

Another thing we can look at is how well the model is able to classify the items into non-DIF and DIF items.

Figure 2 shows the performance of the model in classifying the items as DIF or non-DIF items. On the left, the True Positive and False Positive rates for item classification over all the iterations is shown. The True Positive Rate shows how often the DIF items are correctly identified as being DIF items and takes a value of 0.959. This means that over 95% of the time, DIF items are correctly identified.

On the other hand, the False Positive rate is 0. This means that over all the iterations, a non-DIF item is never incorrectly identified as being a DIF item.

On the right hand side of Figure 2, the detection rate of each individual item being a DIF item is shown. For the first 15 items, the detection rate is 0. These items are the non-DIF items so we do not want them to be erroneously classified as being DIF items. The last 10 items are the DIF items and, as shown by the True positive Rate for the overall DIF item detection performance, these items are on average correctly identified as being DIF items

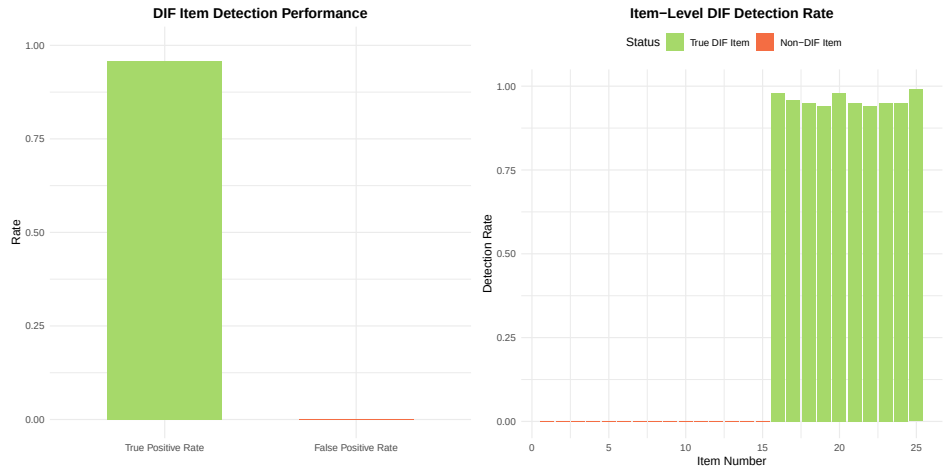

Figure 2: The True Positive and False Positive Rates for detecting DIF items (left), and the frequency of DIF item classification for each item (right).

95.9% of the time. Averaged over 1000 iterations, the detection rate for an individual item is between 94% and 99%.

5 Real Life Example

As we have seen that this method performs well on simulated data, we can now look at how it performs on real data. The most common area for using IRT is in the education setting for the evaluation of educational tests as this allows for the detection of anomalous performances for example due to cheating, or the comparison of performance of individuals from different demographic groups.

However, these techniques can be used in many other scenarios in which there are expected to be DIF items and different group memberships which affect the distributions of the individuals' latent traits. The main criteria is that the data being used to perform the analysis and classification is binary data.

One example of this is in political data, for example the voting patterns of different senators in the US Senate. We can use this data to see if the voting patterns of the different senators can be used to classify them into the two main parties - Republicans and Democrats.

5.1 The Data

The data includes information about 14 bills voted on by the US Senate in the 116th Congress [12]. For each of the bills and senators, a 1 is recorded if the senator votes in favour of a bill ("Yea") and a 0 if the senator votes against ("Nay"). Any abstentions are treated as missing

data. The dataset provides information of the 101 senators who voted on at least one of these 14 bills.

Other information about each of the senators is also provided. This includes their name and the year of their birth as well as what state and party they are representing.

To make it easier to perform the analysis on this data, we will remove any of the senators who did not vote in all 14 of the bills which leaves us with 76 individuals. We want to use this data to see if we can use the voting patterns of senators to work out which party they each represent. Most of the senators either belong to the Republicans or the Democrats, however, there is one senator left in this data set who is an Independent. Because of this we will also remove them from the dataset leaving us with information about 75 senators who voted on 14 bills.

As we know the membership of each of the senators we can use this to compare the truth to the groupings that are predicted using the estimated parameters. Out of the 75 senators of interest, we know that 35 are Democrats and the other 40 are Republicans.

Index	Bill	Total Votes	Democrat Votes	Republican Votes
1	16	59	20	39
2	22	70	35	35
3	129	68	35	33
4	182	33	33	0
5	185	68	30	38
6	188	70	33	37
7	262	56	33	23
8	311	65	35	30
9	442	71	31	40
10	504	69	35	34
11	520	64	26	38
12	549	57	35	22
13	568	67	31	36
14	625	68	35	33

Table 1: The total number of votes, and the number of votes from each party for each of the 14 bills

Table 1 shows how many of the senators voted for each of the bills out of the 75 senators who voted on all bills. Whilst for some of these bills, for example bill 188, a similar percentage of the Democrats (94.3%) and the Republicans (92.5%) voted for the bill, for many others one party favours it more. For example, with bill 182, all of the Republicans voted against the bill, and only 2 of the Democrats who voted in all 14 of the bills did. Bills 262 and 549 were also more favoured by the Democrats, and bill 16 was more favoured by the Republicans. These differences in voting patterns between the two parties suggests that using IRF to identify the

different group memberships may perform well.

5.2 Analysis

Similar to the simulation study, we wish to fit a 2PL model to this data to enable us to be able to predict the party membership of each of the senators. To do this we can use the *mirt* package [4] in *R*. This can be used to find the item parameters (a and d) as well as values for θ which, after initialising some small values for the DIF parameters, δ, can be used as starting values for the expectation-maximisation (EM) algorithm [3]. This is an iterative method to find local maximum likelihood estimates for the parameters.

5.2.1 Initial values for a and d

After fitting the initial model using *mirt*, we can look at the values it produces for both the a and d parameters.

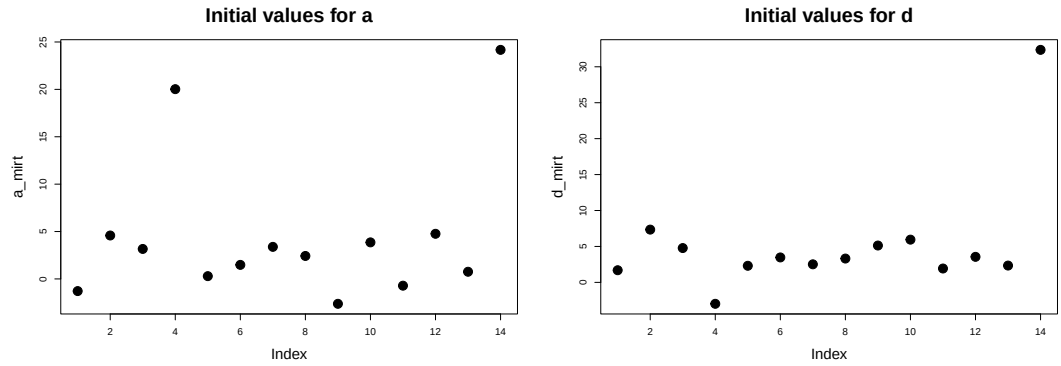

Figure 3: Initial values of a and d produced by the 2PL model.

Figure 3 shows that, whilst most of the values for a lie between -3 and 5, and most of the values for d lie between -3 and 8, there are some outliers. These outliers occur for the a value for item 4, and both the values of a and d for item 14. Having these as the starting points for these parameters for the EM algorithm may lead to some unwanted behaviour as they are much higher than the rest of the parameters, and much higher than we would expect them to be. To solve this, we can instead set a_4, a_{14} equal to 1.5 and d_{14} equal to 3 which are the mean values of a_j and d_j across the other items.

Instead of setting these to a chosen value, multiple possible values for each of the parameters could be tried to see which leads to the best performing final model. However, as we are more interested in seeing if these methods work in areas different to the testing scenario, we can instead choose to set these parameters to values close to the mean values of the parameters corresponding to the rest of the items.

5.3 Results

Once we have built the model, we need to use these results to predict which of the two parties each of the individuals belong to. To do this, the probability of each individual being in each of the two groups - the reference group and the DIF group - needs to be calculated using the estimated values of a_j, d_j, δ_{j1}, μ, σ^2 and the probability of an individual being in the DIF group. These probabilities can then be used to produce a normalised probability of each of the individuals being in the DIF group. A threshold is then needed. This defines what value the normalised probability needs to be larger than for the individual to be classified in the DIF group.

For the two group problem, a common value for this threshold is 0.5. This means that each individual is assigned to the group that they have a bigger probability of being in, according to the model parameters.

5.3.1 Varying the Threshold

As we know the party membership of each of the senators, we know the true groupings and we can use this to see how changing the threshold changes how the model performs when classifying the individuals.

This can be done for each threshold being investigated by dividing the number of individuals correctly classified by the total number of individuals. As the numbers of members of each of the parties are close to equal, this means that it is difficult to know before looking at the data which of the two political parties the model is going to treat as the reference group. This means that if when we plot the percentage of correctly classified individuals, this value is always under 0.5, we need to switch which of the groups refer to each of the parties.

Figure 4 shows how the value of the threshold changes how well the model classifies the individuals into two groups. The red line shows that by using a threshold of 0.435, 84% of the individuals were assigned to the correct political party. If we instead did not know the true groupings so chose to use a threshold of 0.5 as shown by the blue line, this would still manage to correctly classify 79% of the individuals. Even though around 20 % of the individuals are incorrectly classified using this method, this still performs better than if we assigned all individuals to the same, most popular group, which would lead to all of the individuals being classed as Republicans with a correct prediction rate of 53%.

For the rest of the analysis, even though we do know the true groupings so are able to pick the best threshold, we will set the threshold to 0.5 which is a good value for when the true groupings are not known.

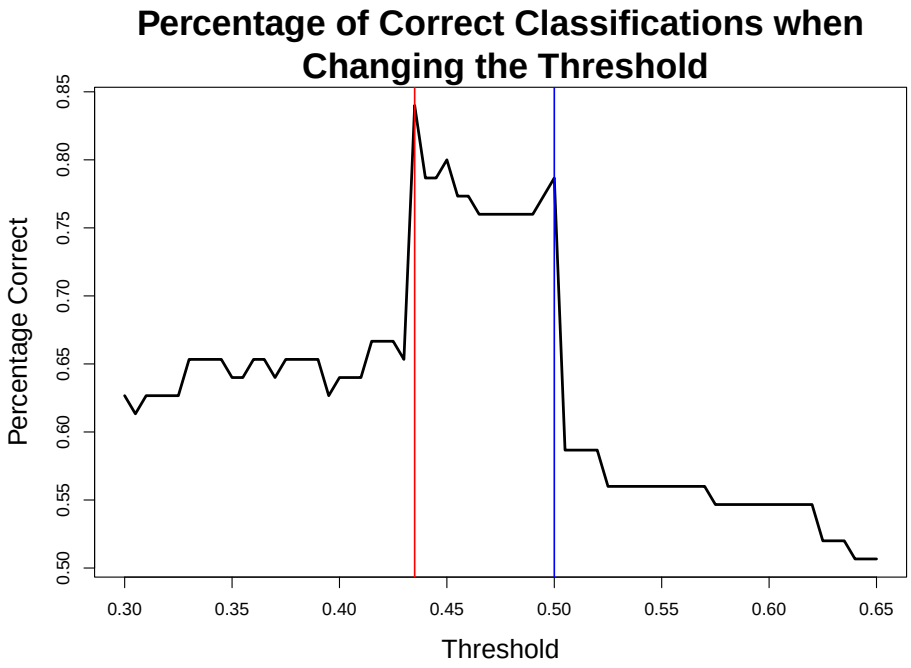

Figure 4: How the percentage of correct classifications varies as the threshold changes. The global maximum is at 0.435 (red) with 84% correct classifications. If we set the threshold to 0.5 (blue), 79% of classifications are correct.

5.3.2 Predicted Classifications

Using a threshold of 0.5 to classify each individual, we can then see how this compares to the true classification.

In our dataset, we had 35 senators who were Democrats, and the other 40 were Republicans. On the other hand, the classification from our model predicts 47 senators being Democrats, and only 28 Republican senators. This could be due to the size of our chosen threshold as the higher the threshold, the fewer the number of individuals that will be assigned to the DIF group. In this scenario the DIF group represents the Republicans. If we instead used a threshold of 0.435, this would give 35 Democrats and 40 Republicans which is the same as in the dataset.

The normalised probabilities for individuals being Republican (DIF group) from those whose true identity is Republican range between 0.303 and 0.675 with a mean of 0.499. These probabilities for Democrats range between 0.283 and 0.517 with a mean of 0.398. This means that whilst there is a difference on average between the values taken by the individuals from each of the two parties, no matter where the threshold is placed, there will be some individuals who the model gives the wrong classification to.

This can be seen by looking at the numbers of individuals from each party who are classified correctly or incorrectly by the model. Out of the 35 Democrats, 33 of them are correctly classified as Democrats (94.3%). On the other hand, out of the 40 Republicans, only 26 are correctly classified as Republicans (65%). To increase this percentage, the threshold can be lowered, but this could lead to a decrease in the number of Democrats correctly classified.

5.3.3 Estimated Parameters

Looking at the values of the parameters that the model estimates for both the individuals and the items can help to see if the model performs well and captures the true behaviour of the scenario we are looking at.

In the testing scenario, the parameters d_j represent the difficulty of each of the items, and the parameters a_j represent the sensitivity to proficiency with higher values meaning that having a better latent ability will highly affect the individuals ability to correctly answer the item [14].

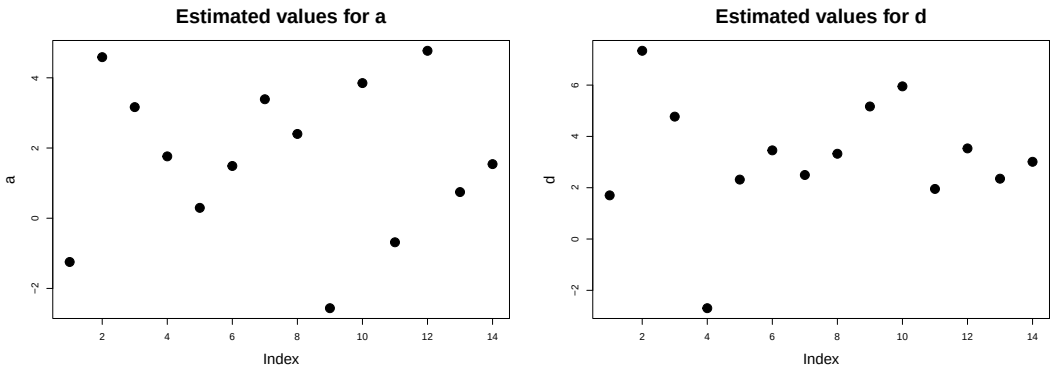

Figure 5: Estimated parameter values of a and d for each item

On the other hand, in this scenario these two parameters are more difficult to interpret, but some general patterns can be identified. Figure 5 shows how the values of a_j and d_j vary for each of the J items.

Values of a_j

In this setting, the values of a_j represent how much the latent trait of each individual affects how they vote for each of the bills. In the simulated data set, all of the values for a_j are positive as having a higher latent ability means that the individual is more likely to get each of the questions correct, no matter how difficult the questions themselves are. In this setting, the latent ability θ_i represents the ideological position of each of the senators. This means that for some of the bills, having a ideological position represented by a higher value may lead

to an individual being less likely to vote for a bill.

If we assume that there will be a difference in ideological position between the two parties, this means that when a_j has a large absolute value, members of one party are a lot more likely to vote for (or against) the bill compared to members of the other party. The lowest value of a_j occurs at $j = 9$ which represents bill 442. From table 1 it can be seen that 100% of the Republican senators in this dataset voted in favour of this bill, but only 88.6% of Democrat senators did. This shows that the ideological position has a strong affect on how the senators are likely to vote for this bill. Bills 16 and 520 also have negative values for a_j which agrees with the voting information that Republicans are more likely to vote for these bills. On the other hand, bills 129 ($j = 2$) and 549 ($j = 12$) have high values which reflect how the Democrats are more likely to vote for these bills than the Republicans.

One data point that is not behaving as what may be expected is $j = 4$ which represents bill 182. Table 1 shows that 33 of the 35 Democrat senators voted for this bill, but all of the 40 Republican senators voted against. This should lead to a high value for a_4, however this lower value could be caused by the starting value for this parameter being manually assigned before performing the EM algorithm leading to a lower value than better performing models may estimate.

A value of a_j close to 0 should mean that there is not much affect on the latent ability, or ideological position on how a senator votes which would be expected for example in bill 188, however, $a_6 = 1.49$ which is larger than expected and might mean that there is some other underlying behaviour that is not being captured well based on other characteristics of the senators.

Values of d_j

Most of the parameters d_j take a value greater than 0. In the testing setting, d_j represents the difficulty of each question with the easier questions which are more likely to be answered correctly having higher values for the parameter d_j. In the voting setting, a high d_j represents that item j is more likely to be voted for by any of the individuals, compared to an item with a lower value.

The one d_j that takes a negative value is d_4 and this represents bill 182. Table 1 shows that it was only voted for by 33 out of the 75 senators which only represents 44% of the votes. As it is less likely to be voted for than the rest of the bills, this is equivalent to a test question being difficult and therefore unlikely to be answered correctly, leading to a high difficulty and a low d_j.

The higher values of d_j, for example d_2 represent the bills that a lot of the senators are likely to vote for. For $j = 2$, 70 out of the 75 senators vote for this bill, which is equivalent to the item having a easy difficulty and therefore leads to large value for d_j.

6 Conclusions and Further Work

It has been shown that IRT models are popular and can perform well in the educational setting, especially with respect to identifying students who have cheated, as well as other anomalous results. The 2PL model used in the simulated data example performs well and manages to correctly identify each of the DIF items over 90% of the time while providing no false positives. Whilst this model performs well and is easily interpretable, there are other models that are used for IRT, for example Rasch models [19] and Probit models [11]. These models may perform better in this scenario depending on the values of the true parameters and the underlying behaviour. This includes the number of different groups that the individuals belong to, as well as the percentage of individuals belonging to the outlier group and the percentage of DIF items. Smaller changes in behaviour between the groups may also be better identified using other models.

For the political scenario, using a 2PL model looks like it could have a high success rate in correctly identifying which parties individuals belong to. Whilst the highest success rate for correct classifications achieved was 84% it is possible that this could be improved if there was more data available. Having voting data for more bills would make it easier to notice the trends in voting patterns, and how the latent variable, θ_i and the group membership affects the voting behaviour of individuals and this could lead to a better classification.

It has been seen that the behaviour of the estimated a_j and d_j parameters mostly matches what is expected based on the voting patterns of the individuals based on their true party membership. However, as there are other variables that may affect the way that the senators may vote, this leads to some individuals from both parties being incorrectly classified. Changing the model to having more than two groups that the individuals can be assigned to can allow the affects of other variables such as age or the State that they represent to be included.

The estimated δ_j values could also be analysed more. A non-zero delta suggests that even after accounting for the ideological position of a senator, one class is more or less likely to support a bill and this leads to the item showing DIF. This is more difficult to understand than in the testing setting as the latent abilities do not behave in the same way as they do in the testing setting. A higher θ_i does not automatically mean that the individual is more likely to vote for a bill.

Finally, the success of this method for modelling different group membership could be tested in more scenarios as the performance it has shown in the political setting suggests that there are other scenarios with binary data and people belonging to different groups that could also use this method, for example in marketing and psychology.

References

[1] Joseph Bafumi, Andrew Gelman, David K Park, and Noah Kaplan. Practical Issues in Implementing and Understanding Bayesian Ideal Point Estimation. *Political Analysis*, 13(2):171–187, 2005.

[2] Michael Birdsall. Implementing Computer Adaptive Testing to Improve Achievement Opportunities. *Office of Qualifications and Examinations Regulation Report*, 2011.

[3] R Darrell Bock and Murray Aitkin. Marginal Maximum Likelihood Estimation of Item Parameters: Application of an EM Algorithm. *Psychometrika*, 46(4):443–459, 1981.

[4] R. Philip Chalmers. mirt: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software*, 48(6):1–29, 2012.

[5] Yunxiao Chen, Xiaoou Li, Jingchen Liu, and Zhiliang Ying. Item Response Theory – A Statistical Framework for Educational and Psychological Measurement, 2021.

[6] Martijn G De Jong, Jan-Benedict EM Steenkamp, Jean-Paul Fox, and Hans Baumgartner. Using Item Response Theory to Measure Extreme Response Style in Marketing Research: A Global Investigation. *Journal of marketing research*, 45(1):104–115, 2008.

[7] Paul W Holland. On the Sampling Theory Roundations of Item Response Theory Models. *Psychometrika*, 55(4):577–601, 1990.

[8] Paul W Holland and Howard Wainer. *Differential Item Functioning*. Routledge, 2012.

[9] Richard N. Jones. Identification of Measurement Differences Between English and Spanish Language Versions of the Mini-Mental State Examination: Detecting Differential Item Functioning Using MIMIC Modeling. *Medical Care*, 44(11):S124–S133, 2006.

[10] Karl G Jöreskog and Arthur S Goldberger. Estimation of a Model with Multiple Indicators and Multiple Causes of a Single Latent Variable. *Journal of the American statistical Association*, 70(351a):631–639, 1975.

[11] Derrick N Lawley. XXIII.— On Problems Connected with Item Selection and Test Construction. *Proceedings of the Royal Society of Edinburgh Section A: Mathematics*, 61(3):273–287, 1943.

[12] Jeffrey B. Lewis, Keith Poole, Howard Rosenthal, Adam Boche, Aaron Rudkin, and Luke Sonnet. Voteview: Congressional Roll-Call Votes Database., 2021.

[13] Frederic M Lord. An Analysis of the Verbal Scholastic Aptitude Test Using Birnbaum's Three-Parameter Logistic Model. *Educational and Psychological Measurement*, 28(4):989–1020, 1968.

[14] Frederic M. Lord and Melvin R. Novick. Statistical Theories of Mental Test Scores, 1968.

[15] David Magis, Francis Tuerlinckx, and Paul De Boeck. Detection of Differential Item Functioning Using the Lasso Approach. *Journal of Educational and Behavioral Statistics*, 40(2):111–135, 2015.

[16] Charles E McCulloch and Shayle R Searle. *Generalized, Linear, and Mixed Models*. John Wiley & Sons, 2004.

[17] Robert J Mislevy. Estimating Latent Distributions. *Psychometrika*, 49(3):359–381, 1984.

[18] Bengt Muthén. A Method for Studying the Homogeneity of Test Items with Respect to Other Relevant Variables. *Journal of educational statistics*, 10(2):121–132, 1985.

[19] Georg Rasch. Studies in Mathematical Psychology: I. Probabilistic Models for Some Intelligence and Attainment Tests. 1960.

[20] Richard A Redner and Homer F Walker. Mixture Densities, Maximum Likelihood and the EM Algorithm. *SIAM review*, 26(2):195–239, 1984.

[21] Gideon Schwarz. Estimating the Dimension of a Model. *The annals of statistics*, pages 461–464, 1978.

[22] Robert Tibshirani. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288, 1996.

[23] Gabriel Wallin, Yunxiao Chen, and Irini Moustaki. DIF Analysis with Unknown Groups and Anchor Items. *Psychometrika*, 89(1):267–295, 2024.

[24] Peng Zhao and Bin Yu. On Model Selection Consistency of Lasso. *Journal of Machine learning research*, 7(Nov):2541–2563, 2006.