Practical Issues in Implementing and Understanding Bayesian Ideal Point Estimation

Joseph Bafumi* Andrew Gelman† David K. Park‡
Noah Kaplan§

June 10, 2004

*Department of Political Science, Columbia University, New York, jb878@columbia.edu
†Department of Statistics and Department of Political Science, Columbia University, New York, gelman@stat.columbia.edu
‡Department of Political Science, Washington University, St. Louis, dpark@artsci.wustl.edu
§Department of Political Science, University of Houston, nkaplan@uh.edu

Abstract

In recent years, logistic regression (Rasch) models have been used in political science for estimating ideal points of legislators and Supreme Court justices. These models present estimation and identifiability challenges, such as improper variance estimates, scale and translation invariance, reflection invariance, and issues with outliers. We resolve these issues using Bayesian hierarchical modeling, linear transformations, informative regression predictors, and explicit modeling for outliers. In addition, we explore new ways to usefully display inferences and check model fit.

1 Introduction

Background

Estimates of legislators' and justices' revealed voting preferences have become an important resource for scholars of legislatures and courts. The most influential method for ideal point estimation[1] in political science was developed by Poole and Rosenthal (1997). Their procedure for scoring American legislators, named NOMINATE (nominal three step estimation), has revolutionized congressional research in the American politics literature.[2]

In the last few years, political scientists have begun using an alternative approach for ideal point estimation, borrowing from the extensive psychometrics literature on logistic regressions (Jackman 2000; Clinton, Jackman and Rivers 2003). They perform these analyses using Bayesian techniques to aid in identification and recast parameter estimation into straightforward missing data problems (as has been shown in the political science context by Jackman (2000)).[3] Although this model offers a number of advantages relative to NOMINATE, it also poses a number of substantive and statistical issues. In this paper we seek to clarify these issues and suggest approaches to addressing the problems they pose.

The basic model with ability and difficulty parameters

We begin with the model as it has been understood and developed for education research (Rasch 1980). A standard model for success or failure in testing situations is the logistic item-response model, also called the Rasch model. Suppose J persons are given a test with K items, with $y_{jk} = 1$ if the response is correct. Then the logistic model can be written as,

$$\Pr(y_{jk} = 1) = \text{logit}^{-1}(\alpha_j - \beta_k), \tag{1}$$

with parameters:

- α_j: the *ability* of person j
- β_k: the *difficulty* of item k.

In general, not every person needs to receive every item, so it is convenient to index the individual responses as $i = 1, \ldots, n$, with each response i associated with a person

[1] An individual's "ideal point" refers to his or her preferences or capacities within a spatial framework. The simplest and most common spatial framework is characterized by a single dimension. Within a political context, this dimension is often conceived of as an ideological continuum, a line whose left end is understood to reflect an extremely liberal position and whose right end corresponds to extreme conservatism. In this one-dimensional spatial model, any person's ideological disposition/preference can be depicted by a point on this line—the person's ideal point.

[2] For a list of many of these works see Clinton, Jackman and Rivers (2003).

[3] The use of Bayesian inference with the Rasch model to estimate ideal points has become increasingly popular in the political science literature. Martin and Quinn (2001, 2002*b*,*a*); Clinton, Jackman and Rivers (2003); Bafumi et al. (2002) have used Bayesian ideal point estimation to scale Supreme Court justices; Clinton, Jackman and Rivers (2003) estimate ideal points in the U.S. House; and Jackman (2001); Clinton (2001); Park (2001) employ estimated ideal points for senators.

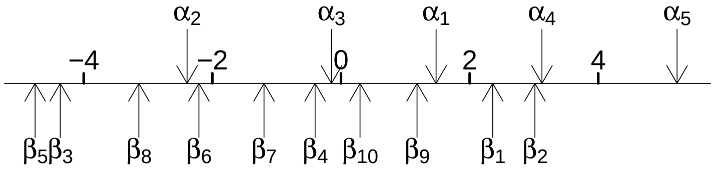

Figure 1: *Illustration of the logistic item-response (Rasch) model,* $\Pr(y_i = 1) = logit^{-1}(\alpha_{j(i)} - \beta_{k(i)})$*, for an example with five persons* j *(with abilities* α_j*) and ten items* k *(with difficulties* β_k*). If your ability* α *is greater than the difficulty* β *of an item, then you have a better-than-even chance of getting that item correct. This graph also illustrates the nonidentifiability in the model: the probabilities depend only on the* relative *positions of the ability and difficulty parameters; thus, a constant could be added to all the* α_j*'s and all the* β_k*'s, and the model would be unchanged. One way to resolve this nonidentifiability is to constrain the* α_j*'s to have mean 0. Another solution is to give the* α_j*'s a distribution with mean fixed at 0.*

$j(i)$ and item $k(i)$. Thus model (1) becomes,

$$\Pr(y_i = 1) = \text{logit}^{-1}(\alpha_{j(i)} - \beta_{k(i)}), \tag{2}$$

Figure 1 illustrates the model as it might be estimated for five persons with abilities α_j and ten items with difficulties β_k. In this particular example, questions 5, 3, and 8 are easy (relative to the abilities of the persons in the study), and all persons except person 2 are expected to answer more than half the items correctly. More precise probabilities can be calculated using the logistic distribution: for example, α_2 is 2.4 higher than β_5, so the probability that person 2 correctly answers item 5 is $\text{logit}^{-1}(2.4) = 0.92$, or 92%.

Interpretation as an ideal point model

The Rasch model can be directly used for ideal point estimation in political science research. Here, subscript j denotes a legislator or justice and subscript k denotes a bill or case. The ability parameter, α_j, measures the liberalness or conservativeness of a legislator and the difficulty parameter, β_k, indicates the ideal point of a legislator who is indifferent on that bill or case. Thus, in Figure 1, α_4 and α_5 could represent highly conservative justices (e.g., Scalia and Thomas) and β_2 would represent a case for which justice 4 would be nearly indifferent. Justice 4 would be more likely to vote in the conservative direction as a case's difficulty parameter moves to the left.

2 Identifiability problems

2.1 Additive aliasing

This model is not identified, whether written as (1) or as (2), because a constant can be added to all the abilities α_j and all the difficulties β_k, and the predictions of the model will not change. The probabilities depend only on the *relative* positions of the ability and difficulty parameters. For example, in Figure 1, the scale could go from -104 to -96 rather than -4 to 4, and the model would be unchanged—a difference of 1 on the original scale is still a difference of 1 on the shifted scale.

From the standpoint of classical logistic regression, this nonidentifiability is a simple case of collinearity and can be resolved by constraining the estimated parameters in some way: for example, setting $\alpha_1 = 0$ (that is, using person 1 as a "baseline"), setting $\beta_1 = 0$ (so that a particular item is the comparison point), constraining the α_j's to sum to 0, or constraining the β_j's to sum to 0. A multilevel model allows for other means of solving the additive aliasing problem, as we discuss next.

Multilevel model

Item-response and ideal-point models are inherently applied to multilevel structures, with data nested within persons and test items, or judges and decisions, or legislators votes. A commonly-used multilevel model for (2) assigns normal distributions to the ability and difficulty parameters:[4]

$$\alpha_j \sim \mathrm{N}(\mu_\alpha, \sigma_\alpha^2), \text{ for } j = 1, \ldots, J$$
$$\beta_k \sim \mathrm{N}(\mu_\beta, \sigma_\beta^2), \text{ for } k = 1, \ldots, K.$$

The model is multilevel because the priors for these parameters are assigned hyperpriors and estimated conditional on the data. This is also referred to as a partial pooling or hierarchical approach (Gelman et al. 2003). The model is nonidentified for the reasons discussed above: this time, it is μ_α and μ_β that are not identified, because a constant can be added to each without changing the predictions. The simplest way to identify the multilevel model is set μ_α to 0 (or to set μ_β to 0, but not both due to collinearity).

Defining the model using redundant parameters

Another way to identify the model is by allowing the parameters α and β to float and then defining new quantities that are well-identified. The new quantities can be defined, for example, by rescaling based on the mean of the α_j's:

$$\alpha_j^{adj} = \alpha_j - \bar{\alpha}, \text{ for } j = 1, \ldots, J$$
$$\beta_k^{adj} = \beta_k - \bar{\alpha}, \text{ for } k = 1, \ldots, K.$$

The new ability parameters α_j^{adj} and difficulty parameters β_k^{adj} are well defined, and they work in place of α and β in the original model:

$$\Pr(y_i = 1) = \mathrm{logit}^{-1}(\alpha_{j(i)}^{adj} - \beta_{k(i)}^{adj}).$$

This holds because we subtracted the same constant from both the α's and β's. It would *not* work to subtract $\bar{\alpha}$ from the α_j's and $\bar{\beta}$ from the β_k's.

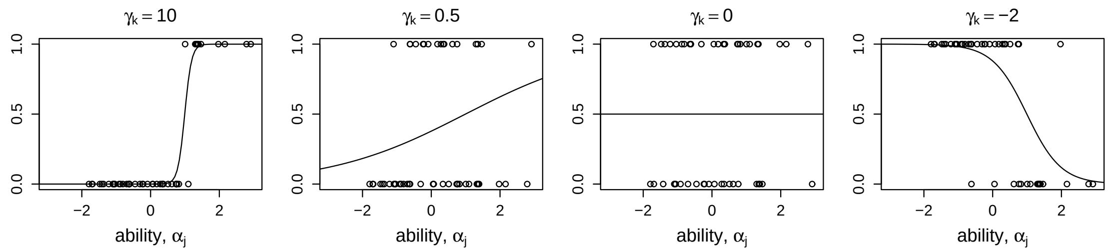

Figure 2: *Curves and simulated data from the logistic item-response (Rasch) model for items k with "difficulty" parameter $\beta_k = 1$ and high, low, zero, and negative "discrimination" parameters γ_k.*

2.2 Multiplicative aliasing

The basic model with a discrimination parameter

The item-response model can be generalized by allowing the slope of the logistic regression to vary by item:

$$\Pr(y_i = 1) = \text{logit}^{-1}(\gamma_{k(i)}(\alpha_{j(i)} - \beta_{k(i)})). \tag{3}$$

In this new model, γ_k is called the *discrimination* of item k: if $\gamma_k = 0$, then the item does not "discriminate" at all ($\Pr(y_i = 1) = 0.5$ for any person), whereas high values of γ_k correspond to a strong relation between ability and the probability of voting as expected or getting a correct response, as the case may be. Figure 2 illustrates. Negative values of γ_k correspond to items where low-ability persons do better. Such items typically represent mistakes in the construction of the test since test designers generally try to create questions with a high positive discrimination value. In ideal point research, the discrimination parameter indicates how well a case or bill discriminates between conservative and liberal justices/legislators. The addition of the discrimination parameter brings about a new invariance problem—scaling invariance or multiplicative aliasing.

Resolving the new source of aliasing

Model (3) has a new source of indeterminacy: a multiplicative aliasing in all three parameters that arises when multiplying the γ's by a constant and dividing the α's and β's by that same constant. We can resolve this indeterminacy by constraining the α_j's to have mean 0 and standard deviation 1 or, in a multilevel context, by giving the α_j's a fixed population distribution (e.g., $\text{N}(0, 1)$).

As an alternative, we propose establishing hyperpriors for all parameters of interest and transforming those parameters via normalization after estimation is complete. For example, we can calculate the mean and standard deviation of α and generate the following normalized parameter:

$$\alpha^{adj}_{j(i)} = (\alpha_{j(i)} - \bar{\alpha})/s_\alpha,$$

[4] This is a model in which there is no distinguishing information on the persons and items (beyond that in the data matrix itself). If additional data are available on the persons and items, this information can be included as predictors in group-level regressions. We present such a model in Section 2.2 and illustrate in a study of Supreme Court justices, using political party as a justice-level predictor.

where the *adj* superscript denotes normalization, and $\bar{\alpha}$ and s_α represent the the mean and standard deviation of of the α_j's.

We also wish to normalize the β's and γ's while retaining a common scale for all parameters. Thus, we transform these parameters using the mean and standard deviation of α as well:

$$\begin{aligned}\beta_{k(i)}^{adj} &= (\beta_{k(i)} - \bar{\alpha})/s_\alpha \\ \gamma_{k(i)}^{adj} &= \gamma_{k(i)} s_\alpha.\end{aligned}$$

This rescaling resolves the multiplicative aliasing problem as well as the additive aliasing problem discussed above. The likelihood is preserved (since $\gamma_k^{adj}(\alpha_j^{adj} - \beta_k^{adj}) = \gamma_k(\alpha_j - \beta_k)$) while allowing computation to proceed more efficiently (this follows the parameter-expansion idea of Liu, Rubin and Wu (1998); also see Gelman et al. (2003). Highly correlated parameters slow down MCMC sampling (Gilks, Richardson and Spiegelhalter 1996), making convergence elusive for very many iterations. The transformations above fix this problem by reducing posterior correlation in posterior densities. For example, Figure 3 plots the potential scale reduction factor $\widehat{R}$ (Gelman and Rubin 1992; Gelman et al. 2003) for unadjusted and adjusted α's representing justices in one natural court.[5] A value of 1 indicates approximate convergence of multiple chains. After 15,000 iterations, the normalized ideal points show much better convergence than the non-normalized scores.

Reflection Invariance

Even after successfully dealing with additive and multiplicative aliasing, one indeterminacy issue remains in model (4): a reflection invariance associated with multiplying all the γ_k's, α_j's, and β_k's by -1. If no additional constraints are assigned to this model, this aliasing will cause a bimodal likelihood and posterior distribution. It is desirable to select just one of these modes for our inferences. (Among other problems, if we include both modes, then each parameter will have two maximum likelihood estimates and a posterior mean of 0.) In a political context, we must identify one direction as "liberal" and the other as "conservative" (or however the principal ideological dimension is understood; see Poole and Rosenthal (1997)).

Before presenting our method for resolving the reflection invariance problem, we briefly discuss two simpler approaches. With appropriately structured data, one can constrain the discrimination parameter (γ's) to all have positive signs. This makes sense when the outcomes have been precoded so that, for example, positive y_i's correspond to conservative votes. However, we do not use this approach because it relies too strongly on the precoding, which, even if it is generally reasonable, is not perfect (as we shall see in our Supreme Court example). We would prefer to estimate the ideological direction of each vote from the data and then compare to the precoding to check that the model makes sense (and to explore any differences found between the estimates and the precoding).

A second approach to resolving the aliasing is to choose one of the α_j's, β_k's, or γ_k's and restrict its sign (Jackman 2001). For example, we could constrain α_j to be negative for the extremely liberal William Douglas, or constrain α_j to be positive for the extremely

[5]Results for the entire dataset of 29 justices are explored in Section 4 below.

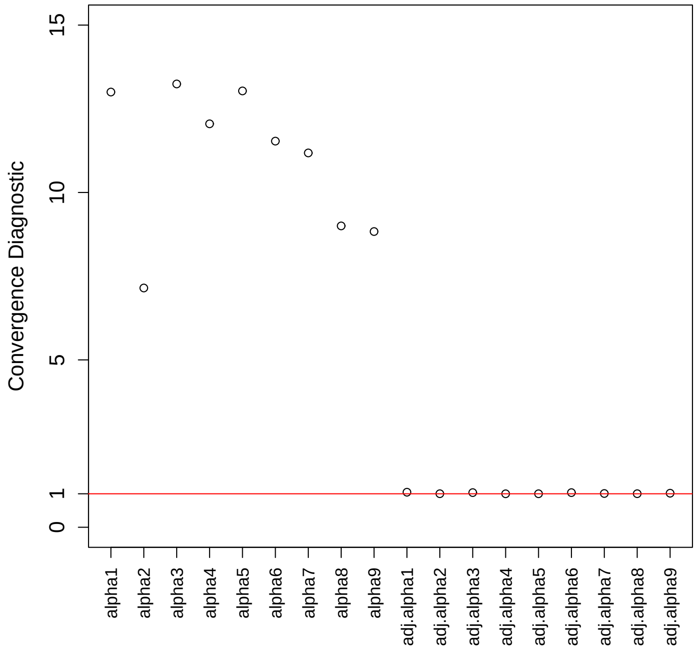

Figure 3: *Convergence of normalized versus non-normalized ideal point estimates after 15,000 iterations. A diagnostic value of 1 indicates mixing of MCMC sequences and thus apparent convergence. The normalized estimates show much better convergence properties.*

conservative Antonin Scalia. Or, we could constrain Douglas's α_j to be less than Scalia's α_j. Only a single constraint is necessary to resolve the two modes problem; however, it should be a clear-cut division. We do not like this strategy as a general approach to resolving indeterminacy. It relies on the existence of a parameter, or a comparison of parameters, that we know a priori will be clearly distinguishable from the data. The existence of such a parameter is not always clear. We cannot simply constrain the sign of α_j for an arbitrary justice, because if we were to pick a centrist such as Sandra Day O'Connor, this could split the likelihood surface across both modes, rather than cleanly selecting a single mode.

Instead, we advocate resolving the reflection invariance problem with a group-level predictor in a level 2 regression equation. Generally, in the Rasch model, the "groups" are the persons and items:

$$\begin{aligned} \alpha_j &\sim \mathrm{N}((X_\alpha \delta_\alpha)_j, \sigma^2_\alpha), \text{ for } j = 1, \ldots, J \\ \beta_k &\sim \mathrm{N}((X_\beta \delta_\beta)_j, \sigma^2_\beta), \text{ for } k = 1, \ldots, K. \\ \gamma_k &\sim \mathrm{N}((X_\gamma \delta_\gamma)_j, \sigma^2_\gamma), \text{ for } k = 1, \ldots, K. \end{aligned}$$

In a model predicting Supreme Court justices' ideal points, the person-level predictors X_α could include age, sex, time in office, party of appointing president, and so forth, and the item-level predictors X_β and X_δ could include characteristics such as indicators for the type of case (for example, civil liberties, federalism, and so forth).

We can use person-level predictors to solve the reflectional invariance problem. For example, we can include the party of the nominating President for each justice as a

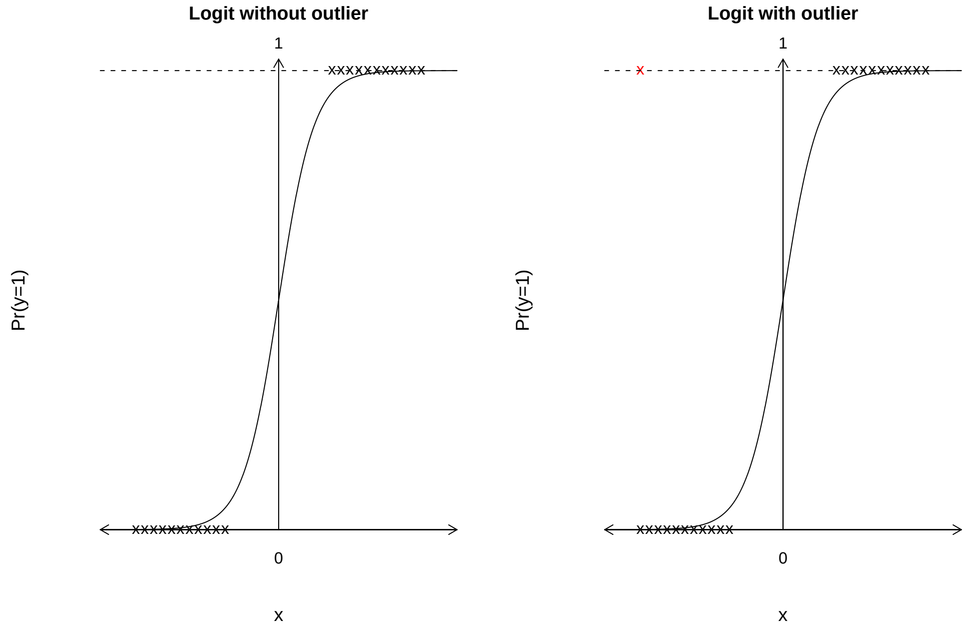

Figure 4: *Plot of hypothetical probability distribution with and without outlier.*

predictor in model (4). The predictor is included in the model at the justice level:

$$\alpha_j \sim \mathrm{Nor}(\delta_0 + \delta_1 x_j),$$

where $x_j = 1$ if the justice was nominated by a Republican and -1 if by a Democrat. Constraining the regression coefficient δ_1 to be positive identifies the model.

This forces ideal points for liberal and conservative justices to be on opposite sides of a scale and in the preferred direction. We prefer this alternative for a number of reasons. First, it utilizes important prior information that may otherwise be ignored. This is likely to improve the fit of the model. Second, it does not require any knowledge of extreme justices or reference cases as with the anchoring strategies discussed above. Third, it is applicable in single or multidimensional contexts (where additional predictors are necessary).

3 Outliers: robust logistic regression

Pregibon (1982) and Liu (2004) have shown that the logit and probit models are not robust to outliers. For binary data, "outliers" correspond not to extreme values of y but rather to values of y that are highly unexpected given the linear predictor $X\beta$ (for example, if $X\beta = 10$ then $\mathrm{logit}^{-1}(10) = 0.99995$, so the observation $y = 0$ would be an "outlier" in this sense). We propose a modified logit model, similar to that of Liu (2004), that allows for outliers (see Figure 4).

We have observed data with n independent observations $(x_i, y_i),\ i = 1, \ldots, n$ with a multidimensional covariate vector x_i and binary response y_i. The logistic regression model is specified by,

$$\Pr(y_i = 1) = \mathrm{logit}^{-1}(\gamma_{k(i)}(\alpha_{j(i)} - \beta_{k(i)})).$$

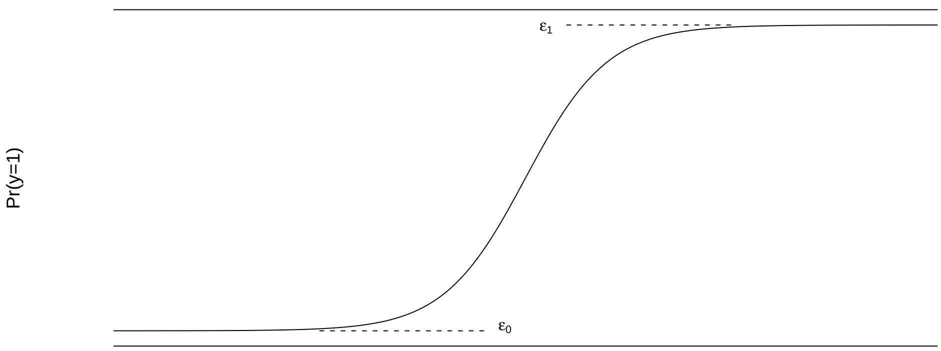

Figure 5: *Logistic distribution which allows for outliers. Outliers are defined as values of the outcome variable (y) that are highly unexpected given the linear prediction. The solid horizontal lines are at y=0 and y=1; the dotted lines are at* ε_0 *and 1 -* ε_1

To have a robust logit model, we simply allow the logit model to contain a level of error, ε_0 and ε_1, as follows (see Figure 5):

$$\Pr(y_i = 1) = \varepsilon_0 + (1 - \varepsilon_0 - \varepsilon_1)\text{logit}^{-1}(\gamma_{k(i)}(\alpha_{j(i)} - \beta_{k(i)})).$$

Within the Bayesian context, we allow the error rates, ε_0 and ε_1, to be estimated from data by assigning them independent Uniform(0,0.1) prior distributions. (If the error rates were much higher than 10%, we would not want to be fitting even an approximate logit model.)

4 Ideal point modeling for U.S. Supreme Court justices

We illustrate with an ideal-point model fit to the voting records of U.S. Supreme Court justices, using all the Court's decisions from 1954 to 2000.[6] Each vote i is associated with a justice $j(i)$ and a case $k(i)$, with an outcome y_i that equals 1 if a justice voted in the conservative direction on a case and 0 if he or she voted in the liberal direction.[7] As discussed in Section 2 of this paper, the data are modeled with a logistic regression,

[6] The data were compiled by Harold J. Spaeth and can be downloaded from `www.polisci.msu.edu/pljp`.

[7] The codings of "liberal" and "conservative" can sometimes be in error. As we shall discuss, the model with its discrimination parameter allows us to handle and even identify possible miscodings of the directions of the votes (Jackman 2001).

with the probability of voting conservatively depending on the "ideal point" α_j for each justice, the "position" β_k for each case, and a "discrimination parameter" γ_k for each case, following the three-parameter logistic model (3):

$$\Pr(y_i = 1) = \text{logit}^{-1}(\gamma_{k(i)}(\alpha_{j(i)} - \beta_{k(i)})). \tag{4}$$

The difference between α_j and β_k indicates the positions of the justices and the cases—if a justice's ideal point is near a case's position, then the case could go either way, but if the ideal point is far from the position, then the justice's vote is highly predictable. The discrimination parameter γ_k captures the importance of the positioning in determining the justices' votes: if $\gamma_k = 0$, the votes on case k are purely random; and if γ_k is very large (in absolute value), then the relative positioning of justice and case wholly determine the outcome. Changing the sign of gamma changes which justices are expected to vote yes and which to vote no.

We fit the model using the Bayesian software WinBUGS (Spiegelhalter, Thomas and Best 1999) as called from R (R Development Core Team 2003; Gelman 2003). Two parallel chains reached approximate convergence (using the adjusted parameterization described in Section 2) after 15,000 iterations. The party of the appointing president is coded such that conservative justices are placed to the right of liberal justices. The posterior distribution for party has a median value of 0.2 and 95 percent confidence intervals around 0 and 0.9. While it does not offer a great deal substantively, it solves the reflectional invariance problem discussed above. Figure 6 takes a random draw from the posterior distribution and plots the difficulty parameters of the cases as well as the median of the posterior distribution for justices' ideal points, both, across court terms. The positions of the cases appear stable over time, but there is a trend toward more conservative justices.[8] This corresponds nicely with the widely accepted notion of a rightward shifting court.

Figure 7 plots a random draw from the discrimination parameters' posterior distribution. Relatively few of the cases have negative discrimination parameters (less than 5 percent). Since this is likely to be a sign of miscoding in the original data, it is a welcome result. Negative discrimination parameters may also result from switching coalitions where conservative justices vote in what would a priori seems to be the liberal direction while liberal justices vote in the conservative direction. Of the cases with negative median discrimination parameter posterior distributions, about 25% are judicial power cases, 18% are economic activity cases and 10% are federalism and interstate relations cases each. Switching coalitions or miscodes prove to be most likely with judicial power cases. The higher the value of the discrimination parameter, the more likely the outcome of the case depends on a left-right ideological construct. Of the discrimination posterior distribution medians that exceed a value of 5, the plurality, 44%, are criminal procedure cases. Not surprisingly, they seem to discriminate best between liberal and conservative justices. Another 20% of these cases are of the civil rights variety.

Such plots serve as a rough verification of model fit. They show results that make sense given what we expect, a priori, about the Supreme Court. However, models designed to generate ideal points, and item response models more generally, can and should undergo more rigorous tests of model fit. To this we turn next.

[8]As can be seen from the horizontal lines in Figure 6, our model constrains each justice to have an unchanging ideology over time.

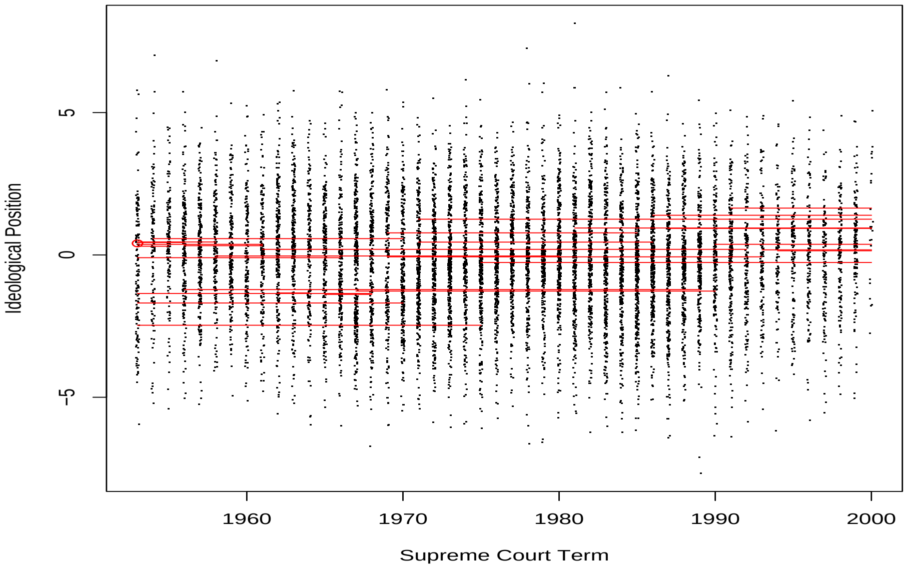

Figure 6: *Posterior median estimates of the positions* β_k *of the cases (the dots on the graph) and the ideal points* α_j *of the 29 justices (the lines) as estimated using the Supreme Court model. Points on the ideal point line reflect cases for which that justice is indifferent.*

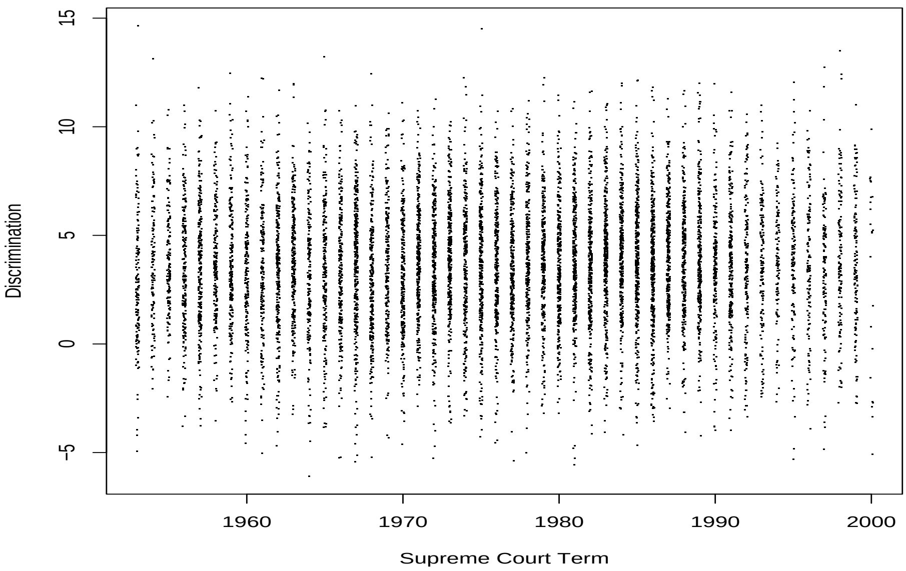

Figure 7: *Draws from standardized discrimination posterior distributions for 29 justices plotted across Supreme Court terms. Dots represent individual cases. The higher values point to cases that better discriminate between liberal and conservative justices.*

5 Assessing Model Fit

Statistical modelers typically spend little time rigorously judging model fit, even when such checks can result in discoveries that greatly improve one's model. For ideal point estimates, one can test one aspect of model fit by checking residuals. They are classically defined as,

$$\hat{r}_i = y_i - \text{logit}^{-1}(\gamma_{k(i)}(\alpha_{j(i)} - \beta_{k(i)}))$$

It is useful to consider the excess error; the error beyond what we would expect given the model's predicted values. First we need to understand how to calculate what we would expect the error to be. The expected error is simply the minimum of the model's prediction and 1 minus this prediction.

$$\text{E}(\hat{r}_i) = \min(\text{logit}^{-1}(\gamma_{k(i)}(\alpha_{j(i)} - \beta_{k(i)})), 1 - \text{logit}^{-1}(\gamma_{k(i)}(\alpha_{j(i)} - \beta_{k(i)})))$$

Thus, the excess error can be formalized as follows,

$$\text{Excess error}_i = \hat{r}_i - \text{E}(\hat{r}_i)$$

Individual-level errors are difficult to interpret usefully. However, averages of errors, which offer 0 as a baseline or expectation, convey more meaningful information. Here, we shall investigate the average excess error for each justice in our data. We plot the excess error rate per justice across the justice's ideal points. We plot the values from 5 separate draws to capture the uncertainty in the posterior distributions. These are the realized residuals (Gelman 2004). To provide a reference distribution, we generate replicated y's from our model and also plot their excess error rate per justice across the justices' ideal points.[9] The replicated y's show the excess rate we would expect if our model were true.

Figure 8 shows five random draws of the realized residuals on the top row, with corresponding draws from the reference distribution on the bottom row. The excess error rate in the reference plots is low implying that ideal points are estimated with great precision. The realized residuals show less precision. In general, the ideal points of conservative justices can be estimated more predictably than liberal justices. Particularly, the ideal points for justices 2 and 3 (Black and Douglas) have high excess error rates. Justice Black has an error rate that is over 50% higher than we would expect if our model were true. Justice Douglas has an error rate that is over one third higher than expected. Douglas's ideal point is probably hard to estimate because there is no one to anchor him to his left (Poole and Rosenthal N.d.). Meanwhile, Black has been shown to undergo significant shifts over time in his ideological ideal point even after controlling for docket effects (Bafumi et al. 2002). Where model fit shows room for improvement, one can revisit the specification of the original model.

The most noticeable pattern in the bottom row of graphs in Figure 8 is that the excess error rates for justices 17, 14, and 15 (Jackson, Fortas, and Goldberg) appear likely to have high absolute values in the replicated datasets. These potentially high errors arise

[9] We generate replicated y's by assigning a random binomial distribution to the vectors of predicted y's.

because our data provides little information on these justices,[10] hence their ideal points are estimated with less accuracy and there is more room for error in the prediction. However, as the top row of Figure 8 shows, the largest data errors are for Justices Black and Douglas, as discussed above.

One can also judge the overall fit of a discrete-data regression using calibration from pooled predictions. A calibration plot allows us to compare the fitted (predicted) versus the actual average values of y within bins. For example, one would begin by selecting the number of bins to analyze; more bins allow for more fine-grained analyses. Then, one would isolate the fitted values for y that fall in each bin. For example, we can examine the number of fitted values for 10 bins: 0–0.1, 0.1–0.2, ... Then we calculate the mean for the fitted y's that fall into each bin. Next, we calculate the mean of the corresponding actual y's. Plotted together, the means of the fitted versus actual y's should fall on the 45° line. This is shown in the first column of Figure 9 below. The actual and fitted y's show about the same vote probability in each of the 10 bins. We can also inspect a binned residual plot by substracting the mean of the fitted y's from the mean of the actual y's across each bin and plotting this new result across the fitted y's. We expect no discernible pattern and residuals close to 0. This is shown in the second column of Figure 9. Such plots offer greater confidence in the fit of the model.

6 Conclusion

The use of ideal point estimates has become common in political science research today. With their increased use, political methodologists have spent more time working to improve the quality of these scores. This paper investigates a series of practical issues that arise with the estimation of ideal points and offers solutions that are not commonly applied to date. Problems include proper variance estimates, scale and translation invariance, reflection invariance and outliers. Resolutions to these issues come in the form of Bayesian hierarchical modeling, linear transformations, informative regression predictors and explicit modeling for outliers. In addition, we explored new ways to usefully display inferences and check model fit.

The procedures investigated above apply to unidimensional models. They do not account for additional dimensions that explain votes or decisions beyond the left-right ideological construct. However, the innovations could be generalized to such multidimensional models. In fact, many could be generalized to Bayesian models of all sorts (for example, transformations that aid in interpretation or convergence and model checking). Similar issues arise in latent-class models and factor analysis (Loken 2004). Also, the substantive model explored in Section 4 (to estimate the ideal points of Supreme Court justices) can be developed much further. For example, it can be expanded to test propositions such as shifting ideal points among justices over time (Martin and Quinn 2001, 2002*b*,*a*). This we leave to future work.

As Congress and judiciary scholarship continue to grow, the demand for high quality ideal point estimates will also grow. These scores are one of several resources that scholars can use to understand the workings of government. Others include in-depth studies

[10] Fortas and Goldberg served for only a few years each, and Jackson's Court service ended shortly after the start of our dataset.

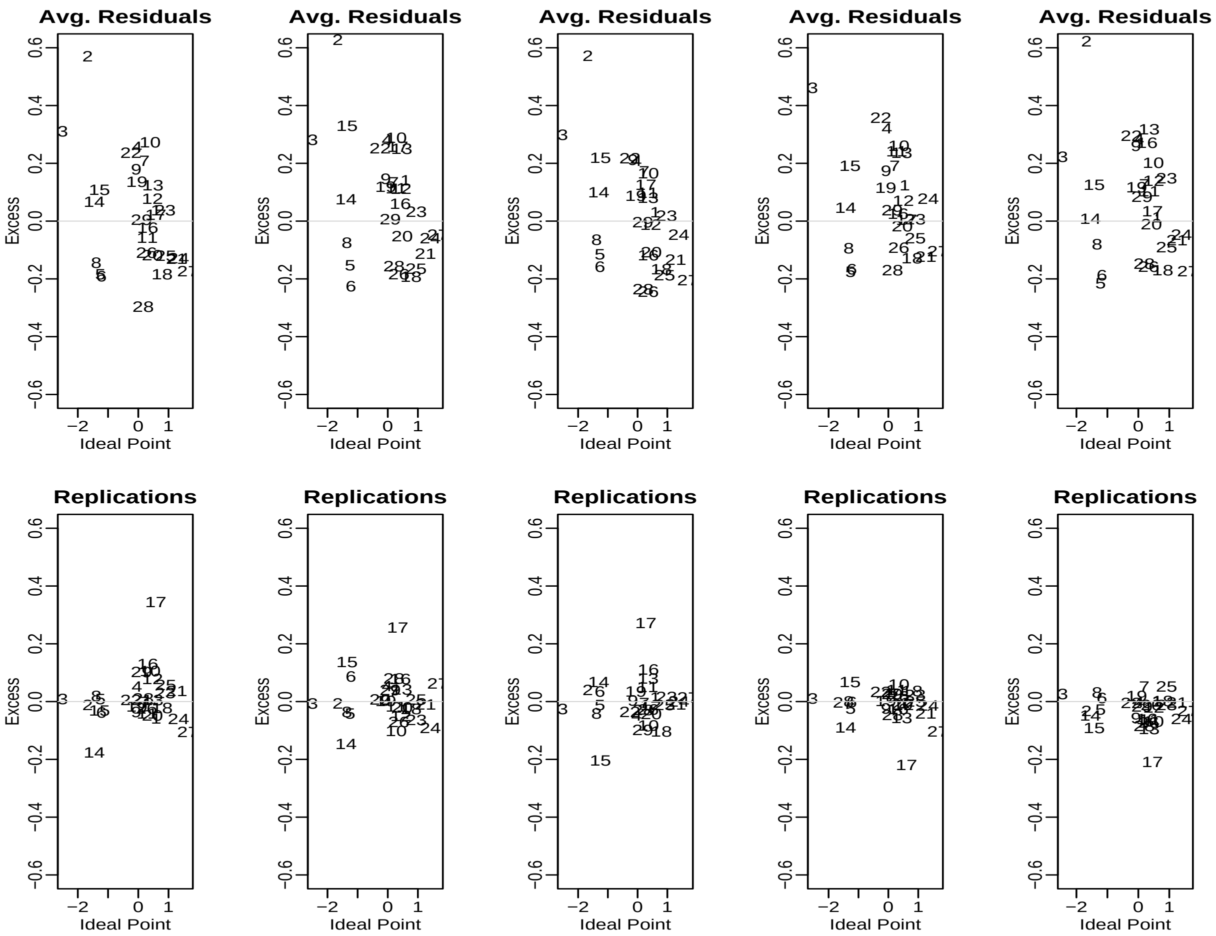

Figure 8: *Plots of excess error rate in real and replicated values of the justices' votes. Numbers label a justices as follows: 1 Harlan, 2 Black, 3 Douglas, 4 Stewart, 5 Marshall, 6 Brennan, 7 White, 8 Warren, 9 Clark, 10 Frankfurther, 11 Whittaker, 12 Burton, 13 Reed, 14 Fortas, 15 Goldberg, 16 Minton, 17 Jackson, 18 Burger, 19 Blackmun, 20 Powell, 21 Rehnquist, 22 Stevens, 23 O'Connor, 24 Scalia, 25 Kennedy, 26 Souter, 27 Thomas, 28 Ginsburg, 29 Breyer. For all plots, the average excess error rate per justice is plotted across the justices' ideal points. The replications show what we would expect if our model were true. The realized residuals show room for model improvement.*

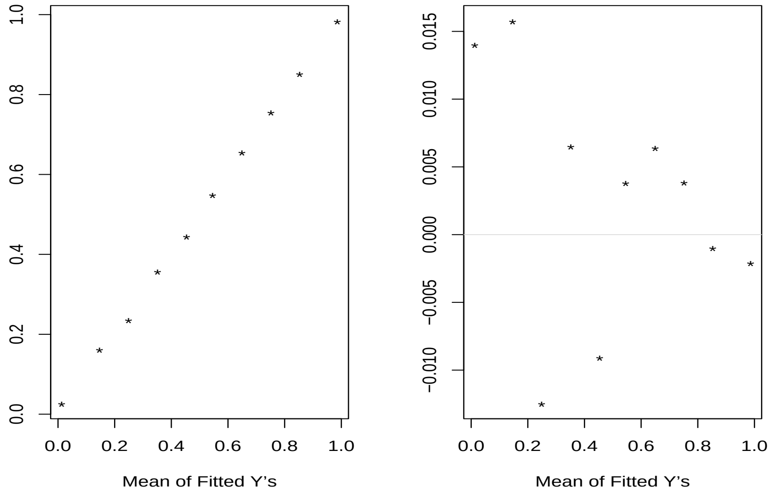

Figure 9: *Calibration and binned residual plots with 10 bins for checking model fit. In a well-fitting model, the mean of the binned actual y's and the mean of the binned fitted y's fall on the* 45° *angle, as seen above. The differences between the two measures are almost all less than 1.5%.*

(Fenno 1978), legislator interviews (Lahav N.d.) and a variety of scores that capture legislators underlying ideological ideal points without the complexity associated with NOMINATE or Bayesian estimates such as content coding, special interest group scores or simple tabulations (Bafumi et al. 2002). Growing research in each of these areas will benefit the scholarship as a whole.

References

Bafumi, Joseph, Noah Kaplan, Nolan McCarty, Keith Poole, Andrew Gelman and Charles Cameron. 2002. "Scaling the Supreme Court: A Comparison of Alternative Measures of the Justices Idealogical Preferences, 1953-2001." Presented at the 2002 Annual Meeting of the Midwest Political Science Association, Chicago, IL.

Clinton, Joshua. 2001. "Legislators and their Constituencies: Representation in the 106th Congress." Presented at the 2001 Annual Meeting of the American Political Science Association, San Francisco, CA.

Clinton, Joshua, Simon Jackman and Douglas Rivers. 2003. "The Statistical Analysis of Roll Call Data." Technical Report: Political Methodology Working Papers.

Fenno, Richard F. 1978. *Home Style: House Members in Their Districts*. Boston: Little, Brown.

Gelman, Andrew. 2003. *Bugs.R: Functions for Calling Bugs from R.*
URL: *http://www.stat.columbia.edu/~gelman/bugsR*

Gelman, Andrew. 2004. "Explanatory Data Analysis for Complex Models (with Discussion)." *Journal of Computational and Graphical Statistics* .

Gelman, Andrew and Donald B. Rubin. 1992. "Inference from Iterative Simulation using Multiple Sequences." *Statistical Science* pp. 457–511.

Gelman, Andrew, John S. Carlin, Hal S. Stern and Donald B. Rubin. 2003. *Bayesian Data Analysis*. Boca Raton: Chapman and Hall.

Gilks, Wally R., Sylvia Richardson and David J. Spiegelhalter. 1996. *Markov Chain Monte Carlo in Practice*. London: Chapman & Hall.

Jackman, Simon. 2000. "Estimation and Inference are Missing Data Problems: Unifying Social Science Statistics via Bayesian Simulation." *Political Analysis* 8:307–332.

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

Lahav, Gallya. N.d. *Immigration and Politics in the New Europe: Reinventing Borders*. London: Cambridge Press.

Liu, Chuanhai. 2004. Robit Regression: A Simple Robust Alternative to Logistic and Probit Regression. In *Applied Bayesian Modeling and Causal Inference from an Incomplete Data Perspective*, ed. Andrew Gelman and X. L. Meng. London: Wiley chapter 21.

Liu, Chuanhai, Donald B. Rubin and Yingnian N. Wu. 1998. "Parameter Expansion to Accelerate EM: The PX-EM Algorithm." *Biometrika* 85:755–770.

Loken, Eric. 2004. Multimodality in Mixture Models and Factor Models. In *Applied Bayesian Modeling and Causal Inference from an Incomplete Data Perspective*, ed. Andrew Gelman and X. L. Meng. London: Wiley chapter 19.

Martin, Andrew D. and Kevin M. Quinn. 2001. "The Dimensions of Supreme Court Decision Making: Again Revisiting *The Judicial Mind*." Presented at the 2001 meeting of the Midwest Political Science Association.

Martin, Andrew D. and Kevin M. Quinn. 2002*a*. "Assessing Preference Change on the U.S. Supreme Court." Presented at the University of Houston, March 15, 2002.

Martin, Andrew D. and Kevin M. Quinn. 2002*b*. "Dynamic Ideal Point Estimation via Markov Chain Monte Carlo for the U.S. Supreme Court, 1953-1999." *Political Analysis* 10:134–153.

Park, David. 2001. "Representation in the American States: The 100th Senate and their Electorate." Working Paper.

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

Poole, Keith T. and Howard Rosenthal. N.d. *Spatial Models of Legislative Voting*.

Pregibon, D. 1982. "Resistant Fits for Some Commonly Used Logistic Models with Medical Applications." *Biometrics* 38:485–498.

R Development Core Team. 2003. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-00-3. **URL:** *http://www.R-project.org*

Rasch, George. 1980. *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago: University of Chicago Press.

Spiegelhalter, David J., Andrew Thomas and Nickey G. Best. 1999. *WinBugs Version 1.4*. Cambridge, UK: MRC Biostatistics Unit.