Estimation and Inference Are Missing Data Problems: Unifying Social Science Statistics via Bayesian Simulation

Simon Jackman
Department of Political Science, Stanford University, Stanford, California 94305-2044
e-mail: jackman@stanford.edu, http://jackman.stanford.edu

Bayesian simulation is increasingly exploited in the social sciences for estimation and inference of model parameters. But an especially useful (if often overlooked) feature of Bayesian simulation is that it can be used to estimate any *function* of model parameters, including "auxiliary" quantities such as goodness-of-fit statistics, predicted values, and residuals. Bayesian simulation treats these quantities as if they were missing data, sampling from their implied posterior densities. Exploiting this principle also lets researchers estimate models via Bayesian simulation where maximum-likelihood estimation would be intractable. Bayesian simulation thus provides a unified solution for quantitative social science. I elaborate these ideas in a variety of contexts: these include generalized linear models for binary responses using data on bill cosponsorship recently reanalyzed in *Political Analysis*, item–response models for the measurement of respondent's levels of political information in public opinion surveys, the estimation and analysis of legislators' ideal points from roll-call data, and outlier-resistant regression estimates of incumbency advantage in U.S. Congressional elections.

1 Bayesian Simulation: Estimation, Inference, and Communication

SOCIAL SCIENTISTS ARE increasingly turning to Bayesian simulation, or, more specifically, Markov chain Monte Carlo (MCMC). This class of methods has an extremely broad range of application and offers a unified framework for social scientific statistical practice. Briefly,[1] in a Bayesian framework, interest focuses on the posterior density for parameters

Author's note: A technical version of this paper is available at the *Political Analysis* website. Portions of this research are from work in progress with Stanford colleagues and students, including Josh Clinton, Doug Grob, Douglas Rivers, and Paul Sniderman. I thank Richard Juster and David Primo for useful comments and suggestions, Keith Krehbiel for sharing some data, and David Draper for his inspiring short-course on "Bayesian Hierarchical Modeling" at Interface 2000, New Orleans, April 2000. I gratefully acknowledge support from a variety of sources at Stanford University: the Department of Political Science, and its chair, Barry Weingast; the Stanford Institute for the Quantitative Study of Society, and its director, Norman Nie; Stanford's Presidential Research Grants for junior faculty; and a grant from Stanford's Office of Technology Licensing. Parts of this research were presented at the meetings of the Southern California Methodology Program (SCAMP) at the University of California, Santa Barbara, May 2000.

[1] A detailed review of the mechanics and statistical heritage of Bayesian simulation is not necessary here. In Jackman (2000) I provide an exposition with applications drawn from political science.

Copyright 2000 by the Society for Political Methodology

$\boldsymbol{\theta}$, written as $\pi(\boldsymbol{\theta} \mid \text{data})$. But a distinguishing feature of Bayesian simulation is its reliance on the following principle.

> *Anything we want to know about a random variable x, we can learn by sampling from $g(x)$, the probability density function of x.*

Moreover, the *precision* with which we learn about features of x is limited only by the number of random samples from $g(x)$ we are prepared to wait for our computer to draw for us. This principle—the Monte Carlo method—has long been known to statisticians (e.g., Metropolis and Ulam 1949), but only recently have increases in computing power made this principle useful to quantitative social scientists.

Bayesian simulation greatly simplifies the interrelated tasks we face as applied statisticians: estimation, inference, and communication. To learn about $\boldsymbol{\theta}$ we simply tell a computer to sample many times from the posterior density for $\boldsymbol{\theta}$. To communicate what we have learned about $\boldsymbol{\theta}$ from the data, we can present summaries of those samples to our readers in a histogram or via some numerical summary. For instance, we can summarize the posterior with indicators of location (e.g., a mean, median or mode) and dispersion (e.g., a standard deviation, or quantiles that bound a confidence interval). Hypothesis testing amounts to nothing more than simply noting how many of the sampled values of $\boldsymbol{\theta}$ lie above or below zero or some other threshold of interest.

1.1 *Why* Bayesian *Simulation?*

The Bayesian underpinnings of Bayesian simulation are most evident in two ways. First, working in the Bayesian framework invites researchers to specify priors for the model parameters. To the extent that researchers employ priors that are not "flat" (it is always possible to use so-called "informative" priors), the posteriors will differ from the results of a traditional likelihood analysis.[2] But for the most part, informative priors are not used by researchers employing these methods.

Rather, what is more uniquely "Bayesian" about Bayesian simulation is the particular way we obtain samples from the joint posterior density $\pi(\boldsymbol{\theta} \mid \text{data})$. The Gibbs sampler—the workhorse of Bayesian simulation—partitions the vector of unknown quantities $\boldsymbol{\theta}$ into components, say, $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_J)'$, and samples from the conditional distributions for each component of $\boldsymbol{\theta}$. The relationships among conditional, joint, and marginal distributions are well understood by Bayesians, and the logic that drives the Gibbs sampler was first exploited in statistical settings by Bayesians. More concretely, if I want a sample from $\pi(\theta_1, \theta_2 \mid \text{data})$, I merely iterate the following two steps. To obtain sample t,

1. sample $\theta_1^{(t)}$ from the conditional density $g_1(\theta_1 \mid \theta_2^{(t-1)}, \text{data})$
2. sample $\theta_2^{(t)}$ from the conditional density $g_2(\theta_2 \mid \theta_1^{(t)}, \text{data})$

Thus at each t, we obtain $\boldsymbol{\theta}^{(t)} = (\theta_1^{(t)}, \theta_2^{(t)})'$. As t gets larger, the Gibbs sampler produces successively better approximations to the desired posterior density, $\pi(\boldsymbol{\theta} \mid \text{data})$. That is, the output of the Gibbs sampler over large values of t can be validly regarded as samples from the posterior. Via the Monte Carlo principle, summaries of these samples amount to summaries of the posterior, which can be used for inference and communication.

[2] Via Bayes' rule, a posterior is proportional to the prior times the likelihood: if the prior is flat over the parameter space supporting the likelihood, then the posterior is simply proportional to the likelihood, and the results of Bayesian and likelihood analyses coincide.

1.2 *Bayesian Simulation Recovers Any Function of the Parameters*

This simulation-based approach has uses far beyond estimation, inference, and communication of model parameters. A popular use of Bayesian simulation is imputation for missing data (e.g., Schafer 1997; Rubin 1987). But *any* function of the model parameters can also be calculated and stored over the iterations of an MCMC algorithm. This is an extraordinarily powerful if underappreciated feature of Bayesian simulation. From the perspective of Bayesian simulation, parameters, missing data, and auxiliary quantities are just "stuff" unobserved by the analyst. "Stuff" is a deliberately broad and vague term, since (as I show in the examples which follow) the class of things that are "auxiliary quantities of interest" is surprisingly wide.

The researcher's uncertainty over this "stuff" is driven entirely by uncertainty in the model parameters. To see this, let $\psi = h(\boldsymbol{\theta})$ be an estimand of interest, functionally related to the model parameters $\boldsymbol{\theta}$. Substantive interest focuses on $\pi(\psi \mid \text{data})$, the posterior density for ψ. This posterior is obtained via the identity $\pi(\psi \mid \text{data}) = \int_{\boldsymbol{\Theta}} \pi(\psi \mid \boldsymbol{\theta})\, \pi(\boldsymbol{\theta} \mid \text{data})\, d\boldsymbol{\theta}$, where $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Bayesian simulation performs this integration by Monte Carlo methods: letting $t = 1, \ldots, T$, (1) sample $\boldsymbol{\theta}^{(t)}$ from $\pi(\boldsymbol{\theta} \mid \text{data})$, the posterior of $\boldsymbol{\theta}$; (2) evaluate $\psi^{(t)} = h(\boldsymbol{\theta}^{(t)})$. The quantities $\psi^{(1)}, \ldots, \psi^{(T)}$ comprise a sample from the target distribution $\pi(\psi \mid \text{data})$, which can be summarized for inference and then communicated to readers.

So, with almost no extra cost, MCMC algorithms can be augmented to output these auxiliary quantities. Depending on the application, these quantities may be vital. For instance, predicted values and comparative statics are critical in policy settings, where policymakers want forecasts for y conditional on configurations of covariates they control. Measures of lack of fit such as a misclassification rate are important, say, when decision makers have asymmetric costs associated with over-/underprediction.

Analysts often forget that auxiliary quantities have any uncertainty associated with them at all. Bayesian simulation methods recover with this uncertainty automatically. Moreover, Bayesian simulation methods provide *arbitrarily precise* approximations to the finite sample distributions of these auxiliary quantities. This is because with a given data set, model, and priors, we can learn about any feature of the posterior of any random quantity up to any degree of precision via the Monte Carlo principle, limited only by the length of the stream of output from the MCMC algorithm.

Contrast this with conventional maximum likelihood estimation (MLE). MLE yields a point estimate of $\boldsymbol{\theta}$ and a standard error. Armed with these two outputs from MLE, analysts then rely on asymptotic normality to characterize uncertainty in $\boldsymbol{\theta}$. But in any finite sample, asymptotic normality is merely an approximation; the quality of the approximation depends not just on the sample size, but also on the nature of the quantity being estimated. Sometimes the exact finite sample posterior may be skewed or have heavy tails, even in a reasonably large sample. If the asymptotic normal approximation is poor, faulty inferences and model predictions can result. This is a real danger for "postestimation" simulation procedures (e.g., Herron 1999; King et al. 2000), which repeatedly sample from the asymptotic multivariate normal distribution for $\boldsymbol{\theta}$ to build an approximation to the posterior for an auxiliary quantity $\psi = h(\boldsymbol{\theta})$.[3] As I show below, Bayesian simulation overcomes these potential pitfalls.

[3] Earlier examples of this approach in political science literature include Bartels (1993, p. 274) and Jackman (1994, p. 327), discussed by Mooney (1997, p. 66).

2 Bayesian Simulation Unifies Social Scientific Statistical Practice

More than 10 years have passed since the publication of Gary King's (1989) *Unifying Political Methodology*. King laid out an influential vision of political methodology in which MLE would occupy center stage. In King's skillful hands, "maximum likelihood" became more than just a property of an estimator, but a *paradigm* for social scientific statistical practice. This is a good thing. MLE encourages social scientists to think about their data from first principles, to specify probability models appropriate to the data, and to do better than cram statistical practice into the framework of least-squares linear regression.

Bayesian simulation has the potential to become the unifying principle for social scientific statistical practice in the early 21st century. Bayesian simulation is at least as general as MLE and is in no way hostile to the maximum-likelihood paradigm. For Bayesians, likelihood is simply how one moves through the data, from priors to posterior beliefs about model parameters. And when prior beliefs about parameters are vague, the likelihood approach and the Bayesian approach yield identical results. But at the frontiers of quantitative social science, MLE can run into trouble. As I have claimed elsewhere,

> Substantively interesting statistical models can give rise to complex likelihood functions, having either lots of parameters or a computationally intractable functional form, or both . . . . Maximization algorithms may reach terminal solutions extremely slowly or not at all . . . . In other cases the likelihood will be known *a priori* not to have a unique maximum . . . . In yet another class of cases, the researcher may want to estimate not just parameters, but the values of missing data points as well, complicating the optimization problem substantially . . . . (Jackman 2000, p. 377)

But where MLE is cumbersome or intractable, Bayesian simulation can help. As King et al. (2000, p. 353) put it,

> . . . There is a simulation-based alternative to nearly every analytical method of computing quantities of interest and conducting statistical tests, but the reverse is not true. Thus, simulation can provide accurate answers even when no analytical solutions exist.

This endorsement of simulation appears in the Harvard team's announcement of *Clarify*, software for simulating from (asymptotically valid) normal approximations for the distribution of β and, in turn, characterizing uncertainty in certain auxiliary quantities. And elsewhere, Gary King and other coauthors have announced the release of *Amelia*, software for generating multiple imputations for missing data (King et al. 1998). But Bayesian simulation encompasses *both* these tasks. As I stressed earlier, from the perspective of Bayesian simulation *they are the same task*. Moreover, Bayesian simulation does not rely on asymptotically valid normal approximations to sampling distributions for model parameters. Bayesian simulation is not something we do "postestimation," requiring that point estimates and standard errors be estimated first via MLE or that we impute missing data prior to parameter estimation. Rather, Bayesian simulation encompasses the tasks of estimation, inference, and communication more or less simultaneously; Bayesian simulation provides answers at least as good as those provided by MLE; and Bayesian simulation lets us estimate models that cannot be handled by other approaches.

In the pages that follow, I use four examples to elaborate these ideas.

1. A study of bill cosponsorship in the U.S. House of Representatives; Bayesian simulation is used to compute residual diagnostics and assess model misspecification.
2. The measurement of political information in a public opinion survey in France: Bayesian simulation is used to recover estimates of the the political information of survey respondents and simultaneously to judge the quality of the survey items.

3. Estimation of the ideological locations of U.S. Senators, as revealed by legislative roll calls.
4. Resistant regression estimates of incumbency advantage in elections to the U.S. House of Representatives, 1956–1994.

3 Probits and Logits Are Missing Data

Probit and logit models (binary response generalized linear models, or GLMs) are regression models in which the dependent variable y_i^* is observed only in terms of its sign:[4] i.e., for probit, $y_i^* = \mathbf{x}_i\boldsymbol{\beta} + \epsilon_i$, $\epsilon_i \overset{\text{iid}}{\sim} N(0,1)$, where $y_i = 0 \Rightarrow y_i^* < 0$ and $y_i = 1 \Rightarrow y_i^* \geq 0$. That is,

> *qualitative response models are regression models, where the dependent variable is missing or partially observed.*

If we could impute values for the "real" dependent variable, y^*, we would have an ordinary regression model, for which estimation of $\boldsymbol{\beta}$ is straightforward. To make these imputations for y^*, note that the probit model implies the following conditional densities for y_i^*:

$$y_i^* \mid (y_i = 1, \mathbf{x}_i\boldsymbol{\beta}) \sim N(\mathbf{x}_i\boldsymbol{\beta}, 1)I(y_i^* \geq 0) \tag{1}$$

$$y_i^* \mid (y_i = 0, \mathbf{x}_i\boldsymbol{\beta}) \sim N(\mathbf{x}_i\boldsymbol{\beta}, 1)I(y_i^* < 0) \tag{2}$$

i.e., truncated normal distributions. Bayesian simulation generates samples from the posterior density of $\boldsymbol{\beta}$ by iterating the following scheme: at iteration t, (a) given $\boldsymbol{\beta}^{(t-1)}$, sample from the distributions in Eqs. (1) and (2), generating the vector $\mathbf{y}^{*(t)}$; (b) sample $\boldsymbol{\beta}^{(t)}$ from the conditional distribution for $\boldsymbol{\beta}$ implied by the regression of $\mathbf{y}^{*(t)}$ on $\mathbf{X}$. Further details appear in the on-line Appendix.

3.1 *Data: Cosponsorship*

The application is a study of legislative behavior by Krehbiel (1995), reanalyzed in this journal by Herron (1999). The dependent variable is a binary indicator, coded 1 if members of the U.S. House of Representatives chose to cosponsor bill H.R. 3266 and 0 otherwise (228 legislators cosponsored H.R. 3266, of the 434 legislators for which data are available). H.R. 3266 was a wide-ranging spending bill designed to circumvent the usual budget-making process, considered by the 103rd House of Representatives in 1993–1994. Krehbiel (1995) describes the broader political context and the eventual fate of H.R. 3266. Seven covariates are used in the analysis: a measure of each member's liberalism as scored by the interest group Americans for Democratic Action (ADA), a measure of fiscal conservatism published by the National Taxpayers' Union (NTU), an indicator variable for Democratic Party membership, a measure of Congressional seniority (years since first election and, thus, inversely proportional to seniority), the electoral margin of the member, an indicator for membership of the House Appropriations Committee, and an indicator for membership of the House Budget Committee.

[4] Bayesian simulation algorithms for binary response models are described by Albert and Chib (1993) and Johnson and Albert (1999); I also used Bayesian simulation for a probit model in an expository paper (Jackman 2000).

These data and maximum-likelihood estimates of the probit model are summarized in the on-line Appendix. My primary interest here is not in retelling the substantive story; Krehbiel's multivariate analysis suggests that after controlling for legislators' policy preferences (as measured with ADA and NTU scores), Democrats were actually more likely to support H.R. 3266 than Republicans. Seniority is also a key predictor, with junior members more likely to cosponsor this legislation than members with greater seniority.

3.2 *Goodness-of-Fit Summaries*

Herron (1999) reminds us that summaries of model fit from binary response GLMs are subject to uncertainty. Herron focuses on a goodness-of-fit measure popular within political science: "percentage correctly predicted" (PCP), the percentage of cases that have predicted probabilities p_i lying on the correct side of a classification threshold c such that if $y_i = 1$ and $p_i > c$, a correct prediction is recorded; $p_i \leq c$ is considered a correct prediction if $y_i = 0$. Usually c is set at 0.5, although there is seldom a good justification for this particular choice, or for counting misclassification of $y_i = 1$ the same as misclassification of $y_i = 0$.

Uncertainty in measures such as PCP stems from uncertainty in the predicted probabilities in these models, which in turn stems from uncertainty in the model parameters. Specifically, since $p_i \equiv \Pr(y_i = 1) = F(\mathbf{x}_i\boldsymbol{\beta})$, uncertainty in $\boldsymbol{\beta}$ generates uncertainty in the p_i and in any summary measure using the p_i, such as PCP. From a Bayesian perspective, uncertainty in $\boldsymbol{\beta}$ is completely characterized by the posterior distribution $\pi(\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{X})$. Bayesian simulation provides an arbitrarily precise approximation to $\pi(\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{X})$, drawing an arbitrarily large number of samples, $\boldsymbol{\beta}^{(1)}, \ldots, \boldsymbol{\beta}^{(T)}$. With each sampled $\boldsymbol{\beta}^{(t)}$, any function of $\boldsymbol{\beta}$ such as the predicted probabilities p_i can be calculated, yielding an arbitrarily precise characterization of uncertainty over these auxiliary quantities. In turn, functions of the p_i such as PCP can also be calculated at each iteration, yielding $\text{PCP}^{(1)}, \ldots, \text{PCP}^{(T)}$ for summarizing and communicating to readers.

In contrast, working in a classical setting, Herron (1999) labels uncertainty in auxiliary quantities such as p_i as "postestimation uncertainty." *After* estimating the model parameters $\boldsymbol{\beta}$ by MLE, Herron samples from the (asymptotic) multivariate normal distribution for $\boldsymbol{\beta}$ and then generates predicted probabilities and PCP with each sampled $\boldsymbol{\beta}$; this is also the idea underlying the *Clarify* program (King et al. 2000, p. 353).

Bayesian simulation shows that the posterior density for PCP is centered at 90.8% (close to the 90.6% point estimate implied by the MLEs) and with a 95% coverage interval ranging from 89.6 to 91.7%, closely corresponding to Herron's estimates. After taking into account the uncertainty in the parameter estimates, the model's predictive performance remains extremely high, at least as gauged by PCP. Bayesian simulation offers little over the two-step "estimate-then-simulate" procedure for this simple example; with $n > 400$, the (asymptotically valid) normal approximation for the posterior of $\boldsymbol{\beta}$ is good. But the point of this simple example is to highlight that in the Bayesian simulation context, we learn about quantities such as p_i while simultaneously learning about the model parameters; in fact, the phrase "postestimation uncertainty" is not meaningful in the context of Bayesian simulation.

3.3 *Residuals Are Missing Data*

Also extremely useful in diagnosing model fit (or lack thereof) are residuals. Of course, inspection of residuals is perhaps the most widely known and frequently employed diagnostic

aid for ordinary regression models; many types of model misspecification are readily identified by visual inspection of a regression's residuals. But how do we perform residual analysis in the case of a GLM?

Most political scientists are unaware of the fact that residuals can be recovered from a probit or logit. But statisticians have defined several types of residuals for binary response GLMs (e.g., Venables and Ripley 1999, p. 217). Here I focus on a type of residual that Bayesian simulation makes available to us: "latent" or "Bayesian" residuals (e.g., Albert and Chib 1995). These quantities are simply the difference between the latent y_i^* and their fitted values, $\mathbf{x}_i\beta$. In a typical logit or probit analysis, y_i^* is completely ignored by the analyst. But for Bayesian simulation, y_i^* is central to the recovery of the parameters themselves: recall that Bayesian simulation in this context amounts largely to treating the y_i^* as missing data. Hence not only are the y_i^* recovered, but with very little additional effort, we can also recover $e_i^{*(t)} = y_i^{*(t)} - \mathbf{x}_i\beta^{(t)}$, the latent residuals, at each iteration t. Moreover, these latent residuals have uncertainty associated with them, stemming from uncertainty in β, and so we recover not just point estimates of the e_i^*, but their posterior densities.

Armed with the posterior densities of the latent residuals, we are immediately in a position to authoritatively identify outliers and misclassifications. We can construct 95% intervals and the like for the e_i^* and identify cases with confidence intervals that do not overlap zero as outliers. This provides a much more nuanced and scientific approach to diagnosing lack of fit in binary response models than is typical.

Figure 1 presents two views of the latent residuals from the cosponsorship probit model. The top panel shows the latent residuals by observation number. Each plotted point is the median of 1000 draws from the posterior distribution for each latent residual. The vertical lines indicate 95% confidence intervals on those residuals distinguishable from zero. The four largest negative and positive residuals are labeled with the legislator's name. In the lower panel in Fig. 1, I plot the latent residuals by the model's predicted probabilities, roughly analogous to the way we might plot residuals against y or $\hat{y}$ in a Gaussian regression setting. Again, residuals with confidence intervals distinguishable from zero are highlighted, this time with large plotting symbols (open circles indicate $y_i = 1$, cosponsorship; filled circles indicate $y_i = 0$).

Most of the residuals are closely clustered around zero and are indistinguishable from zero, confirming that the probit model fits the data well. But inspection of the latent residuals shows us that the model makes not just a number of classification errors, but a number of *large* classification errors. The most striking errors are for 2 Republicans, and indeed, there are just 3 Republicans among the 32 observations with latent residuals distinguishable from zero. Republicans Hoke (OH-20) and R. Baker (LA-6) are overwhelmingly predicted to have cosponsored H.R. 3266 but did not do so. Both are relatively junior members of Congress, although Hoke is a member of the Budget Committee, and both have NTU scores slightly higher than the average Republican NTU score. In fact, no one with a NTU score higher than 59 failed to cosponsor H.R. 3266 *except* for these two Republicans. In addition, nine Democrats who also did *not* cosponsor H.R. 3266 pick up latent residuals significantly different from zero, the largest being that for J. Long, an Indiana Democrat.

On the side of underprediction (positive latent residuals), the four largest latent residuals (by posterior medians) are those for D. Peterson (FL-2), R. Wyden (OR-3), M. Lloyd (TN-3), and G. Green (TX-29); all cosponsored H.R. 3266, while the model predicts that they would not. All score quite low on the NTU measure (14, 24, 25, and 18, respectively; the median Democratic NTU score is 20); Peterson has the lowest NTU score among all

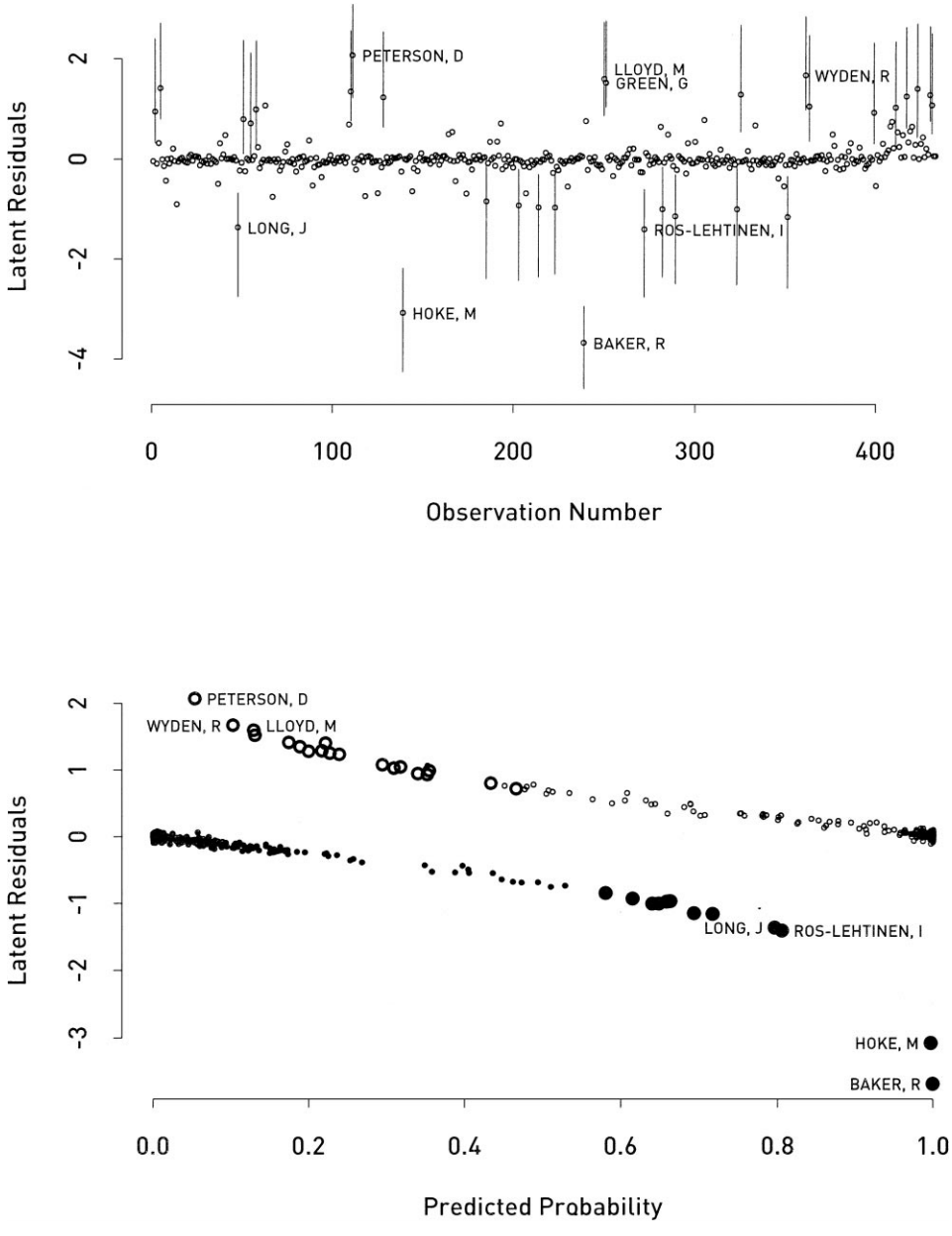

Fig. 1 Residual diagnostics, cosponsorship data.

cosponsors, followed by Green's score of 18. However, these errors are nowhere near as striking as those for the two Republicans Hoke and Baker.

At this stage we might go back to the data and reconsider the specification we estimated. With the model correctly classifying roughly 90% of the cosponsorship decisions, we might not be troubled by the enormous residuals for these two Republicans and the cluster of poor predictions on the Democratic side. But looking at these outlying data points more closely might prompt us to reconsider the model specification. Is the mapping from NTU score

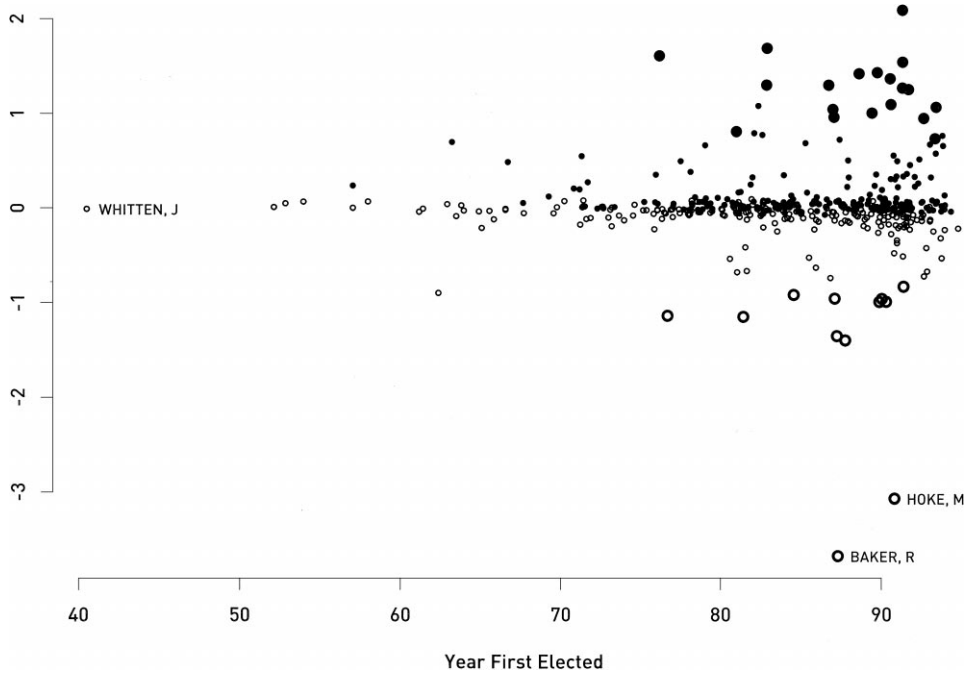

Fig. 2 Latent residuals, by year first elected. Smaller plotting symbols indicate latent residuals indistinguishable from zero (their 95% confidence interval overlaps zero); larger plotting symbols indicate outlying observations. Filled symbols are instances of cosponsorship ($y_i = 1$); open symbols indicate noncosponsors.

to the latent variables y_i^* linear or conditional on partisanship? For instance, including the square of the NTU score produces a slight improvement in fit, and the implied quadratic in NTU shows a steep impact on y_i^* at low to moderate NTU scores, leveling out at high NTU scores. Similarly, we could estimate a nonparametric fit in NTU scores (e.g., Beck and Jackman 1998).

To provide a taste of a reanalysis along these lines, consider Fig. 2, where I plot the latent residuals against the predictor year first elected. This plot is revealing, showing that larger, significant latent residuals are concentrated among relatively junior legislators. In fact, this diagnostic plot shows an almost-textbook example of heteroskedasticity, with the cosponsorship decisions of more junior legislators being significantly less predictable than their more senior colleagues. That is, while we find a strong tendency for more junior legislators to be more likely to cosponsor H.R. 3266, we also find more *variation* in their cosponsorship decisions.

This finding is quite striking in substantive terms (suggesting that legislators become more predictable with seniority), but also showcases the power of Bayesian simulation in recovering the posterior densities of the latent residuals. There was little reason to suspect heteroskedasticity with respect to seniority until we actually recovered the latent residuals and determined which residuals were large and significant, putting us in a position to generate some graphical diagnostics. In response to this finding we might consider fitting a richer specification, trying to capture this apparent heterogeneity, perhaps via a heteroskedastic probit model (e.g., Alvarez and Brehm 1995). I reserve consideration of such a model for another time and place.

4 Political Information Is Missing Data

Political information lies at the heart of modern understandings of public opinion.[5] For instance, this variable plays a critical role in Zaller's (1992) account of contemporary American public opinion. Similarly, in the analysis of question wording experiments embedded in sample surveys, Paul Sniderman and his partners use measures of political information as a critical *conditioning* variable. That is, levels of political information govern the magnitude of responses to experimental treatments, with low-information respondents generally more swayed by ideologically disconsonant treatments than high-information respondents (e.g., Sniderman and Theriault 1999).

The American National Election Studies have helped pioneer a very popular measure of political information via a series of "objective" items, that is, questions about politics and political leaders that have correct answers. For instance, respondents are asked what job William Rehnquist has, which party controls both houses of Congress, and so on. These items have become "industry standard" measures of political information.

4.1 *Data: Political Information in France*

As part of a recent study of public opinion in France, Paul Sniderman, myself, and our French partners came up with a list of 12 political information items. We administered these items to our respondents as "true" or "false" propositions toward the end of the interview. This is the first time that "objective" measurements of political information have been attempted in France, and we faced considerable uncertainty as to how our test items would fare. Were our items too hard or too easy? Do some items tap political information more so than others? What are the properties of any resulting scale measure?

The items appear in Table 1. Two pretests of 26 and 25 interviews, respectively, were conducted in April 2000. Each respondent was administered 10 items, and the items varied over the two pretests according to the pattern shown in Table 1. We immediately see considerable variation in the rate of correct responses; just 18% of respondents knew that the election of deputies to the National Assembly does not take place via proportional representation (item 8), while 86% correctly responded "false" to proposition 6, regarding the length of a presidential term.

4.2 *Item–Response Model*

To assess more scientifically the utility of our political information items, I use the following *two-parameter item–response model*:

$$p_{ij} \equiv \Pr[y_{ij} = 1] = F(\beta_{j1} x_i - \beta_{j2}) \tag{3}$$

where

- y_{ij} is the ith respondent's answer to the jth political information item (1 if correct, 0 if incorrect, with no answers considered an incorrect response);
- x_i is the ith respondent's level of political information;

[5] Mondak (2000) provides an excellent summary of the ways political information has been used in recent empirical studies of public opinion and political behavior.

Table 1 Political information items, pretest of French study

	Item	*Correct response*	*Percentage correct*	*n*
1.	Michèle Alliot Marie est la présidente du RPR	True	77	26
2.	La Finlande fait partie de l'Union européene	True	47	51
3.	Jean Pierre Chevènement appartient au Parti socialiste	False	35	26
4.	Le premier ministre a le droit de dissoudre l'Assemblée nationale	False	59	51
5.	Il y a des ministres communistes dans le gouvernement de Lionel Jospin	True	78	51
6.	Le Président de la République est élu pour un mandat de 5 ans	False	86	51
7.	Le Sénat a le pouvoir de renverser le gouvernement	False	55	51
8.	Les députés sont élus au scrutin proportionnel	False	18	51
9.	Les étrangers qui résident en France depuis 5 ans ont le droit de voter à l'élection présidentielle	False	57	51
10.	L'Etat aide financièrement les partis politique	True	71	51
11.	Laurent Fabius appartient au Parti socialiste	True	76	25
12.	Jorg Haider est le leader du parti libéral autrichien[a]	False	20	25

[a]Haider had resigned the leadership of Austria's Freedom Party when we pretested these items.

- β_{j1} is an unknown parameter, tapping the *item discrimination* of the jth item, the extent to which the probability of a correct answer responds to levels of political information;
- β_{j2} is an unknown *item difficulty* parameter, tapping the probability of a correct answer irrespective of levels of political information; and
- $F(\cdot)$ maps from the real line to the unit probability interval.

Here I set $F(\cdot)$ to the normal CDF $\Phi(\cdot)$, making Eq. (3) a probit model, with separate slope and intercept parameters for each item.[6] But this is a rather unusual binary response model in which the entire right-hand side is missing "stuff." Notice that we do not observe the "covariate" x_i, the respondent's level of political information, which can be regarded as missing data.

An additional problem is that the model parameters are not identified. Any rescaling of the latent traits x_i is consistent with the observed binary responses, via offsetting rescalings of the item parameters. This problem is sometimes solved by imposing an identifying additivity constraint $\sum_{i=1}^{n} x_i = 0$ (e.g., Bock and Aitken 1981). In the Bayesian approach a proper prior density over the x_i solves this problem, providing a scale for the x_i.

4.3 *Estimation*

Prior to Bayesian simulation methods, estimating the two parameter item–response model was quite formidable with large data sets. For instance, with, say 1000 subjects and 50 items, we have 1100 parameters to estimate (1000 x_i parameters and $2 \times 50 = 100$ item

[6]The choice of $F(\cdot)$ is not particularly consequential for these data, and logit and probit models are standard in the item–response literature.

parameters). Directly attempting to maximize the likelihood for the data gives rise to an 1100-dimensional optimization problem, which is hardly trivial, even by the standards of modern computing. Even the relatively small French pretest data set involves estimating 51 latent traits and 24 item parameters, for a total of 75 parameters. Prior to Bayesian simulation, researchers were forced to make compromises; item parameters were estimated via *marginal* maximum likelihood, ignoring uncertainty in the latent traits.

Bayesian solutions that overcome these weaknesses are available (e.g., Johnson and Albert 1999, Chap. 6) The intuition is to treat the latent traits (in my case, political information) as missing data. In addition, the probit models in Eq. (3) are linear regressions for the missing variable y_{ij}^*: i.e., $\Phi^{-1}(p_{ij}) \equiv y_{ij}^* = \beta_{j1}x_i - \beta_{j2} + \epsilon_{ij}$, where $\epsilon_{ij} \overset{\text{iid}}{\sim} N(0, 1)$. As in any binary response model, y_{ij}^* is observed only in terms of its sign: if respondent i answers item j correctly, then $y_{ij} = 1$ and $y_{ij}^* \geq 0$, and otherwise $y_{ij}^* < 0$. The idea underlying the Bayesian simulation approach is simple: with imputations for the missing x_i and the latent y_{ij}^*, estimate $\boldsymbol{\beta}_j$ by running regressions of y_{ij}^* on the x_i, $j = 1, \ldots, m$.

More formally, we seek the joint posterior of all random quantities in the model: (1) the missing levels of political information $\mathbf{x} = (x_1, \ldots, x_n)'$, (2) the item parameters $\boldsymbol{\beta} = (\beta_{11}, \beta_{12}, \ldots, \beta_{J1}, \beta_{J2})'$, and (3) the partially observed $\mathbf{Y}^* = \{y_{ij}^*\}$. Denote this joint posterior $\pi(\mathbf{x}, \boldsymbol{\beta}, \mathbf{Y}^* \mid \mathbf{Y})$ where $\mathbf{Y} = \{y_{ij}\}$ are the observed binary responses (1 if correct, 0 otherwise). A Gibbs sampler yields an arbitrarily exact approximation to this posterior density, by sequentially sampling from the following conditional distributions: (1) sample $\mathbf{Y}^{*(t)}$ from $g_1(\mathbf{Y}^* \mid \mathbf{x}^{(t-1)}, \boldsymbol{\beta}^{(t-1)}, \mathbf{Y})$, (2) sample $\mathbf{x}^{(t)}$ from $g_2(\mathbf{x} \mid \mathbf{Y}^{*(t)}, \boldsymbol{\beta}^{(t-1)}, \mathbf{Y})$, and (3) sample $\boldsymbol{\beta}^{(t)}$ from $g_3(\boldsymbol{\beta} \mid \mathbf{Y}^{*(t)}, \mathbf{x}^{(t)}, \mathbf{Y})$, where t indexes iterations of the Gibbs sampler. These conditional distributions and computational details are described in the on-line Appendix.

Vague priors are specified for the item parameters, and a $N(0, 1)$ prior is used for the political information items. Inspection of the Gibbs sampler output indicates that the algorithm is slowly meandering through the parameter space and that a long string of output is required to fully recover the posterior density of the parameters. After a burn-in of 10,000 iterations, I base inference and communication on a thinned series of simulations, generating 250,000 Gibbs samples and retaining every 250th sample.

4.4 *Results*

The posterior densities of the x_i are summarized in Fig. 3, and graphical summaries of the bill parameters appear in Fig. 4. Plotted points indicate the median of the posterior, and the lines extend out to cover a 95% confidence interval. For the political information scores (Fig. 3), the posteriors are considerably more narrow than the prior (the 95% interval for the $N(0, 1)$ prior is indicated by the dotted vertical lines at -1.96 and 1.96), confirming that the data do help us learn about the respondents' levels of political information. In general, we are most certain about those respondents in the middle of the distribution, while the confidence intervals grow large for respondents with low levels of political information and respondents with high levels of political information. For these respondents the test items are not particularly discriminatory. High-information respondents almost always answer correctly, and so we learn little about the upper bound of their political information. Conversely, low-information respondents almost always answer incorrectly, and so while we know that they are located at the low end of our distribution, we are not particularly sure just *how low* their scores are. Thus the confidence intervals for these extreme respondents are slightly asymmetric about their posterior medians.

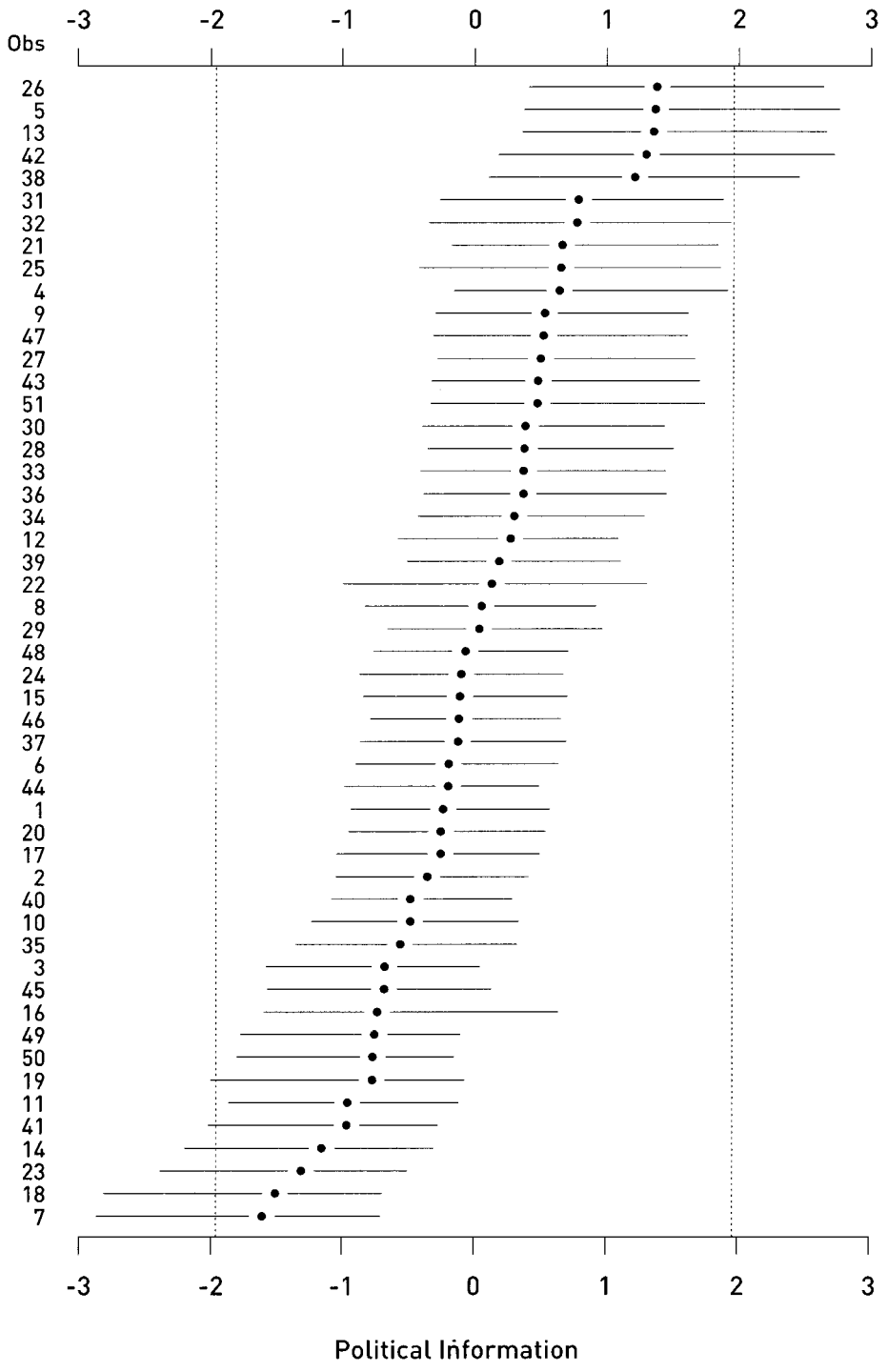

Fig. 3 Posterior densities, political information, French pretest data. Each plotted point is the median of 1000 samples from the posterior density of each respondent's political information score, x_i; the horizontal lines cover 95% confidence intervals. The observations have been ordered by the posterior medians. The dotted vertical lines show the 95% confidence interval for the $N(0, 1)$ prior.

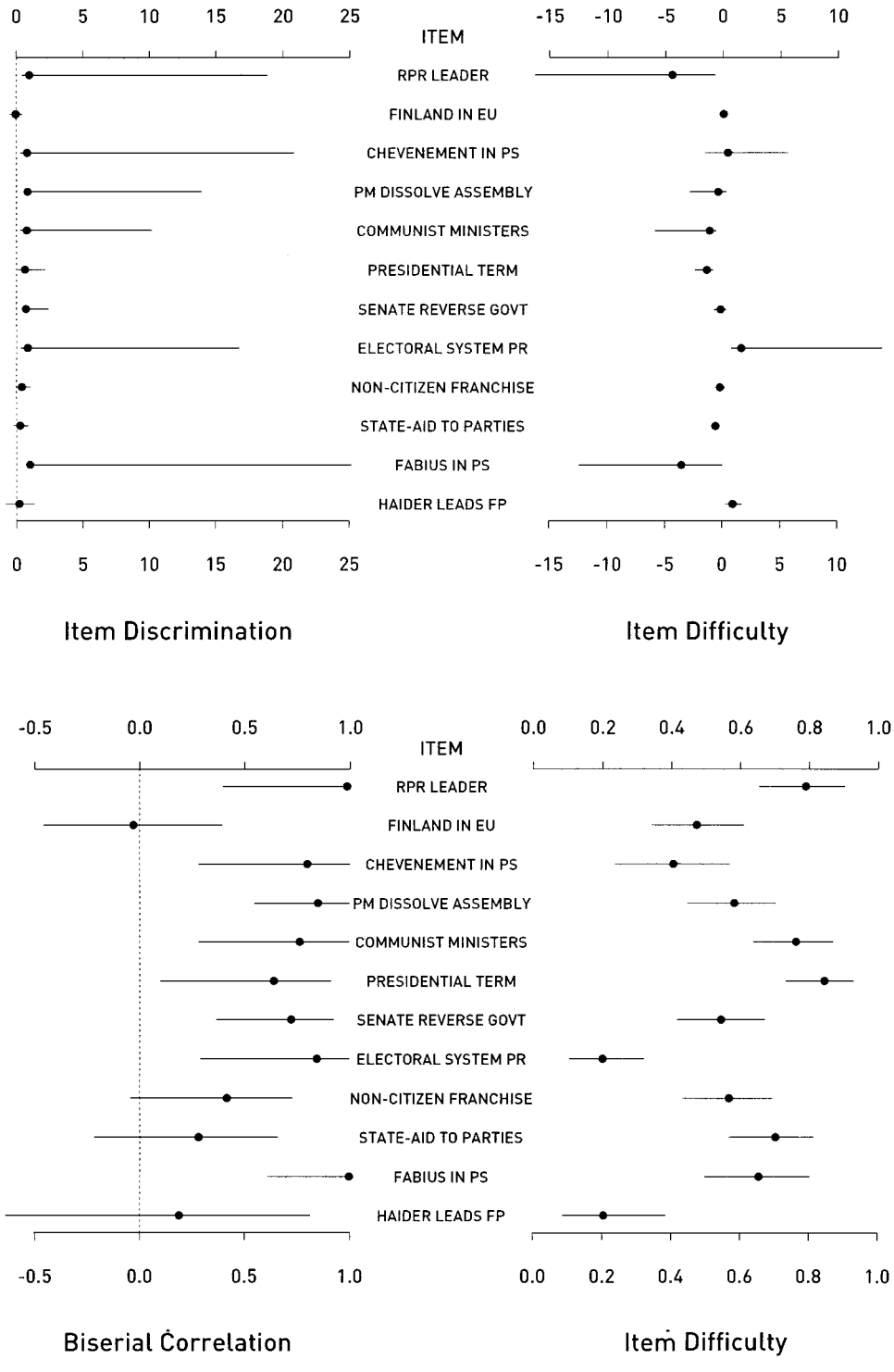

Fig. 4 Posterior densities, political information item parameters, French pretest data. Each plotted point is the median of 1000 samples from each quantity's marginal posterior density; the lines cover a 95% confidence interval.

To help gauge the quality of the items, I rely on two auxiliary quantities that are functions of the item parameters. First, the *biserial correlation* between the observed binary responses y_{ij} and the latent trait x_i is $r_j = \beta_{j1}/\sqrt{1+\beta_{j1}^2}$. If r_j is indistinguishable from zero, then item j and the latent trait are uncorrelated, that is, the item provides no information about the latent trait. On the other hand, r_j close to one implies that item j provides good discrimination with respect to the underlying trait. Since r_j is simply a function of β_{j1}, uncertainty in r_j is completely characterized by uncertainty in β_{j1}, so sampling repeatedly from the posterior for β_{j1} provides all the information required for inferences for r_j. Second, I calculate the expected proportion of respondents who would get the item correct (an alternative measure of item difficulty). This auxiliary quantity is $p_j = \Phi(-\beta_{j2}/\sqrt{1+\beta_{j1}^2})$, a function of both item parameters. Posterior summaries for these auxiliary quantities appear in the bottom half of Fig. 4, while the posteriors for the raw item parameters appear in the top half. Many of the posterior densities summarized in Fig. 4 are clearly not normal: asymptotically valid normal approximations would be a poor summary of uncertainty for the item parameters, and, in turn, auxiliary quantities that are functions of the item parameters.

The items about Finland's membership of the European Union and Haider's leadership of Austria's Freedom Party perform rather poorly, exhibiting correlations with the latent trait that are essentially zero: indeed, respondents seem to be guessing their responses to the Finland question, with the expected percentage correct for the Finland item right around 50%. Items that perform well include the question about the electoral system used in elections for the National Assembly and whether the Prime Minister can dissolve the National Assembly. Both these items yield good discrimination and large and significant correlations with the latent trait and yet are not too easy. The electoral system item is roughly the hardest item on the survey and so helps us discriminate among respondents with high levels of political information. We appear to have too many easy items with high discrimination: these include the item identifying the RPR leader (Michèle Alliot Marie), the party membership of Laurent Fabius, and, to a lesser extent, the questions about communist ministers and the length of the presidential term. These items help us discriminate among our most poorly informed respondents, but we probably do not need 3 of our 10 items to possess these properties. In short, this analysis suggests that we have not done too badly with this initial attempt at measuring political information in France: a nice distinction is apparent between the items about institutions and procedures and the items about parties and personalities, with respondents finding the latter items significantly easier than the former.

5 Legislative Ideal Points Are Missing Data

Estimates of legislative ideal points are extremely important to political science. These measures of preferences play two important functions. First, as measures per se, they permit substantively interesting *descriptions* of the ideological positions of elected representatives in the world's legislatures. But, second, armed with these measures, researchers are then able to *test* theories of legislative behavior, using ideal point estimates as "data" for use in further statistical work.

Keith Poole and Howard Rosenthal's NOMINATE measures of Congressional ideal points are virtually canonical. The NOMINATE model begins with standard postulates of utility maximization and works forward to a complex likelihood function that is estimable with roll call data. But a major shortcoming of their method is that the measures

it produces come with no uncertainty assessments, such as standard errors or confidence intervals.

It is well known that ideal point estimation in one dimension is *exactly the same statistical model* as the two-parameter item–response model presented in the previous section.[7] So just as we were able to exploit Bayesian ideas in the estimation of the item–response model, the same ideas let us estimate features of the legislative ideal point model never before possible. In particular, I present estimates of legislative ideal points accompanied by confidence intervals. These results are a sample of collaborative work currently under way with colleagues and students at Stanford (see Clinton et al. 2000).

To see how the ideal point model works, let n denote the number of legislators and m the number of roll calls included in the analysis. In most applications to legislatures, both n and m will be large. For example, in the U.S. Senate, $n = 100$, and for a single session, m is about 500. On the other hand, if we are attempting to recover the ideal points of, say, Supreme Court justices, then n will be quite small, as could be m.

Roll-call data can be arranged as an $n \times m$ matrix of zeros and ones $\mathbf{Y} = \{y_{ij}\}$, where y_{ij} indicates whether the ith legislator votes "yea" (1) or "nay" (0) on the jth roll call. For abstentions and other forms of missing roll calls, our simulation-based estimation approach allows us to make imputations for these data points. Keeping all data points in this way is a novel step in the ideal point literature but remarkably easy to implement once we recast the problem in a Bayesian context.

Roll calls are modeled using a unidimensional spatial model. Each legislator's ideal point x_i indicates his or her most preferred position in a subspace of Euclidean space. Since we examine a one-dimensional subspace, the natural interpretation is that this subspace represents the liberal–conservative continuum. Associated with each roll call j is a pair of positions in the subspace indicating the location associated with the passage (θ_j) and the rejection (ϕ_j) of roll call j. Legislators have quadratic utilities over the proposals, plus a random disturbance. The utility legislator i gets from the jth proposal is

$$u_i(\theta_j) = -(x_i - \theta_j)^2 + \eta_{ij} \tag{4}$$

while the utility obtained from retention of the status quo is

$$u_i(\psi_j) = -(x_i - \psi_j)^2 + \nu_{ij} \tag{5}$$

We assume that (η_{ij}, ν_{ij}) has a bivariate normal distribution, and, as is standard in the literature, we assume independence and homoskedasticity across legislators and roll calls. Now let $\sigma_j^2 = V(\eta_{ij}) - 2C(\eta_{ij}, \nu_{ij}) + V(\nu_{ij})$ and define $\epsilon_{ij} = (\eta_{ij} - \nu_{ij})/\sigma_j$ so ϵ_{ij} has unit variance. Let y_{ij}^* denote the (latent) utility difference between the proposal and status quo positions for the ith legislator, $y_{ij}^* = u_i(\theta_j) - u_i(\psi_j)$, and so we observe a "yea" vote when $y_{ij}^* > 0$ and a "nay" vote otherwise. Substituting Eqs. (4) and (5) and rearranging yields

$$y_{ij}^* = \beta_{j1} x_i - \beta_{j2} + \epsilon_{ij} \tag{6}$$

[7] Poole and Rosenthal (1997, p. 247) note this connection, citing work by Lord (1975) and Ladha (1991). See also Bailey and Rivers (1997). Londregan (2000) also provides an excellent summary of the connection between the two models.

where $\beta_{j1} = 2(\theta_j - \psi_j)$ and $\beta_{j2} = \theta_j^2 - \psi_j^2$. Thus, the probability of a "yea" vote is

$$Pr(\text{"Yea"}_{ij}) = Pr(y_{ij}^* > 0) = Pr(\epsilon_{ij} > -(\beta_{j1}x_i - \beta_{j2})) = F(\beta_{j1}x_i - \beta_{j2})$$

where F is the standard normal CDF Φ for our probit model. This is a hierarchical probit model with the complication that the "covariate" x_i is unobserved by the analyst and is, instead, treated as missing data. As we noted earlier, this model is equivalent to a two-parameter item–response model. For instance, β_{j1} is the *item discrimination* parameter for test item j, and the intercept term β_{j2} is the *item difficulty* parameter. In our setup x_i is the unobserved ideal point of legislator i, while in the item–response model this is the latent ability of test-taker i. Note also that we recover the midpoint between the proposal and the status quo, $(\theta_j + \psi_j)/2$ as β_{j2}/β_{j1}.

5.1 *Weaknesses of MLE*

As for the item–response model, direct MLE of all the unknown parameters is infeasible, and Poole and Rosenthal use an EM algorithm, alternating between two marginal maximum-likelihood problems: first, estimate the bill parameters conditional on the current estimates of the ideal points, then update the estimate of the ideal points conditional on the estimates of the bill parameters and iterate until convergence. And since the unidimensional ideal point model is isomorphic with a two-parameter item–response model, it comes as no surprise to note that this EM approach was a popular approach to the estimation of item–response models (e.g., Bock and Aitken 1981), up until the advent of Bayesian simulation methods. A key weakness of the EM approach is that it provides no assessments of uncertainty in the model parameters: that is, the EM algorithm finds its way to the top of the joint likelihood surface but does not offer a way to characterize uncertainty in the estimates. This is why NOMINATE scores are unaccompanied by uncertainty assessments such as standard errors or confidence intervals. In addition, the method flounders badly on unanimous votes (or, equivalently, test items that all subjects "pass" or all subjects "fail").

5.2 *The Bayesian Approach*

The Bayesian simulation approach overcomes the weaknesses of MLE or EM, recasting the ideal point model as a series of interrelated missing data problems. To see this, note that Eq. (6) for y_{ij}^* is a linear regression with the unobserved ideal points x_i and the unknown parameters $\boldsymbol{\beta}_j = (\beta_{j1}, \beta_{j2})'$ constituting the right-hand side. This simple linear model allows us to apply textbook results on the Bayesian linear model to recover the posterior densities for $\boldsymbol{\beta}_j$, conditional on imputations for the y_{ij}^* and the x_i. Put differently, treating the y_{ij}^* as missing data to be imputed makes the ideal point problem relatively simple: (1) conditional on imputations for the y_{ij}^* and the x_i, we have a regression structure in which $\boldsymbol{\beta}_j$ can be easily estimated; (2) conditional on imputations for the y_{ij}^* and the $\boldsymbol{\beta}_j$, we can rearrange Eq. (6) to obtain the "reverse regression" $y_{ij}^* + \beta_{j2} = \beta_{j1}x_i + \epsilon_{ij}$, treating the ideal point x_i as a parameter to be estimated in each of $i = 1, \ldots, n$ regressions. Armed with updated estimates for x_i and the $\boldsymbol{\beta}_j$, we can obtain imputations for the y_{ij}^* and repeat the previous steps. Of course, all of these imputations and estimates are obtained by *sampling* from conditional distributions for each quantity, as described in the previous section on item–response models. Further details appear in the on-line Appendix.

Finally, as with the item–response model, the Bayesian approach obliges us to specify priors for these unknown parameters. As in the item–response setup, I specify independent $N(0, 1)$ priors for each x_i. Vague priors are chosen for the bill parameters: $\boldsymbol{\beta}_j \sim N(\mathbf{0}, \kappa \cdot \mathbf{I})$, $\kappa = 100$.

5.3 *Data*

Here I work with a relatively small data set: all nonlopsided roll calls from the 105th U.S. Senate, which sat from January 1997 through October 1998. Lopsided roll calls were defined as those decided with fewer than one senator either for or against, yielding 534 roll calls from the full set of 612.[8] Note that the dimensions of this problem ($n = 100$ and $m = 534$) are small compared to multiyear analyses of the U.S. House of Representatives.

Of the 53,400 individual voting decisions analyzed, 941 or 1.8% are missing (abstentions and absences) and are imputed by treating them as additional quantities to be estimated conditional on the observed data, model structure, and parameter values. The median missing data rate is 0.94% by senator, but some notable outliers include John Glenn (missed 13.7% of the nonlopsided roll calls analyzed), Jesse Helms (12.2%), and Daniel Inouye (10.1%). Since Helms is a preference outlier, inferences could be sensitive to how we handle his missing roll calls; the methodology employed here allows us to impute these roll calls while simultaneously estimating all the other parameters of interest.

5.4 *Results*

The Gibbs sampler converges on the target posterior after a couple of hundred iterations and from a variety of different starting points. Nonetheless, I let the sampler run for 10,000 iterations and then saved an additional 2000 iterations for inference and communication.

Summaries of the ideal point posteriors appear in Fig. 5; the plotted points are the medians of 2000 Gibbs samples, and the horizontal lines correspond to 95% posterior confidence intervals. The point estimates are extremely close to Poole and Rosenthal's, save for differences in scaling; but the contribution here is to demonstrate the uncertainty accompanying the ideal point estimates. The stark differentiation by party is perhaps the most striking feature of the results, with no party overlap whatsoever. Armed only with point estimates for the x_i, we could not tell if this gap between the parties was statistically significant: now we know that it is not. Also noteworthy is the way the confidence intervals grow wider on the extremes of the distribution. The range of variability is quite large, with the confidence bounds for our most extreme legislators up to *three times* as wide as those for the legislators in the middle of the space. This is a seldom-noticed feature of ideal point estimates. Data analyses that treat the ideal point estimates as fixed are making heroic assumptions. In fact, whenever ideal points are used as dependent variables on the left-hand side of a regression equation, they are virtually guaranteed to generate heteroskedastic disturbances and threaten the validity of any resulting inferences.

In Fig. 6 I show the *ranks* of the ideal point estimates, along with confidence intervals on the ranks. These confidence intervals are obtained by sorting the legislators from left to right over 2000 draws from the posterior density of their ideal points. Here is yet another example of using Bayesian simulation to obtain auxiliary quantities of interest that are functions of the model parameters: in this case, we obtain the order statistics for the x_i. Uncertainty in these quantities is generated solely by uncertainty in the x_i themselves. In the Congress literature, interest focuses on those legislators occupying critical locations on the left–right ideological dimension presumed to underlie the ideal points. These critical locations include the chamber median, the party medians, and, in the

[8] Poole and Rosenthal define lopsided roll calls as those decided with fewer than 2.5% either for or against; our definition is slightly more expansive, allowing us to obtain slightly better resolution of extreme ideal points. With bounded priors on the bill parameters we could plausibly include the remaining 78 lopsided roll calls.

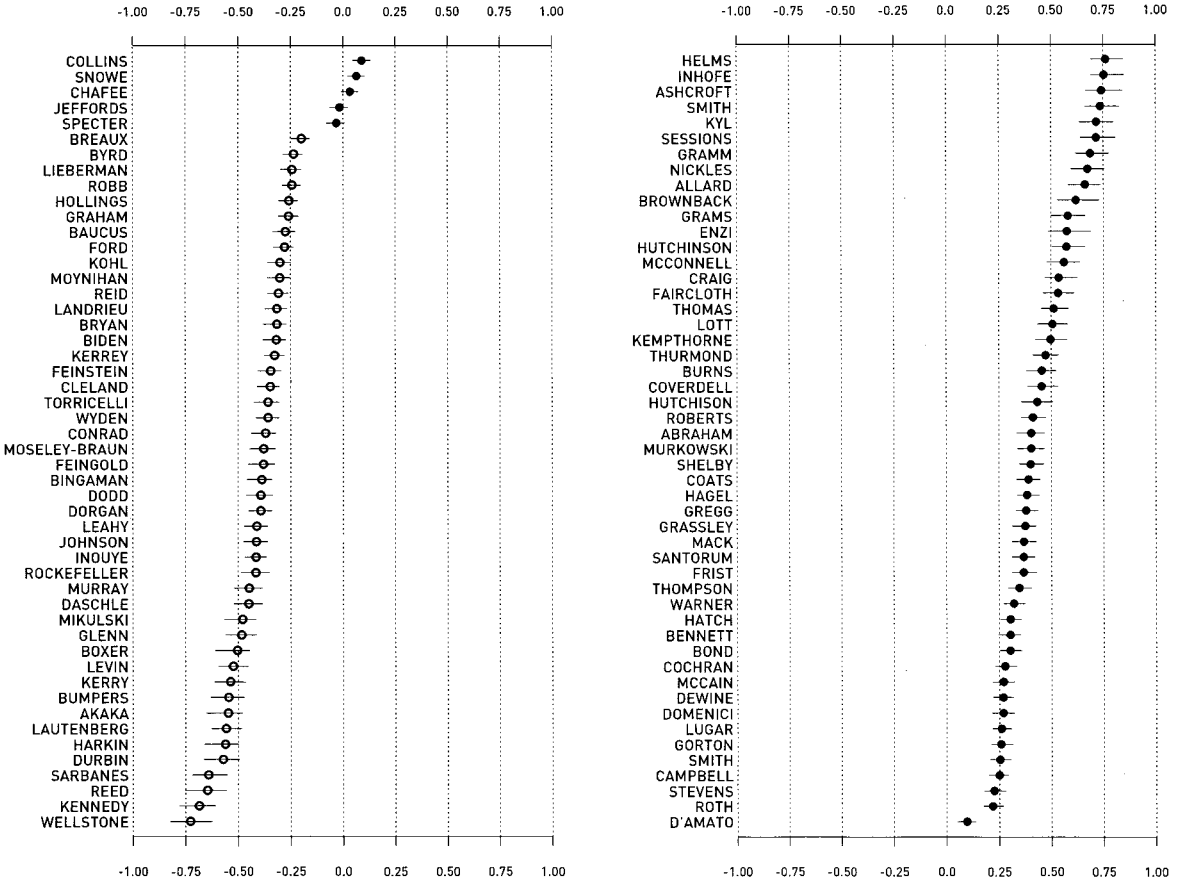

Fig. 5 Posterior Summaries, Ideal Points of the 105th U.S. Senate. Each point indicates the median of 2,000 draws from the posterior for each Senator's ideal point; the lines extend to the 2.5th and 97.5th percentiles, corresponding to a 95% posterior confidence interval. Solid dots indicate Republicans (to the right); open dots indicate Democrats (to the left).

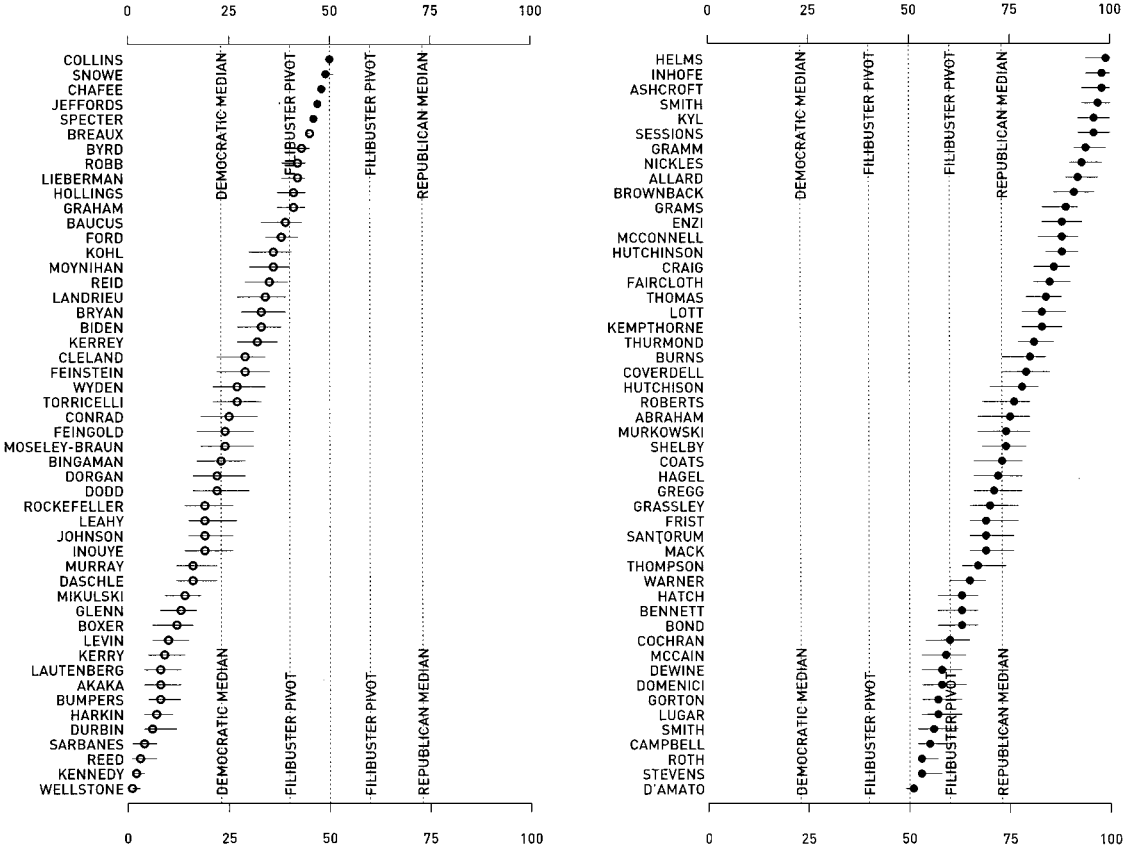

Fig. 6 Posterior Summaries, Ranks of Ideal Points of the 105th U.S. Senate. Each point indicates the median rank across legislators of 2,000 draws from the posterior for each Senator's ideal point; the lines extend to the 2.5th and 97.5th percentiles, corresponding to a 95% posterior confidence interval on each Senators' rank. Solid dots indicate Republicans (to the right); open dots indicate Democrats (to the left).

Senate, the filibuster pivots. There is considerable uncertainty in some of these quantities. There is very little doubt as to the identity of the chamber median: Republican senators Snowe, Collins, and D'Amato are the only senators whose estimated ranks are indistinguishable from 50. But any of 14 Democratic senators could plausibly be the Democratic median, as might a similar number of Republicans qualify as the Republican median. The filibuster pivots are also estimated with a reasonable degree of uncertainty. Nine Democratic senators have ranks that are indistinguishable from 40, the "left-hand" filibuster pivot, ranging from Robert Byrd to Daniel Patrick Moynihan. On the Republican side, 12 senators could be the filibuster pivot, with ranks indistinguishable from 60.

In short, relying on estimated ideal points alone to *order* legislators gives a false sense of precision. When we take into account the uncertainty associated with each ideal point, we notice that making authoritative distinctions among senators is not possible, at least not at conventional levels of statistical significance. We can distinguish a Ted Kennedy or a Paul Wellstone from many other Democrats, or a Jesse Helms from most other Republicans. But theories of legislative behavior typically do not concern these legislators in the tails of the distribution of ideal points. Other than the chamber median, it is quite difficult to resolve just which Senators are the party medians and filibuster pivots or, indeed, make any authoritative ordinal distinctions among the vast majority of legislators. This is not widely recognized but is made quite clear via the Bayesian simulation approach presented here.

This is just a taste of what Bayesian simulation can do for us in the context of estimating legislative ideal points. Other directions we are pursuing include

1. estimating the effects of district-specific and legislator-specific covariates on ideal points, effectively integrating the tasks of *measuring* and *analyzing* ideal points;
2. testing for dimensionality of the policy space, further exploiting the Bayesian simulation methods we employ here;
3. testing ideas about the progress of the agenda over the life of a legislature, exploiting our ability to recover estimates of bill parameters accompanied by confidence intervals; and
4. applying this methodology to smaller legislatures and judicial bodies.

6 Robust Estimation as a Missing Data Problem

Outlying data points can seriously distort estimates of *location*, such as means or regression coefficients. This is consequential when using normal distributions to characterize data that are more heavily tailed. Location estimates obtained via a normal likelihood for such data will be quite sensitive to the data in the tails, risking faulty inferences.

The body of literature on robust statistics is colossal, and I do not attempt to summarize it here.[9] But a popular alternative to normal-based regression analysis is to fit the data using a t-density with an unknown degrees of freedom parameter, ν. The t_ν density has heavier tails than the normal but tends to the normal as $\nu \to \infty$. Thus the t_ν density is an ideal candidate for fitting data suspected to be heavier-tailed (i.e., data with outliers in the dependent variable), with the unknown parameter ν providing a way to test for the appropriateness of the normal density. Moreover, by embedding a model with location parameters in the t_ν density, we obtain outlier-resistant estimates of location parameters when ν is small.

[9] Western (1995) provides a recent review, geared for a political science audience.

Formally, if we have a t_ν regression model for the y_i, $E(y_i) = \mathbf{x}_i\boldsymbol{\beta}$, $i = 1, \ldots, n$, the ith observation makes the likelihood contribution

$$f(y_i \mid \mathbf{x}_i, \boldsymbol{\beta}, \sigma^2, \nu) = \frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)\sqrt{\pi\,\nu}\,\sigma^2}\left[1 + \frac{(y_i - \mathbf{x}_i\boldsymbol{\beta})^2}{\nu\sigma^2}\right]^{-(\nu+1)/2} \tag{7}$$

where $\Gamma(\cdot)$ is the gamma function (e.g., Geweke 1993). Joint estimation of $\boldsymbol{\beta}$, σ^2, and ν is not difficult via direct optimization of the log-likelihood function, with the constraint $\nu > 2$.[10] In many applications, ν is preset to small integer values, which amounts simply to replacing one strong assertion (normality) with another. Despite the ease with which we might estimate the t_ν regression model, I am aware of only one political science application where ν is estimated (Katz and King 1999).

An interesting issue when working with this model is characterizing uncertainty in the degrees of freedom parameter, ν. Even with a moderately sized data set (say, several hundred observations), there is typically not a lot of information with which to estimate ν along with location and scale parameters. In addition, the posterior distribution of ν is almost surely not normal in a small data set and is generally skewed right. This is consequential, since as ν shifts to the right (gets larger) the t_ν model reduces to a normal model and, for $\nu > 30$, is essentially indistinguishable from a normal model. We also want to be sure that uncertainty in the ν fully propagates into inferences for $\boldsymbol{\beta}$. And in addition to making inferences about $\boldsymbol{\beta}$ resistant to outliers, we might also want authoritatively to identify outlying cases in the data.

Bayesian simulation is well posed to handle all these goals simultaneously. I begin by exploiting an interesting property of the t distribution: the t distribution is actually a *scale mixture* of normal distributions. That is, if $y_i \overset{\text{iid}}{\sim} t_\nu(\mathbf{x}_i\boldsymbol{\beta}, \sigma^2)$, then $y_i \mid \lambda_i \sim N(\mathbf{x}_i\boldsymbol{\beta}, \sigma^2/\lambda_i)$ and $\lambda_i \mid \nu \sim \text{Gamma}(\nu/2, 2/\nu)$, where λ_i is "missing data," indicating how any particular y_i is more or less dispersed relative to the normal. Outlying data points will have relatively small λ_i, effectively down-weighting their influence on the location estimates. One could consider the λ_i as auxiliary quantities indicating the extent to which observation i is an outlier. In short, outlier-resistant regression has been reduced to yet another missing data problem, well suited for Bayesian simulation.

6.1 *Data and Model: Incumbency Advantage in Congressional Elections*

The application involves estimating incumbency advantage in American congressional elections, with a data set of 5090 observations from 20 elections (1956–1994). The dependent variable is the proportion of the two-party vote won by the Democratic candidate in district i at election t (uncontested districts are dropped from the analysis). The following regression structure is used to model these data, based loosely on work by Gelman and King (1990, Eq. 6):

$$V_{it} = \alpha_t + \beta_1 V_{i,t-1} + \beta_2 \text{PWP}_{it} + \beta_3 \text{DemInc}_{it} + \beta_4 \text{RepInc}_{it} + \epsilon_{it} \tag{8}$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{20})'$ are election-specific fixed effects; $V_{i,t-1}$ is the Democratic vote share in district i at the previous House of Representatives election (dropping cases where a

[10] The variance of the t distribution is undefined for $\nu \leq 2$ and the Cauchy distribution (with no moments defined) results when $\nu = 1$.

redistricting intervenes); PWP_{it} is "previous winning party," coded -1 for Republicans and 1 for Democrats; and DemInc and RepInc are dummies indicating the party affiliation of an incumbent (if present). A detailed justification for this specification is given by Gelman and King (1990).

The quantity $\delta = \beta_3 + \beta_4$ is *asymmetry* in incumbency advantage, since the dependent variable is Democratic vote share and we expect $\beta_3 > 0$ and $\beta_4 < 0$. Finding $\delta > 0$ would imply that Democrats enjoy a larger incumbency advantage than Republican incumbents. Gelman and King (1990) operationalized incumbency status with a single variable coded -1 for Republican incumbents, 0 for open seats, and 1 for Democratic incumbents, effectively constraining $\delta = 0$.

6.2 *Results*

Equation (8) is first estimated via ordinary least squares, or, equivalently, via maximum likelihood assuming $\epsilon_{it} \sim N(0, \sigma^2), \forall\, i, t$. Results appear in the second column in Table 2. Republican incumbents enjoy an incumbency advantage 1.31 percentage points greater than that for Democratic incumbents. This difference is statistically significant, suggesting that the restriction $\delta = 0$ is unsupported by the data. But it is plausible that these vote data are heavy-tailed, even conditional on the regression specification presented above: strategic entry decisions by challengers and risk aversion by well-resourced incumbents combine to generate election outcomes that are lopsided, or the phenomenon of "vanishing

Table 2 Regression analysis, congressional elections data[a]

	Normal	t_ν
Previous vote	0.677	0.726
	[0.655, 0.698]	[0.705, 0.748]
Previous winning party	−2.77	−3.35
	[−3.39, −2.16]	[−3.99, −2.74]
Democratic incumbent	7.35	7.96
	[6.46, 8.24]	[7.02, 8.94]
Republican incumbent	−8.65	−8.00
	[−9.49, −7.82]	[−8.83, −7.17]
δ	−1.31	−.04
	[−2.59, −.028]	[−1.36, 1.28]
σ	6.98	5.81
	[6.85, 7.12]	[5.61, 6.00]
ν	∞	6.36
		[5.36, 7.51]
log-likelihood	−17,103.20	−16,994.85

[a]Ninety-five percent confidence bounds are reported in brackets. The estimates of the t_ν model are generated by Bayesian simulation; the point estimates are the median of 80,000 Gibbs samples, and the 95% confidence intervals are the 2.5 and 97.5 quantiles. Coefficients on 20 mutually exclusive and exhaustive year dummies are not reported. The quantity δ is the difference between the party-specific estimates of incumbency advantage. The estimates of the scale parameter σ are not comparable between the normal and the t_ν models. The reported log-likelihoods are calculated by setting the parameters equal to their joint posterior mode; with flat priors, these correspond to the maximum values of the respective log-likelihoods.

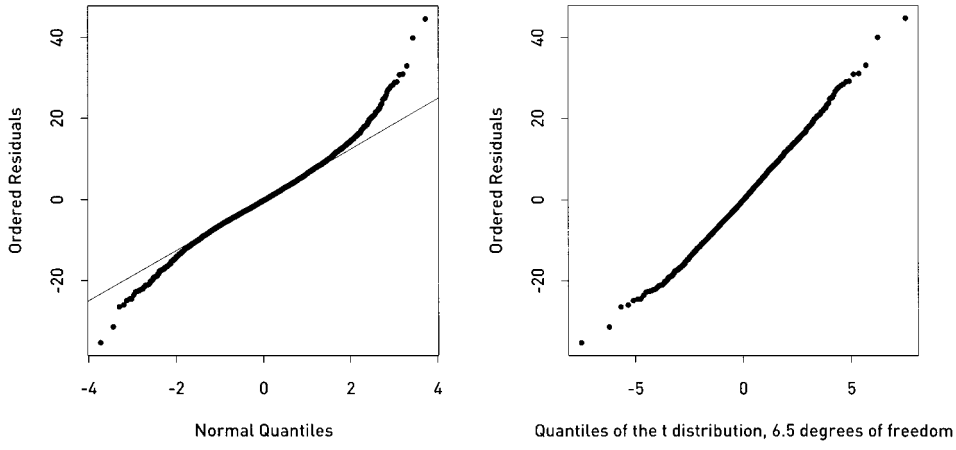

Fig. 7 Quantile–quantile plots, residuals from regression analysis of congressional elections data, normal and $t_{6.5}$ distribution. Curvature in the plotted points indicates departure from the indicated distribution; if the residuals were normal, they would all lie close to the straight line in the left panel.

marginal seats." Note that measures of these aspects of congressional races such as campaign expenditures or challenger quality are absent from the adopted specification. Statistically, this may generate residuals that are overdispersed relative to the normal. That is, estimates of incumbency advantage may be dominated by outliers, and the other regression parameters may be similarly afflicted.

As a preliminary step, I employ some traditional diagnostic tests for nonnormality. The residuals from the normal regression model overwhelmingly fail the Jarque–Bera, Shapiro–Wilk, and Kolmogorov–Smirnov tests of normality. Graphical inspection of the residuals also strongly suggest that their distribution has heavier tails than the normal. The quantile–quantile plot in Fig. 7 strongly suggests that the data are considerably more heavy-tailed than the normal and that a heavier-tailed density is a better characterization. The graph also makes clear the numerous outliers in the data.

I estimated the t_ν regression model via Bayesian simulation, using a Gibbs sampler described in the on-line Appendix. Two samplers were run for 50,000 iterations. Each sample was started from two quite different starting points, with the last 40,000 samples of each chain retained for inference and communication. Estimates of the t_ν model appear in the last column in Table 2. The data strongly support a model based on the t distribution, with the posterior mode of the degrees of freedom parameter approximately 6.5. This is a reasonably heavy-tailed distribution, confirming that the normal is almost certainly the wrong distribution model for these data. This finding is itself novel and echoes Katz and King's (1999) work in multiparty contexts, in which they found a multivariate t density to be an appropriate model for log-odds ratios of vote shares. Estimates of the regression parameters change appreciably from the normal model. Since one feature of the t_ν model is to down-weight outliers for the normal model, past vote shares become a stronger predictor of current vote. The incumbency advantage estimates also change noticeably: while the normal-based analysis suggests that Republican incumbency is worth more than Democratic incumbency, analysis based on a more appropriate set of distributional assumptions shows that this is not the case. The t_ν estimates of incumbency advantage are symmetric with respect to party and worth 8 percentage points of the vote share. Researchers modeling vote

shares ought to be alert to the possibility of outliers influencing their results, and that the t_ν model is an easily implemented alternative to normal based models.

Finally, Bayesian simulation shows the posterior for ν to be slightly skewed right, revealing just how slowly finite sample distributions can converge on asymptotic normality for certain parameters. Even in this data set of over 5000 observations, asymptotics have not yet "kicked in" for this parameter: a normal distribution is not an appropriate characterization of uncertainty in ν. In smaller data sets this nonnormality can be expected to be more pronounced. But this poses no problem for Bayesian simulation, which produces draws from the exact finite sample distributions of all random quantities.

7 Conclusion

I have presented a number of examples to illustrate the strengths of Bayesian simulation. I relied on two simple principles. First, estimation and inference can often be simplified by recasting these tasks as missing data problems. Second, anything we wish to know about a random variable can be discovered up to any degree of accuracy via random sampling from the density for that variable. The combination of these two principles is the essence of Bayesian simulation.

The range of problems that are amenable to estimation, inference, and communication via Bayesian simulation is astonishingly large. Most of the examples here have a latent variable at work (the GLM for bill cosponsorship, correct responses to the political information items in France, and the legislative ideal point model), making recourse to the "missing data" principle somewhat natural. But the last example—outlier resistant estimates of incumbency advantage—turns out to be a "missing data problem" as well. Many other models can be considered in these terms.

In addition, the simulation-based approach employed here generates exact inferences in every instance. In this way Bayesian simulation might be considered a refinement of MLE, removing the reliance on asymptotically valid normality. While this is true, there is much more to Bayesian simulation. Bayesian simulation lets us estimate models where MLE simply is not feasible. This was demonstrated in the political information and ideal-point examples: Bayesian simulation can provide estimates (and confidence bounds) where maximum likelihood struggles or fails altogether.

To be sure, maximum likelihood provided quantitative political scientists with a "great leap forward," opening up the vast statistical terrain beyond linear regression and model fitting by least squares. But Bayesian simulation lets us explore this new terrain more thoroughly and will allow us to go even farther forward in the years ahead.

References

Albert, James H., and Siddhartha Chib. 1993. "Bayesian Analysis of Binary and Polychotomous Response Data." *Journal of the American Statistical Association* 88:669–679.

Albert, James H., and Siddhartha Chib. 1995. "Bayesian Residual Analysis for Binary Response Regression Models." *Biometrika* 82:747–759.

Alvarez, R. Michael, and John Brehm. 1995. "American Ambivalence Towards Abortion Policy: Devlopment of a Heteroskedastic Probit Model of Competing Values." *American Journal of Political Science* 39:1055–1082.

Bailey, Michael, and Douglas Rivers. 1997. "Ideal Point Estimation: A Survey." Paper presented at the Annual Meetings of the Midwest Political Science Association, Chicago.

Bartels, Larry M. 1993. "Messages Received: The Political Impact of Media Exposure." *American Political Science Review* 87:267–285.

Beck, Nathaniel, and Simon Jackman. 1998. "Beyond Linearity by Default: Generalized Additive Models." *American Journal of Political Science* 42:596–627.

Bock, R. D., and M. Aitken. 1981. “Marginal Maximum Likelihood Estimation of Item Parameters: Application of An EM Algorithm.” *Psychometrika* 46:443–459.

Clinton, Joshua, Simon Jackman, and Douglas Rivers. 2000. “The Statistical Analysis of Legislative Behavior: A Unified Approach.” Paper presented at the Southern California Area Methodology Program, University of California, Santa Barbara, May 12–13, 2000.

Gelman, Andrew, and Gary King. 1990. “Estimating Incumbency Advantage Without Bias.” *American Journal of Political Science* 34:1142–1164.

Geweke, J. 1993. “Bayesian Treatment of the Independent Student-*t* Linear Model.” *Journal of Applied Econometrics* 8:S19–S40.

Herron, Michael C. 1999. “Postestimation Uncertainty in Limited Dependent Variable Models.” *Political Analysis* 8:83–98.

Jackman, Simon. 1994. ”Measuring Electoral Bias: Australia, 1949–1993.” *British Journal of Political Science* 24:319–357.

Jackman, Simon. 2000. “Estimation and Inference via Bayesian Simulation: An Introduction to Markov Chain Monte Carlo.” *American Journal of Political Science* 44:375–404.

Johnson, Valen E., and James H. Albert. 1999. *Ordinal Data Modeling*. New York: Springer-Verlag.

Katz, Jonathan N., and Gary King. 1999. “A Statistical Model for Multiparty Electoral Data.” *American Political Science Review* 93:15–32.

King, Gary. 1989. *Unifying Political Methodology*. New York: Cambridge University Press.

King, Gary, James Honaker, Anne Joesph, and Kenneth Scheve. 1998. “Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation,” Typescript. Cambridge, MA: Department of Government, Harvard University. `http://GKing.Harvard.Edu/preprints.shtml`.

King, Gary, Michael Tomz, and Jason Wittenberg. 2000. “Making the Most of Statistical Analysis: Improving Interpretation and Presentation.” *American Journal of Political Science* 44:341–355.

Krehbiel, Keith. 1995. “Cosponsors and Wafflers from A to Z.” *American Journal of Political Science* 39:906–923.

Ladha, Krishna. 1991. “A Spatial Model of Voting with Perceptual Error.” *Public Choice* 78:43–64.

Londregan, John. 2000. “Estimating Legislator’s Preferred Points.” *Political Analysis* 8:35–56.

Lord, F. M. 1975. *Evaluation with Artificial Data of a Procedure for Estimating Ability and Item-Characteristic Curve Parameters*. Princeton, NJ: Educational Testing Service.

Metropolis, N., and S. Ulam. 1949. “The Monte Carlo Method.” *Journal of the American Statistical Association* 44:335–341.

Mondak, Jeffrey J. 2000. “Reconsidering the Measurement of Political Knowledge.” *Political Analysis* 8:57–82.

Mooney, Christopher Z. 1997. *Monte Carlo Simulation*. Thousand Oaks, CA: Sage.

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

Rubin, Donald B. 1987. *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

Schafer, J. L. 1997. *Analysis of Incomplete Multivariate Data*. London: Chapman and Hall.

Sniderman, Paul M., and Sean M. Theriault. 1999. “The Structure of Political Argument and the Logic of Issue Framing.” Presented at the International Society of Political Psychology, Amsterdam.

Venables, William N., and Brian D. Ripley. 1999. *Modern Applied Statistics with S-PLUS*, 3rd ed. New York: Springer-Verlag.

Western, Bruce. 1995. “Concepts and Suggestions for Robust Regression Analysis.” *American Journal of Political Science* 39:786–817.

Zaller, John R. 1992. *The Nature and Origins of Mass Opinion*. New York: Cambridge University Press.