Parameterization and Bayesian Modeling

Andrew GELMAN

Progress in statistical computation often leads to advances in statistical modeling. For example, it is surprisingly common that an existing model is reparameterized, solely for computational purposes, but then this new configuration motivates a new family of models that is useful in applied statistics. One reason why this phenomenon may not have been noticed in statistics is that reparameterizations do not change the likelihood. In a Bayesian framework, however, a transformation of parameters typically suggests a new family of prior distributions. We discuss examples in censored and truncated data, mixture modeling, multivariate imputation, stochastic processes, and multilevel models.

KEY WORDS: Censored data; Data augmentation; Gibbs sampler; Hierarchical model; Missing-data imputation; Parameter expansion; Prior distribution; Truncated data.

1. INTRODUCTION

Progress in statistical computation often leads to advances in statistical modeling. We explore this idea in the context of data and parameter augmentation—techniques in which latent data or parameters are added to a model. Data and parameter augmentation are methods for reparameterizing a model, not changing its description of data, but allowing computation to proceed more easily, quickly, or reliably. In a Bayesian context, however, these latent data and parameters can often be given substantive interpretations in a way that expands the model's practical utility.

1.1 Data Augmentation and Parameter Expansion in Likelihood and Bayesian Inference

Data augmentation (Tanner and Wong 1987) refers to a family of computational methods that typically add new latent data that are partially identified by the data. By "partially identified," we mean that there is some information about these new variables, but as sample size increases, the amount of information about each variable does not increase. Examples included in this article include censored data, latent mixture indicators, and latent continuous variables for discrete regressions. Data augmentation is designed to allow simulation-based computations to be performed more simply on the larger space of "complete data," by analogy to the workings of the EM algorithm for maximum likelihood (Dempster, Laird, and Rubin 1977).

Parameter expansion (Liu, Rubin, and Wu 1998) typically adds new parameters that are nonidentified—in Bayesian terms, if they have improper prior distributions (as they typically do), then they have improper posterior distributions. An important example is replacing a parameter θ by a product, $\phi\psi$, so that inference can be obtained about the product but not about either individual parameter. Parameter expansion can be viewed as part of a larger perspective on iterative simulation (see van Dyk and Meng 2001; Liu 2003), but our focus here is on its construction of nonidentifiable parameters as a computational aid.

Both data augmentation and parameter expansion are exciting tools for increasing the simplicity and speed of computations. In a likelihood-inference framework, that is all they can be—computational tools. By design, these methods do not change the likelihood; they only change its parameterization. The same goes for simpler computational methods, such as standardization of predictors in regression models and rotations to speed Gibbs samplers (e.g., Hills and Smith 1992; Boscardin 1996; Roberts and Sahu 1997).

From a Bayesian perspective, however, new parameterizations can lead to new prior distributions and thus new models. One way in which this often occurs is if the prior distribution for a parameter is *conditionally conjugate* (i.e., conjugate in the conditional posterior distribution), given the data and all other parameters in the model. In Gibbs sampler computation, conditional conjugacy can allow more efficient computation of posterior moments using "Rao–Blackwellization" (see Gelfand and Smith 1990). This technique is also useful for performing inferences on latent parameters in mixture models, conditional on convergence of the simulations for the hyperparameters. As we discuss in Section 5, parameter expansion leads to new families of conditionally conjugate models.

Once again, there is an analogy to Bayesian inference in simpler settings. For example, in classical regression, applying a linear transformation to regression predictors has no effect on the predictions. But in a Bayesian regression with a hierarchical prior distribution, rescaling and other linear transformations can pull parameters closer together so that shrinkage is more effective, as we discuss in Section 5.1.

1.2 Model Expansion for Substantive or Computational Reasons

The usual reason for expanding a model is for substantive reasons—to better capture an underlying model of interest, to better fit existing data, or both. Interesting statistical issues arise when balancing these goals, and Bayesian inference with proper prior distributions can resolve the potential nonidentifiability problems that can arise.

A Bayesian model can be expanded by adding parameters, or a set of candidate models can be bridged using discrete model averaging or continuous model expansion. Recent treatments of these approaches from a Bayesian perspective have been presented by Madigan and Raftery (1994), Hoeting, Madigan, Raftery, and Volinsky (1999), Draper (1995), and Gelman, Huang, van Dyk, and Boscardin (2003, secs. 6.6 and 6.7). In the context of data fitting, Green and Richardson (1997) showed how Bayesian model mixing can be used to perform the equivalent of nonparametric density estimation and regression.

Another important form of model expansion is for sensitivity to potential nonignorability in data collection (see Rubin 1976; Little and Rubin 1987). The additional parameters in

Andrew Gelman is Professor, Department of Statistics, Columbia University, New York, NY 10027 (E-mail: *gelman@stat.columbia.edu*). The author thanks Xiao-Li Meng and several reviewers for helpful discussions and the U.S. National Science Foundation for financial support.

Journal of the American Statistical Association
June 2004, Vol. 99, No. 466, Review Article
DOI 10.1198/016214504000000458

these models cannot be identified but are varied to explore sensitivity of inferences to assumptions about selection (see Diggle and Kenward 1994; Kenward 1998; Rotnitzky, Robins, and Scharfstein 1998; Troxel, Ma, and Heitjan 2003). Nandaram and Choi (2002) argued that continuous model expansion with an informative prior distribution is appropriate for modeling potential nonignorability in nonresponse, and showed how the variation in nonignorability can be estimated from a hierarchical data structure (see also Little and Gelman 1998).

In this article, we do not further consider these substantive examples of model expansion, but rather discuss several classes of models for which *theoretically* or *computationally* motivated model expansion has unexpectedly led to new insights or new classes of models. All of these examples feature new parameters or model structures that could be considered a purely mathematical constructions but gain new life when given direct interpretations. Where possible, we illustrate with applications from our own research so that we have more certainty in our claims about the original motivation and ultimate uses of the reparameterizations and model expansions.

2. TRUNCATED AND CENSORED DATA

It is well understood that the censored data model is like the truncated data model, but with additional information. With censoring, certain specific measurements are missing. Here we further explore the connection between the two models.

2.1 Truncated and Censored Data Models

We work in the context of a simple example (see Gelman et al. 2003, sec. 7.8). A random sample of N animals are weighed on a digital scale. The scale is accurate, but it does not give a reading for objects that weigh more than 200 pounds. Of the N animals, $n = 91$ are successfully weighed; their weights are $y_1, \ldots, y_{91}$. We assume that the weights of the animals in the population are normally distributed with mean μ and standard deviation σ.

In the "truncated data" scenario, N is unknown and the posterior distribution of the unknown parameters μ and σ of the data is

$$p(\mu, \sigma|y) \propto p(\mu, \sigma)\left[1 - \Phi\left(\frac{\mu - 200}{\sigma}\right)\right]^{-91} \prod_{i=1}^{91} \mathrm{N}(y_i|\mu, \sigma^2). \quad (1)$$

In the "censored data" scenario, N is known and the posterior distribution is

$$p(\mu, \sigma|y, N) \propto p(\mu, \sigma)\left[\Phi\left(\frac{\mu - 200}{\sigma}\right)\right]^{N-91} \prod_{i=1}^{91} \mathrm{N}(y_i|\mu, \sigma^2). \quad (2)$$

By using $p(\mu, \sigma)$ rather than $p(\mu, \sigma|N)$, we are assuming that N provides no direct information about μ and σ.

2.2 Modeling Truncated Data as Censored but With an Unknown Number of Censored Data Points

Now suppose that N is unknown. We can consider two options for modeling the data:

1. Using the truncated-data model (1)
2. Using the censored-data model (2), treating the original sample size N as missing data.

The second option requires a probability distribution for N. The complete posterior distribution of the observed data y and missing data N is then

$$p(\mu, \sigma, N|y) = p(N)p(\mu, \sigma)\binom{N}{91} \times \left(\Phi\left(\frac{\mu - 200}{\sigma}\right)\right)^{N-91} \prod_{i=1}^{91} \mathrm{N}(y_i|\mu, \sigma^2).$$

We can obtain the marginal posterior density of (μ, σ) by summing over N,

$$\begin{aligned} p(\mu, \sigma|y) &\propto \sum_{N=91}^{\infty} p(N)p(\mu, \sigma)\binom{N}{91}\left(\Phi\left(\frac{\mu - 200}{\sigma}\right)\right)^{N-91} \\ &\quad \times \prod_{i=1}^{91} \mathrm{N}(y_i|\mu, \sigma^2) \\ &= p(\mu, \sigma) \prod_{i=1}^{91} \mathrm{N}(y_i|\mu, \sigma^2) \\ &\quad \times \sum_{N=91}^{\infty} p(N)\binom{N}{91}\left(\Phi\left(\frac{\mu - 200}{\sigma}\right)\right)^{N-91}. \quad (3) \end{aligned}$$

It turns out that if $p(N) \propto 1/N$, then the expression inside the summation in (3) has the form of a negative binomial density with $\theta = N - 1$, $\alpha = 91$, and $\frac{1}{\beta+1} = \Phi((\mu - 200)/\sigma))$. The expression inside the summation is proportional to $(1 - \Phi(\frac{\mu-200}{\sigma}))^{-91}$, so that for this particular choice of noninformative prior distribution, the entire expression (3) becomes proportional to the simple truncated-data posterior density (1). Meng and Zaslavsky (2002) discussed other properties of the $1/N$ prior distribution and the negative binomial model.

It seems completely sensible that if we add a parameter N to the model and average it out, then we should return to the original model (1). What is odd, however, is that this works only with one particular prior distribution, $p(N) \propto 1/N$. This seems almost to cast doubt on the original truncated-data model, in that it has this hidden assumption about N.

The truncated-data expression of the censored data model adds generality but introduces a sensitivity to the prior distribution of the new parameter N.

2.3 Connection to the Themes of This Article

Model (1) is the basic posterior distribution for truncated data. Going to the censored-data formulation is a model expansion; N is an additional piece of information that is not needed in the model but allows it to be analyzed using a different method (in this case the censored-data model). In examples more complicated than the normal, this model expansion may

be computationally useful, because it removes the integral in the denominator of the truncated-data likelihood.

However, once we consider the model expansion, it reveals the original truncated-data likelihood as just one possibility in a class of models. Depending on the information available in any particular problem, it could make sense to use different prior distributions for N and thus different truncated-data models. It is hard to return to the original "state of innocence" in which N did not need to be modeled.

A similar modeling issue arises in contingency table models (see, e.g., Fienberg 1977), where the multinomial model, $y_1, \ldots, y_J \sim \text{multinomial}(n; \pi_1, \ldots, \pi_J)$ for counts can be derived from the Poisson model, $y_1, \ldots, y_J \sim \text{Poisson}(\lambda_1, \ldots, \lambda_J)$, conditioning on $n = \sum_j y_j$ from the Poisson$(\sum_j \lambda_j)$ distribution and with $\pi_j = \lambda_j / \sum \lambda$ for each j. However, other models with the same conditional distributions are possible, corresponding to different marginal models for the total n.

3. LATENT VARIABLES

By definition, latent variables are constructions that add unknowns to the model without changing the marginal likelihood. Here we consider two common ways in which latent variables are constructed: (1) discrete labeling of components in a finite mixture model and (2) hypothesizing continuous unobserved data underlying discrete-data models, such as logistic regression. We illustrate with examples from our research in voting and public opinion.

3.1 Latent Discrete Variables in Mixture Models

Figure 1 shows a histogram of the Democratic party's share of the vote in about 400 elections to the U.S. House of Representatives in 1988. In an earlier work (Gelman and King 1990), we modeled similar data from many other elections to study the properties of the electoral system under various assumptions about electoral swings. Traditionally this problem had been studied by fitting a normal distribution to district vote proportions and then shifting this distribution to the left or right to simulate alternative hypothetical election outcomes (see Kendall and Stuart 1950; Gudgin and Taylor 1979). There had been discussion in the political science literature of more general models (e.g., King and Browning 1987), and data such as that shown in Figure 1 persuaded us that mixture distributions might be appropriate. We set up a model for these sorts of election data

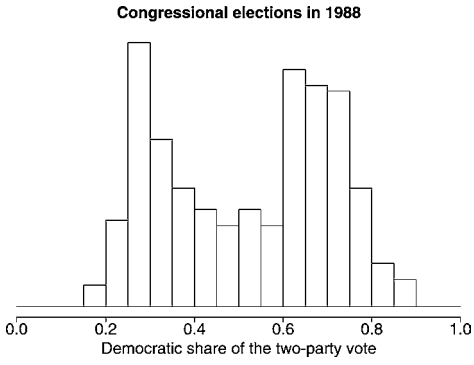

Figure 1. Histogram of Democratic Share of the Two-Party Vote in Congressional Elections in 1988. Only districts that were contested by both major parties are shown here.

using a mixture of three normal distributions, with two large components corresponding to the two modes and one lower and broader component to pick up outlying districts that would not be well captured by the two main components.

For electoral data u_i (defined as the logit of the Democratic share of the two-party vote in district i, excluding uncontested districts), we fit an error model, $u_i \sim \text{N}(\alpha_i, \sigma^2)$, where α_i represented the "usual vote" in the district i (on the logit scale) that would persist in future elections or under counterfactual scenarios. (The variance σ^2 was estimated from variation in district votes between successive elections.) We continued by modeling the usual votes with a mixture distribution,

$$p(\alpha_i) = \sum_{j=1}^{3} \lambda_j \text{N}(\alpha_i | \mu_j, \tau_j^2).$$

It is unstable to estimate mixture parameters by maximum likelihood (see, e.g., Titterington, Smith, and Makov 1985; Gelman 2003), a problem compounded by the fact that the data α_i are observed only indirectly, and so we regularized the estimates of the hyperparameters λ, μ, and σ by assigning informative prior distributions. The prior mode corresponded to a hump at about 40% Democratic vote with standard deviation 10%, another hump symmetrically located at about 60% Democratic vote, and a third hump, centered at 50% and much wider, to catch the outliers that are not close to either major mode. The prior distributions were given enough uncertainty that the variances for the major humps could differ, and the combined distribution could be either unimodal or bimodal, depending on the data. The model was completed with a Dirichlet(19, 19, 4) distribution on λ_1, λ_2, and λ_3, which allowed the major modes to differ a bit in mass but constrained the third mode to its designated minor role of picking up outliers.

We fit the model using data augmentation, following Tanner and Wong (1987). The data augmentation scheme alternately updated the model parameters and the latent mixture indicators for each district. The mixture model fit well and did a good job estimating the distribution of the "usual vote" parameters, α_i, which in turn was useful in answering questions of political interest, such as what proportion of districts we would expect to switch parties if the national vote were to shift by $x\%$.

The next step was to take the mixture indicators seriously. What does it mean that some elections are in the "Republican hump," some are in the "Democratic hump," and others are in the group of "extras"? We realized that there was a key piece of information not included in our model that took on three possible values: *incumbency.* An election could have a Republican incumbent (i.e., the existing Congress member could be a Republican, running for reelection), a Democratic incumbent, or an open seat (no incumbent running).

Conditional on incumbency, the election data were fit reasonably well by unimodal distributions, as shown by Figure 2. The individual distributions are not quite normal, which is perhaps an interesting feature, but it is certainly an improvement to go from a three-component mixture model to a regression-type model conditional on a predictor that takes on three values. The model with incumbency gives more accurate predictions and makes more political sense (see Gelman and King 1994). It came about because we were trying to make sense of a mixture

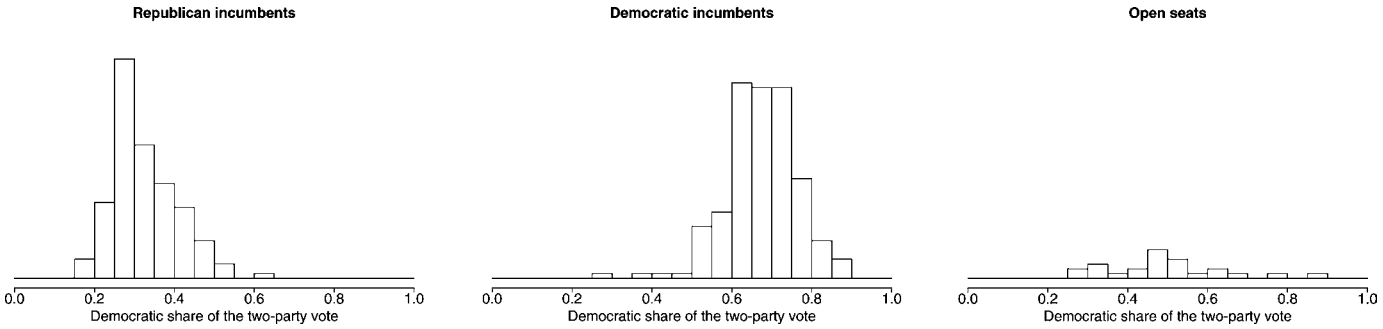

Figure 2. Histogram of Democratic Share of the Two-Party Vote in Congressional Elections in 1988, in Districts With (a) Republican Incumbents, (b) Democratic Incumbents, and (c) Open Seats. Combined, the three distributions yield the bimodal distribution in Figure 1.

model—not to estimate the latent categories, but merely to better fit the distribution of data (as was done using more advanced computational techniques by Green and Richardson 1997).

3.2 Latent Continuous Variables in Logistic Regression

We now consider the opposite sort of model: discrete data usefully expressed in terms of a latent continuous variable. It is well known in statistical computation that logistic or probit regression can be defined as linear regressions with latent continuous variables (see Finney 1947; Ashford and Sowden 1970). More recently, this parameterization has been applied to the Gibbs sampler for binary regression with probit and Student t links (see Albert and Chib 1993; Liu 2004). We give an example to illustrate how the latent variables, which might have been viewed as a computational convenience, can take on a life of their own.

In earlier work (Gelman and King 1993) we described a study of opinion trends in a series of political polls during the 1988 Presidential election campaign. In our analysis we focused on the discrete binary outcome of whether a survey respondent favored George Bush or Michael Dukakis for President, at one point fitting a series of logistic regressions. (The article also briefly considered the "no opinion" responses.)

Although the latent variable formulation was not needed in our modeling, it provides some useful insights. For example, Figure 3 shows some opinion trends among different subgroups of the population of potential voters at three key periods: the Democratic convention, the Republican convention, and the final 40 days of the campaign. The most notable patterns occur during the two conventions; at each time there is a strong swing toward the nominee of the party throwing the convention. The swings occur among almost all subgroups but most strongly among groups of Independents. Similar patterns were found by Hillygus and Jackman (2003) in a study of the 2000 election.

The pattern of Independents showing the greatest swing was first a surprise to us from a political perspective. We had expected that the strongest swings during each convention would be within the party holding the convention—thus greater movements among Democrats in the Democratic convention and Republicans during their convention.

However, from a latent-variable perspective, the patterns make sense. Think of each voter as having a continuous variable that is positive for a Bush supported and negative for a Dukakis supporter, and suppose that campaign events have additive effects on the latent variable. Then Independents will tend to be near 0 on the continuous preference scale (we in fact learn this from the logistic regressions—party identification is a strong predictor of vote preference) and thus will be more likely to be shifted in their preference by an event such as a convention

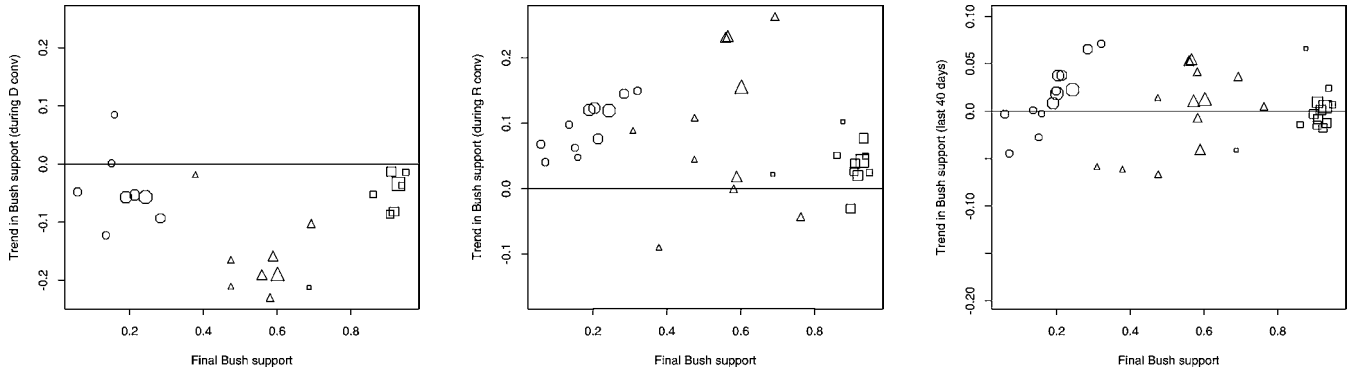

Figure 3. Changes in Public Opinion During Three Periods in the 1988 Presidential Election Campaign: (a) The Democratic Convention, (b) the Republican Convention, and (c) the Final 40 Days of the Campaign. In each graph each symbol represents a different subset of the adult population. The circles represent different subsets of Democrats (e.g., white Democrats, college-educated Democrats, Democrats between 30 and 44), the triangles represent different subsets of Independents, and the squares represent Republicans. Party labels are self-defined from survey responses. In each graph the change in support for Bush during the period is plotted versus the final support for Bush at the time of the election. The different groups tend to move together, but the largest changes tend to be among the Independents.

that is small but in a uniform direction. In contrast, the shifts of the last 40 days were not consistently in favor of a single party, and so groups of Independents were not so likely to show large swings.

A natural model for the preference y_i of an individual voter i at time t is then

$$y_{it} = \begin{cases} \text{Bush supporter} & \text{if } z_{it} > 0 \\ \text{Dukakis supporter} & \text{if } z_{it} \le 0, \end{cases}$$

$$z_{it} = (X\beta)_i + \delta_t + \epsilon_{it},$$

where the individual predictors X contain such information as party identification. The national swings controlled by the continuous time series δ_t will then have the largest effects on individuals i for whom $(X\beta)_i$ is near 0. The model can be further elaborated with individual random effects and time-varying predictors.

We are not claiming here that latent variables are *necessary* for quantitative or political understanding of these opinion trends. Rather, we are pointing out that once the latent variables have been set up, they take on direct political interpretations far beyond what might have been expected had they been treated only as computational devices. In this case the connection could be made more completely by including survey responses on "feeling thermometers"—that is, questions such as, "Rate George Bush on a 1–10 scale, with 1 being most negative and 10 being most positive."

4. MULTIVARIATE MISSING–DATA IMPUTATION

4.1 Iterative Univariate Imputations

Van Buuren, Boshuizen, and Knook (1999) and Raghunathan, Lepkowski, Solenberger, and Van Hoewyk (2001) have formalized iterative algorithms for imputing multivariate missing data using the following method. The algorithm starts by imputing simple guessed values for all of the missing data. Then each variable, one at a time, is modeled by a regression conditional on all of the other variables in the model. The regression is fit and used to impute random values from the predictive distribution for all of the missing data for that particular variable. The algorithm then cycles through the variables and updates them iteratively in Gibbs sampler fashion until approximate convergence.

This "ordered pseudo-Gibbs sampler" (Heckerman, Chickering, Meek, Rounthwaite, and Kadie 2001) is a Markov chain algorithm, and if the values of the parameters are recorded at the end of each loop of iterations, then they will converge to a distribution. This distribution may be inconsistent with various of the conditional distributions used in its implementation, but in general there will be convergence to *something* (assuming that the usual conditions hold for convergence of a Markov chain to a unique nondegenerate stationary distribution).

For the special case in which each regression model includes all of the other variables with no interactions and is linear with normal errors, then this *is* a Gibbs sampler, in which the missing data are imputed based on an underlying multivariate normal model that is encoded as a set of interlocking linear regressions.

In fact, the iterative imputation approach is more general, first because it allows nonnormal models (e.g., logistic regression for binary outcomes and truncated distributions for bounded variables) and also because it allows different sets of predictors for different regression models. These two features allow imputation models to be more realistic for a wider class of data problems, and they have been implemented in S (Van Buuren and Oudshoom 2000) and SAS (Raghunathan, Solenberger, and Van Hoewyk 2002) and used in applications. In comparison, it would be much more difficult to set up, and compute inferences from, a fully multivariate probability model for fitting discrete, continuous, and bounded variables.

4.2 Inconsistent Conditional Distributions

But there is a potential problem with this computationally convenient imputation algorithm: It is so flexible in its specifications that its interlocking conditional distributions will not in general be compatible. That is, there is no implicit joint distribution underlying the imputation models.

This flexibility leading to incompatibility (Arnold, Castillo, and Sarabia 1999) is a theoretical flaw, but in earlier work (Gelman and Raghunathan 2001) we argued that in practice, it can be useful in modeling data structures that could not be easily fit by more standard models. Separate regressions often make more sense than joint models that assume normality and hope for the best (Gelman, King, and Liu 1998), mix normality with completely unstructured discrete distributions (Schafer 1997), mix normality (with random effects) and log-linear structures for discrete distributions (Raghunathan and Grizzle 1995), or generalize with the t distribution (Liu 1995). One may argue that having a joint distribution in the imputation is less important than incorporating information from other variables and unique features of the dataset (e.g., zero/nonzero features in income components, bounds, skip patterns, nonlinearity, interactions). Conditional modeling allows enormous flexibility in dealing with practical problems. In applied work, we have never been able to fit the joint models to a real dataset without making drastic simplifications.

Thus we are suggesting the use of a new class of models—inconsistent conditional distributions—that were initially motivated by computational and analytical convenience. However, as in the other examples of this article, once we accept these new models, they take on their own substantive interpretations.

4.3 Example From Survey Imputation

We consider an example in which we found that structural features of the conditional models can affect the distributions of the imputations in ways that are not always obvious. In the New York City Social Indicators Survey (Garfinkel and Meyers 1999), it was necessary to impute missing responses for family income conditional on demographics and information, such as whether or not anyone in the family received government welfare benefits. Conversely, if the "welfare benefits" indicator is missing, then family income is clearly a useful predictor. (We ignore here the complicating factor that the survey asked about several different sources of income, and these questions had different patterns of nonresponse.)

To simplify, suppose that we are imputing a continuous income variable y_1 and a binary indicator y_2 for welfare benefits, conditional on a set X of fully-observed covariates. We can consider two natural approaches. Perhaps the simplest is a direct

model where, for example, $p(y_1|y_2, X)$ is a normal distribution (perhaps a regression model on y_2, X, and the interactions of y_2 and X) and $p(y_2|y_1, X)$ is a logistic regression on y_1, X, and the interactions of y_1 and X. (For simplicity, we ignore the issues of nonnegativity and possible zero values of y_1.)

A more elaborate and perhaps more appealing model uses hidden variables. Let z_2 be a latent continuous variable, defined so that

$$y_2 = \begin{cases} 1 & \text{if } z_2 \geq 0 \\ 0 & \text{if } z_2 < 0. \end{cases} \tag{4}$$

We can then model $p(y_1, z_2|X)$ as a joint normal distribution (i.e., a multivariate regression). Compared with the direct model, this latent-variable approach has the advantage of a consistent joint distribution. And once inference for (y_1, z_2) has been obtained, we can directly infer about y_2 using (4). In addition, this model has the conceptual appeal that z_2 can be interpreted as some sort of continuous "proclivity" for welfare that is activated only if it exceeds a certain threshold. In fact, the relationship between z_2 and y_2 can be made stochastic if such a model would appear more realistic.

So the latent-variable model is better (except for possible computational difficulties), right? Not necessarily. A perhaps-disagreeable byproduct of the latent model is that, because of the joint normality, the distributions of income among the welfare and nonwelfare groups—that is, the distributions $p(y_1|y_2 = 1, X)$ and $p(y_1|y_2 = 0, X)$—must substantially overlap. In contrast, the direct model allows the overlap to be large or small, depending on the data. Thus the class of inconsistent models, introduced for computational reasons, is more general than previously considered models in ways that are relevant for modeling multivariate data.

4.4 Potential Consequences of Incompatibility

In the multivariate missing-data problem, we are going beyond Bayesian analysis to computation from an inconsistent model—but to the extent that inferences are being summarized by simulations, these can be considered approximate Bayes. The non-Bayesianity should show up as incoherence of inferences, which suggests that estimates of the extent of the problem and approaches to correcting it could be obtained by *data partitioning*—that is, combining analyses from different subsets of the data—another statistical method that has been motivated by computational reasons (see Chopin 2002; Ridgeway and Madigan 2003; Gelman and Huang 2003).

5. MULTILEVEL REGRESSION MODELS

5.1 Rescaling Regression Predictors

When linear transformations are applied to predictors in a classical regression or generalized linear model, it is for reasons of computational efficiency or interpretive convenience. Two familiar examples are the expression of indicators in an ANOVA setting in terms of orthogonal contrasts and standardization of continuous predictors to have mean 0 and variance 1. These and other linear transformations can make regression coefficients easier to understand, but they have no effect on the predictive inferences from the model. From an inferential standpoint, linear transformations do nothing in classical regression, which is one reason why statistical theory has focused on nonlinear transformations (see Atkinson 1985; Carroll and Ruppert 1988). In classical inference, reparameterizations such as centering affect the likelihood for the model parameters but do not affect predictive inference.

In contrast, even simple linear transformations can substantively change inferences in hierarchical Bayesian regressions. For a simple example, consider a regression predicting students' grades from a set of pretests, in which the coefficients β_j for the pretests j are given a common $N(\mu, \tau^2)$ prior distribution, where μ and τ are hyperparameters estimated from the data. Further suppose that the pretests were originally on several different scales and that each pretest has been rescaled to the range 0–100. If the tests measure similar abilities, then it would make sense for the coefficients for the rescaled test scores to be similar. The Bayesian model will then perform more shrinkage on the new scale and lead to more accurate predictions compared with the untransformed model. This is an example of a reparameterization that has no effect on the likelihood but is potentially important in Bayesian inference.

5.2 Parameter Expansion

A more elaborate example of linear reparameterization for hierarchical regression is the parameter-expanded EM algorithm proposed by Liu et al. (1998) that has since been adapted to Gibbs sampler and Metropolis algorithms (Liu and Wu 1999; van Dyk and Meng 2001; Liu 2003; Gelman et al. 2003). We briefly review the PX-Gibbs model here and then describe how it has motivated model expansions.

5.2.1 Hierarchical Model With Potentially Slow Convergence. The starting point is the hierarchical model expressed as a regression with coefficients in M exchangeable batches,

$$y = \sum_{m=1}^{M} X^{(m)} \beta^{(m)} + \text{error},$$

where $X^{(m)}$ is the mth submatrix of predictors and $\beta^{(m)}$ is the mth subvector of regression coefficients. The J_m coefficients in each subvector $\beta^{(m)}$ have exchangeable prior distributions,

$$\beta_j^{(m)} \sim N(0, \sigma_m^2), \quad \text{for } j = 1, \ldots, J_m.$$

More generally, the distributions could have nonzero means; we use the simpler form here for ease in exploring the key issues. Even so, this model is quite general and includes models with several batches of random effects—for example, a model for replicated two-way data can have row effects, column effects, and two-way interactions. In a mixed-effects model, the coefficients for the fixed effects can be collected into a batch with standard deviation σ_m set to infinity. For the random effects, the σ_m parameters will be estimated from data.

For this class of hierarchical models, Gibbs sampler calculations can get stuck, as follows. If a particular standard deviation parameter σ_m happens to be started near 0, then in the updating stage, the batch of coefficients in $\beta^{(m)}$ will be shrunk toward 0. Then at the next update, σ_m will be estimated near 0, and so on. Eventually, the simulation will spiral out of this trap, but under some conditions this can take a long time (Gelman et al. 2003).

5.2.2 Parameter Expansion to Speed Computation. The slow convergence can be fixed using the following parameter expansion scheme, adapted from a similar algorithm for the EM algorithm given by Liu et al. (1998). In the parameter-expanded model, each component m of the regression model is multiplied by a new parameter, α_m,

$$y = \sum_{m=1}^{M} \alpha_m X^{(m)} \beta^{(m)} + \text{error}, \tag{5}$$

and then the parameters α_m are given uniform prior distributions on $(-\infty, \infty)$. This new model is equivalent to the original model but with a new parameterization: $|\alpha_m|\sigma_m$ in the new model maps to σ_m in the original formulation, and each new $\alpha_m \beta_j^{(m)}$ maps to the old $\beta_j^{(m)}$. The parameters α_m and σ_m are not jointly identified, but we can simply run the Gibbs sampler on the expanded set of parameters (α, β, σ) and transform back to the original scale to get inferences under the original model.

In the Gibbs sampler for the new model, the components of β and σ^2 are updated as before—normal and inverse chi-squared distributions—and the new parameter vector α is updated using a normal distribution. This extra step stops the algorithm from getting stuck near 0 (Gelman, Huang, van Dyk, and Boscardin 2004; Liu and Wu 1999).

5.2.3 New Model Families Implied by Parameter Expansion. By giving the new parameters α_m uniform prior distributions, we can perform more efficient computations for the hierarchical model. The parameter α has no meaning and is used just as a multiplier to create the parameters β and σ from the original model.

But now suppose that we take the parameters α_m seriously, giving them informative prior distributions. What does that get us?

First, the variance parameter σ_m^2 in the old model has been split into $\alpha_m^2 \sigma_m^2$ in the new model, which expands the family of conditionally conjugate distributions from inverse chi-squared to products of inverse chi-squared with squared normal distributions.

Second, and potentially more important, we can interpret the parameters α_m themselves in the context of the multiplicative model (5). For each m, $X^{(m)}\beta^{(m)}$ represents a "factor" formed by a linear combination of the J_m individual predictors, and α_m represents the importance of that factor.

For the factor-analytic interpretation of the model to be useful, we need informative prior distributions on the parameters α or β. Otherwise, the model is nonidentified, and we cannot distinguish between the importance α_m of a factor and the coefficients $\beta_j^{(m)}$ of its individual components. This indeterminacy can be resolved by, for example, constraining $\sum_j \beta_j^{(m)} = 1$ for each m, thus giving each factor the form of a weighted average, or by setting a "soft constraint" in the form of a proper prior distribution on the multiplicative parameters α_m.

We illustrate with the problem of forecasting presidential elections by state. A forecasting model based on 12 recent national elections would have 600 "data points"—state-level elections—and could then potentially include many state-level predictors measuring such factors as economic performance, incumbency, and popularity (see, e.g., Rosenstone 1983; Campbell 1992). However, at the national level there are really only 12 observations, and so one must be parsimonious with national-level predictors. In practice, this means performing some previous data analysis to pick a single economic predictor, a single popularity predictor, and maybe one or two other predictors based on incumbency and political ideology (Gelman and King 1993; Campbell, Ali, and Jalalzai 2003).

A more general approach to including national predictors would use the parameter-expanded model. For example, suppose that we wish to include five measures of the national economy (e.g., change in per-capita GDP, change in unemployment). We could first standardize them (see Sec. 5.1) and then include them in the model exchangeably in a batch m, as $\beta_1^{(m)}, \ldots, \beta_5^{(m)}$, each with a $N(\frac{1}{5}, \sigma_m^2)$ prior distribution. This prior distribution bridges the gap between the two extremes of simply using the average of the five measures as a predictor (that would be $\sigma_m = 0$), and including them as five independent predictors ($\sigma_m = \infty$). The multiplicative form of model (5) allows us to set up the full model using conditionally-conjugate distributions. For example, we could predict the election outcome in year t in state s within region $r(s)$ as

$$y_{st} = X_{st}^{(0)} \beta^{(0)} + \alpha_1 \sum_{j=1}^{5} X_{jt}^{(1)} \beta_j^{(1)} + \alpha_2 \gamma_t + \alpha_3 \delta_{r(s),t} + \epsilon_{st},$$

where $X^{(0)}$ is the matrix of state × year-level predictors, $X^{(1)}$ is the matrix of year-level predictors, and γ, δ, and ϵ are national, regional, and statewide error terms. In this model the auxiliary parameters α_2 and α_3 exist for purely computational reasons, and they can be given noninformative prior distributions with the understanding that we are interested only in the products $\alpha_2\gamma_t$ and $\alpha_3\delta_{r,t}$. More interestingly, α_1 serves both a computational and modeling rule—with the $\beta_j^{(1)}$ parameters having a common $N(\frac{1}{5}, \sigma_m^2)$ prior distribution, α_1 has the interpretation as the overall coefficient for the economic predictors. The parameters α and β are conditionally conjugate and form a more general model than would be possible using marginally conjugate families with the usual linear model parameterization.

Both the parameter-expanded computations and the factor-analytic model work with generalized linear models as well. They connect somewhat to the recent approach of West (2003) to using linear combinations to include more predictors in a model than data points. The approach of West (2003) is more completely data based, whereas the class of models that we have developed here is structured using prior information (e.g., knowing which predictors correspond to economic conditions, which correspond to presidential popularity, and so forth).

6. ITERATIVE ALGORITHMS AND TIME PROCESSES

Spatial models have motivated developments in iterative simulation algorithms from Metropolis, Rosenbluth, Rosenbluth, Teller, and Teller (1953) to Geman and Geman (1984) and beyond. It has been pointed out that iterative simulation has the role of transforming an intractable spatial process into a tractable space–time process. Such developments as hybrid sampling (Duane, Kennedy, Pendleton, and Roweth 1987; Neal 1994) take this idea even further by connecting the space–time process to a physical model of particles with "momentum" as

well as "position" in simulation space. In a simulation with multiple sequences, algorithms have been developed in which the "particles" from the different sequences can interact.

Perhaps insight can be gained by taking this time-series process seriously. How would this be done? We first must distinguish between two phases of the iterative simulation. First, there is the initial mixing of the sequences, moving from their starting points to the target distribution. Second, the simulations move around within the target distribution.

The first stage could be identified with "learning," or information propagating from the model statement to the inferences. This could be relevant in a setting in which new information is coming in or in which a process is being tracked that is itself changing (Gilks and Berzuini 2001). In that case the current stage of the simulation could represent some current state of knowledge. Another scenario could be a fixed truth, but with actors who are learning about it by gathering information. This might be of interest to economists. For example, if solving a particular regression problem requires 10,000 Gibbs sampler iterations, then perhaps it is a problem that in real time cannot be solved immediately by economic actors. In this sort of application, the Gibbs sampling (or other iterative algorithm) should map, at least approximately, to a real-time learning process.

The second stage of an iterative algorithm—a walk through the target distribution—could have a direct interpretation if *uncertainty* about parameters in a model could be mapped onto *variability* over time. For example, when in studying exposures to radon gas, we found a high level of uncertainty in home radon levels, even given some geographic information (Lin, Gelman, Price, and Krantz 1999). It is also known that radon levels vary over time (both on weekly and annual scales). Perhaps it could be possible to set up an iterative sampling algorithm so that the jumps in updating home radon levels correspond to temporal variation.

7. CONCLUSION

It is increasingly common in statistical modeling to add latent variables and parameters that do not change the likelihood function, but rather improve computational tractability. Sometimes these additional variables are partially estimable (as with censored data, latent mixture components, and latent continuous variables for discrete regression); in other settings they are completely nonidentified (as in parameter expansion for hierarchical linear models). In either case, we have found insights from treating these computational parameters as "real." Taking these parameters seriously can lead to deeper understanding of the original model (as with truncated data) or data (as with the public opinion study) or suggest ways in which the model can be improved by adding information (as with the Congressional elections example). In other settings, improved algorithms can actually lead to new statistical models that can fit data more accurately, as with multivariate missing-data imputation and multilevel regression.

At a theoretical level, reparameterization has a special new role in Bayesian analysis because it tends to lead to new classes of prior distributions (especially true if the distributions are originally motivated by computational concerns) and thus new models, even if the likelihood is unchanged. At a practical level, when using Bugs (Spiegelhalter, Thomas, Best, Gilks, and Lunn 1994, 2003), which requires specifications for all parameters in a model, it is natural to give substantive models and interpretations to the additional random variables that arise with data augmentation and parameter expansion.

[Received September 2002. Revised November 2003.]

REFERENCES

Albert, J. H., and Chib, S. (1993), "Bayesian Analysis of Binary and Polychotomous Response Data," *Journal of the American Statistical Association*, 88, 669–679.

Arnold, B. C., Castillo, E., and Sarabia, J. M. (1999), *Conditional Specification of Statistical Models*, New York: Springer-Verlag.

Ashford, J. R., and Sowden, R. R. (1970), "Multi-Variate Probit Analysis," *Biometrics*, 26, 535–546.

Atkinson, A. C. (1985), *Plots, Transformations, and Regression*, Oxford, U.K.: Oxford University Press.

Boscardin, W. J. (1996), "Bayesian Analysis for Some Hierarchical Linear Models," unpublished doctoral thesis, University of California-Berkeley, Dept. of Statistics.

Campbell, J. E. (1992), "Forecasting the Presidential Vote in the States," *American Journal of Political Science*, 36, 386–407.

Campbell, J. E., Ali, S., and Jalalzai, F. (2003), "Forecasting the Presidential Vote in the States, 1948–2000," technical report, University of Buffalo, Dept. of Political Science.

Carroll, R. J., and Ruppert, D. (1988), *Transformation and Weighting in Regression*, London: Chapman & Hall.

Chopin, N. (2002), "A Sequential Particle Filter Method for Static Models," *Biometrika*, 89, 539–552.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum Likelihood From Incomplete Data via the EM Algorithm" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 39, 1–38.

Diggle, P., and Kenward, M. G. (1994), "Informative Drop-out in Longitudinal Data Analysis," *Journal of the Royal Statistical Society*, Ser. B, 43, 49–73.

Draper, D. (1995), "Assessment and Propagation of Model Uncertainty" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 57, 45–97.

Duane, S., Kennedy, A. D., Pendleton, B. J., and Roweth, D. (1987), "Hybrid Monte Carlo," *Physics Letters*, Ser. B, 195, 216–222.

Fienberg, S. E. (1977), *The Analysis of Cross-Classified Categorical Data*, Cambridge, MA: MIT Press.

Finney, D. J. (1947), "The Estimation From Individual Records of the Relationship Between Dose and Quantal Response," *Biometrika*, 34, 320–334.

Garfinkel, I., and Meyers, M. K. (1999), "A Tale of Many Cities: The New York City Social Indicators Survey," Columbia University, School of Social Work.

Gelfand, A. E., and Smith, A. F. M. (1990), "Sampling-Based Approaches to Calculating Marginal Densities," *Journal of the American Statistical Association*, 85, 398–409.

Gelman, A. (2003), "A Bayesian Formulation of Exploratory Data Analysis and Goodness-of-Fit Testing," *International Statistical Review*, 71, 369–382.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995), *Bayesian Data Analysis*, London: Chapman & Hall.

Gelman, A., and Huang, Z. (2003), "Sampling for Bayesian Computation With Large Datasets," technical report, Columbia University, Dept. of Statistics.

Gelman, A., Huang, Z., van Dyk, D. A., and Boscardin, W. J. (2004), "Transformed and Parameter-Expanded Gibbs Samplers for Hierarchical Linear and Generalized Linear Models," technical report, Department of Statistics, Columbia University.

Gelman, A., and King, G. (1990), "Estimating the Electoral Consequences of Legislative Redistricting," *Journal of the American Statistical Association*, 85, 274–282.

——— (1993), "Why Are American Presidential Election Campaign Polls So Variable When Votes Are So Predictable? "*British Journal of Political Science*, 23, 409–451.

——— (1994), "A Unified Model for Evaluating Electoral Systems and Redistricting Plans," *American Journal of Political Science*, 38, 514–554.

Gelman, A., King, G., and Liu, C. (1998), "Not Asked and Not Answered: Multiple Imputation for Multiple Surveys" (with discussion and rejoinder), *Journal of the American Statistical Association*, 93, 846–874.

Gelman, A., and Raghunathan, T. E. (2001), "Using Conditional Distributions for Missing-Data Imputation," discussion of "Conditionally Specified Distributions," by Arnold et al., *Statistical Science*, 3, 268–269.

Geman, S., and Geman, D. (1984), "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721–741.

Gilks, W. R., and Berzuini, C. (2001), "Following a Moving Target: Monte Carlo Inference for Dynamic Bayesian Models," *Journal of Royal Statistical Society*, Ser. B, 63, 127–146.

Green, P. J., and Richardson, S. (1997), "On Bayesian Analysis of Mixtures With an Unknown Number of Components" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 59, 731–792.

Gudgin, G., and Taylor, P. J. (1979), *Seats, Votes, and the Spatial Organisation of Elections*, London: Pion.

Heckerman, D., Chickering, D. M., Meek, C., Rounthwaite, R., and Kadie, C. (2001), "Dependency Networks for Inference, Collaborative Filtering, and Data Visualization," *Journal of Machine Learning Research*, 1, 49–75.

Hills, S. E., and Smith, A. F. M. (1992), "Parameterization Issues in Bayesian Inference" (with discussion), in *Bayesian Statistics 4*, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, Oxford, U.K.: Oxford University Press, pp. 227–246.

Hillygus, D. S., and Jackman, S. (2003), "Voter Decision Making in Election 2000: Campaign Effects, Partisan Activation, and the Clinton Legacy," *American Journal of Political Science*, 47, 583–596.

Hoeting, J., Madigan, D., Raftery, A. E., and Volinsky, C. (1999), "Bayesian Model Averaging," *Statistical Science*, 14, 382–401.

Kendall, M. G., and Stuart, A. (1950), "The Law of the Cubic Proportion in Election Results," *British Journal of Sociology*, 1, 193–196.

Kenward, M. G. (1998), "Selection Models for Repeated Measurements With Non-Random Dropout: An Illustration of Sensitivity," *Statistics in Medicine*, 17, 2723–2732.

King, G., and Browning, R. X. (1987), "Democratic Representation and Partisan Bias in Congressional Elections," *American Political Science Review*, 81, 1251–1276.

Lin, C. Y., Gelman, A., Price, P. N., and Krantz, D. H. (1999), "Analysis of Local Decisions Using Hierarchical Modeling, Applied to Home Radon Measurement and Remediation" (with discussion), *Statistical Science*, 14, 305–337.

Little, R. J. A., and Rubin, D. B. (1987), *Statistical Analysis With Missing Data*, New York: Wiley.

Little, T. C., and Gelman, A. (1998), "Modeling Differential Nonresponse in Sample Surveys," *Sankhyā*, 60, 101–126.

Liu, C. (1995), "Missing Data Imputation Using the Multivariate t Distribution," *Journal of Multivariate Analysis*, 48, 198–206.

——— (2003), "Alternating Subspace-Spanning Resampling to Accelerate Markov Chain Monte Carlo Simulation," *Journal of the American Statistical Association*, 98, 110–117.

——— (2004), "Robit Regression: A Simple Robust Alternative to Logit and Probit," in *Missing Data and Bayesian Methods in Practice*.

Liu, C., Rubin, D. B., and Wu, Y. N. (1998), "Parameter Expansion to Accelerate EM: The PX–EM Algorithm," *Biometrika*, 85, 755–770.

Liu, J., and Wu, Y. N. (1999), "Parameter Expansion for Data Augmentation," *Journal of the American Statistical Association*, 94, 1264–1274.

Madigan, D., and Raftery, A. E. (1994), "Model Selection and Accounting for Model Uncertainty in Graphical Models Using Occam's Window," *Journal of the American Statistical Association*, 89, 1535–1546.

Meng, X. L., and Zaslavsky, A. M. (2002), "Single Observation Unbiased Priors," *The Annals of Statistics*, 30, 1345–1375.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953), "Equation of State Calculations by Fast Computing Machines," *Journal of Chemical Physics*, 21, 1087–1092.

Nandaram, B., and Choi, J. W. (2002), "Hierarchical Bayesian Nonresponse Models for Binary Data From Small Areas With Uncertainty About Ignorability," *Journal of the American Statistical Association*, 97, 381–388.

Neal, R. M. (1994), "An Improved Acceptance Procedure for the Hybrid Monte Carlo Algorithm," *Journal of Computational Physics*, 111, 194–203.

Raghunathan, T. E., and Grizzle, J. E. (1995), "A Split Questionnaire Survey Design," *Journal of American Statistical Association*, 90, 55–63.

Raghunathan, T. E., Lepkowski, J. E., Solenberger, P. W., and Van Hoewyk, J. H. (2001), "A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models," *Survey Methodology*, 27, 85–95.

Raghunathan, T. E., Solenberger, P. W., and Van Hoewyk, J. (2002), "IVEware (SAS software for missing-data imputation)," available at *www.isr.umich.edu/src/smp/ive/*.

Ridgeway, G., and Madigan, D. (2003), "A Sequential Monte Carlo Method for Bayesian Analysis of Massive Datasets," *Journal of Data Mining and Knowledge Discovery*, 7, 301–319.

Roberts, G. O., and Sahu, S. K. (1997), "Updating Schemes, Correlation Structure, Blocking and Parameterization for the Gibbs Sampler," *Journal of the Royal Statistical Society*, Ser. B, 59, 291–317.

Rosenstone, S. J. (1983), *Forecasting Presidential Elections*, New Haven, CT: Yale University Press.

Rotnitzky, A., Robins, J. M., and Scharfstein, D. O. (1998), "Semiparametric Regression for Repeated Outcomes With Nonignorable Nonresponse," *Journal of American Statistical Association*, 94, 1096–1146.

Rubin, D. B. (1976), "Inference and Missing Data," *Biometrika*, 63, 581–592.

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

Spiegelhalter, D., Thomas, A., Best, N., Gilks, W., and Lunn, D. (1994, 2003), "BUGS: Bayesian Inference Using Gibbs Sampling," MRC Biostatistics Unit, Cambridge, U.K., available at *www.mrc-bsu.cam.ac.uk/bugs/*.

Tanner, M. A., and Wong, W. H. (1987), "The Calculation of Posterior Distributions by Data Augmentation" (with discussion), *Journal of the American Statistical Association*, 82, 528–550.

Titterington, D. M., Smith, A. F. M., and Makov, U. E. (1985), *Statistical Analysis of Finite Mixture Distributions*, New York: Wiley.

Troxel, A., Ma, G., and Heitjan, D. F. (2003), "An Index of Local Sensitivity to Nonignorability," *Statistica Sinica*, to appear.

Van Buuren, S., Boshuizen, H. C., and Knook, D. L. (1999), "Multiple Imputation of Missing Blood Pressure Covariates in Survival Analysis," *Statistics in Medicine*, 18, 681–694.

Van Buuren, S., and Oudshoom, C. G. M. (2000), "MICE: Multivariate Imputation by Chained Equations (S software for missing-data imputation)," available at *web.inter.nl.net/users/S.van.Buuren/mi/*.

van Dyk, D. A., and Meng, X. L. (2001), "The Art of Data Augmentation" (with discussion), *Journal of Computational and Graphical Statistics*, 10, 1–111.

West, M. (2003), "Bayesian Factor Regression Models in the 'Large p, Small n' Paradigm," in *Bayesian Statistics 7*, eds. S. Bayarri, J. O. Berger, J. M. Bernardo, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West, Oxford, U.K.: Oxford University Press.