# Chapter 11
# Bayesian Inference in Political Science, Finance, and Marketing Research

Many current research challenges in Bayesian analysis arise in applications. A beauty of the Bayesian approach is that it facilitates principled inference in essentially any well-specified probability model or decision problem. In principle one could consider arbitrarily complicated priors, probability models and decision problems. However, not even the most creatively convoluted mind could dream up the complexities, wrinkles and complications that arise in actual applications. In this chapter we discuss typical examples of such challenges, ranging from prior constructions in political science applications, to model based data transformation for the display of multivariate marketing data, to challenging posterior simulation for state space models in finance and to expected utility maximization for portfolio selection.

## 11.1 Prior Distributions for Bayesian Data Analysis in Political Science

*Andrew Gelman*

Jim Berger has made important contributions in many areas of Bayesian statistics, most notably on the topics of statistical decision theory and prior distributions. It is the latter subject which I shall discuss here. I will focus on the applied work of my collaborators and myself, not out of any claim for its special importance but because these are the examples with which I am most familiar. A discussion of the role of the prior distribution in several applied examples will perhaps be more interesting than the alternative of surveying the gradual progress of Bayesian inference in political science (or any other specific applied field).

I will go through four examples that illustrate different sorts of prior distributions as well as my own progress—in parallel with the rest of the statistical research

M.-H. Chen et al. (eds.), *Frontiers of Statistical Decision Making and Bayesian Analysis*, DOI 10.1007/978-1-4419-6944-6_11,

community—in developing tools for including prior information in statistical analyses:

- In 1990, we fit a hierarchical model for election outcomes in congressional districts, using a mixture distribution with an informative prior distribution to model districts held by Democrats and Republicans.
- In 1994, we returned to this example, replacing the mixture model with a regression using incumbency as a predictor, with a flat prior distribution on the regression coefficient.
- In 1997, we used a hierarchical model with poststratification to estimate state-level public opinion from national polls. Formally, the model used noninformative prior distributions, but our poststratification actually used lots of external information from the Census.
- In 2008, we used a varying-intercept, varying slope model to explore the relation between income and voting in U.S. states. An attempt to extend this model to include additional predictors revealed the limitations of our default approach of marginal maximum likelihood.

### *11.1.1 Statistics in Political Science*

Is there anything about the study of public opinion and politics (as compared to economics, psychology, sociology, or history, for example) that would show up in the statistical modeling, in particular in prior distributions? I don't think so.

Important statistical issues do arise in particular examples, however. For example, there have not been many national elections, but the fifty states are a natural setting for hierarchical modeling—the states are hardly exchangeable but it can be reasonable to model them with exchangeable errors after controlling for regional indicators and other state-level predictors[1]. Much work in political science goes into increasing the sample size, for example studying other countries (or, within the United States, by studying state and local elections) or replacing binary data with continuous variables. For example, students of the so-called "democratic peace" use continuous measures for democracy and peace, allowing quantitative researchers to examine more sophisticated hypotheses (see Garktze, 2007).

I now return to the statistical specifics of the examples listed above. As we shall see, our models do not show any linear or even monotonic development. Rather,

[1] I used to say that Alabama and Mississippi were exchangeable, along with North and South Dakota, until Brad Carlin—a resident of the neighboring state of Minnesota—explained to me the differences between these two sparsely populated northern states, thus also educated me in the general principle, emphasized by Bayesians from De Finetti to Berger and beyond, that exchangeability is a state of mind as much as it is a description of the physical and social world. To this day I remain blissfully ignorant of any important features distinguishing the two southern states mentioned above. Not so many years ago many would've considered New Hampshire and Vermont to be exchangeable as well, but the expanding Boston suburbs on one side and Ben & Jerry's on the other have made such a model untenable.

we have used more informative prior distributions where needed because of data limitations.

### *11.1.2 Mixture Models and Different Ways of Encoding Prior Information*

Gelman and King (1990) present a model for estimating the so-called seats-votes curve: the expected percentage of seats won by a political party in a legislative election, as a function of the party's share of the national vote. For example, in 2008 the Democrats won 59% of the seats in the U.S. House of Representatives based on an average of 55% of the vote in House elections (after adjusting for uncontested seats; see Kastellec, Gelman, and Chandler, 2009). In 2006 they garnered 53% of the seats based on 52% of the vote. More generally, we can estimate a stochastic seats-votes relation-and thus compute its expectation, the seats-votes curve—by setting up a probability model for the vector of 435 congressional election outcomes.

For a Bayesian such as Jim Berger (or myself), inference under a probability model is conceptually straightforward (even though it might require computational effort and even research). The real challenge is setting up the model.

To use statistical notation, we have districts $i = 1,2,\dots,435$, and in each there is $y_i$, the proportion of votes received in that district by the Democrats in the most recent election (as noted above, our model corrects for uncontested races, a detail which we ignore in our treatment here). We model

$$y_i \sim N(\theta_i, \sigma_y^2),$$

where $\theta_i$ represents the expected level of support for the Democrats in that district and year, with $\sigma_y$ representing variation specific to that election. We estimated $\sigma_y$ by looking at the residual variance predicting an election from the election six years ago, four years ago, and two years ago, and extrapolating this down to predict a hypothetical variance at lag zero.

With one data point $y_i$ for each parameter $\theta_i$, we certainly needed a prior distribution, and what we used was a mixture model with three components: two major modes roughly corresponding to Democratic and Republican-leaning districts, and a third component with a higher variance to capture districts that did not fit in either of the two main modes. This mixture of three normal distributions had eight hyperparameters, which we gave pretty strong prior distributions in order to separately identify the modes from a single election's worth of data.

Much has been written about the difficulty of estimating mixture models and the failure of maximum likelihood or noninformative Bayesian inference in this setting; here, we had to go even further because our mixture components had particular interpretations that we did not want to lose. To be specific, we assigned following informative prior distribution:

- Mixture component 1: mean had a $N(-0.4, 0.4^2)$ prior distribution, standard deviation had an inverse-$\chi^2(4, 0.4^2)$ prior distribution;
- Mixture component 2: mean had a $N(+0.4, 0.4^2)$ prior distribution, standard deviation had an inverse-$\chi^2(4, 0.4^2)$ prior distribution;
- Mixture component 3: mean had a $N(0, 3^2)$ prior distribution, standard deviation had an inverse-$\chi^2(4, 0.8^2)$ prior distribution; and
- The three mixture parameters had a Dirichlet$(19, 19, 4)$ prior distribution.

(The model was on the logit scale, which was why the priors for the modes were centered at −0.4, +0.4 rather than at 0.4, 0.6 as they would have been had the data been untransformed.)

Finally, having performed inference for the model using the Gibbs sampler (or, as we called it in those pre-1990 days, "the data augmentation method of Tanner and Wong (1987)"), we can simulate hypothetical replications of the election under different conditions and then map out a seats-votes curve by allowing different nationwide vote swings.

We followed up this study a few years later (Gelman and King, 1994) with a very similar model differing in only two particulars: First, we set up our model as a regression in which for each data point $y_i$ there could be district-level predictors $x_i$, and as a predictor we took incumbency status: a variable that equaled 1 for Democratic congress members running for reelection, −1 for Republican incumbents, and 0 for "open seats"—those districts with no incumbents running. The model was essentially the same as before, except that the district-level variance represented unexplained variation after accounting for this (and any other) predictors.

The other way in our 1994 model differed from that published four years earlier was that we got rid of the mixture model and its associated informative prior distribution! It turned out that all the information captured therein—and more—was contained in the incumbency predictor. This illustrates the general point that what is important is the information, not whether it is in the form of a "prior distribution" or a "likelihood[2]."

### 11.1.3 Incorporating Extra Information Using Poststratification

In the wake of the successes of hierarchical Bayes for agricultural, social, and educational research (see, for example, Lindley and Smith (1972), and the accompanying references and discussion), survey researchers began using these methods for small-area estimation (Fay and Herriot, 1979).

Gelman and Little (1997) applied these models to the problem of estimating state-level opinions from national surveys, using hierarchical logistic regression to obtain estimates of the average survey response within population subgroups defined by sex, ethnicity (2 categories), age (4 categories), education (4 categories), and

[2] Contrary to what Bayesians sometimes say, however, neither a loss function nor any formal decision analytic framework was needed to set up the model and use it to perform useful inferences.

state (51, including the District of Columbia)—3264 cells in all, and thus certainly a case of small-area estimation—and then summing those estimates over the 64 cells within each state to estimate state-level averages.

Looked at in the traditional Bayesian way, the regression model was innocuous, with predictors including sex×ethnicity, age×education, and state indicators, fitted normal prior distributions for the 16 age×education, and a group-level regression with normal errors for the 51 state predictors. The unmodeled coefficients and hyperparameters were given noninformative uniform prior distributions, and it was easy enough to program a Metropolis algorithm that converged well and yielded simulation-based inference for all the regression parameters, simulations that we directly propagated to obtain inference for the 3264 population cells—a nice trick, given that the procedure performs well even when fit to samples of 1500 or less.

A key place where external information enters into this example, though, is in the next step, in which we construct inferences for the 51 states. The key step is poststratification: summing over the cells in proportion to their population sizes within each state. This step is not particularly Bayesian—given the computations already done, it's nothing more than the computation of 51 weighted averages for each of our posterior simulations—but it does use prior information, in this case the population counts from the Census. The poststratification framework allows us to include external information structurally, as it were, in a way more natural than would be the formal elicitation of a prior distribution.

This multilevel regression and poststratification approach has been useful in other studies of public opinion. For example, Lax and Phillips (2009a) estimate state-level opinion on several gay-rights issues and compare to state policies in this area. Lax and Phillips (2009b) demonstrate that this approach outperforms classical methods while using far smaller samples. The formal prior distribution is not important here, but what is crucial is the use of prior information in the form of state-level predictors (along with the external information from the Census, which is also implicitly used in survey weighting).

### 11.1.4 Prior Distributions for Varying-Intercept, Varying-Slope Multilevel Regressions

A striking feature of the American political map in the twenty-first century is that the Democratic Party does best in the richer states of the northeast and west coast, while the Republicans' strength is in the poorer states in the south and middle of the country—even while the parties retain their traditional economic bases, with Democrats and Republicans continuing to win the votes of poorer and richer voters, respectively. Gelman et al. (2008b) use Bayesian multilevel modeling to explore this juxtaposition, using both individual-level and state-level incomes to predict vote choice in a logistic regression model that includes unexplained variation at both levels. The coefficients for state and individual incomes go in opposite directions, corresponding to rich Democratic states with rich Republican voters within each state.

Our central model included varying intercepts and slopes—that is, the relation between income and voting was allowed to be different in each state—and we blithely fit it using noninformative uniform prior distributions for the hyperparameters, which for this model included the unmodeled regression coefficients, the group-level standard deviation parameters, and the correlation between the errors in the state-level intercepts and slopes. All worked well, and we had the agreeable choice of fitting the full Bayesian model in Bugs (Spiegelhalter et al., 1994, 2002) or running a quick approximate fit using a program in R that computed marginal maximum likelihood estimates (Bates, 2005).

But we ran into trouble when we tried to extend the model by adding religious attendance as a predictor (Gelman et al., 2008a, Chapter 6), thus requiring four varying coefficients per state (income, religious attendance, their interaction, and a constant term). A group-level covariance of dimension 4 × 4 was just too much for a noninformative prior distribution to handle. Bugs simply choked—the program ran extremely slowly and failed to move well through the posterior distribution—and the marginal maximum likelihood estimate moved straight to the boundary of parameter space, yielding an estimated covariance matrix that was not positive definite. These problems arose even with sample sizes in the tens of thousands; apparently, the hyperparameters of even moderately-dimensional hierarchical regression models are not well identified from data.

In our particular example of modeling vote choice given income and religious attendance, we managed to work around the problem by accepting this flawed estimate—our focus here was on the four coefficients for each state rather than on the hyperparameters themselves—but we are convinced that a good general solution to this problem requires an informative prior distribution for the group-level covariance matrix, possibly using the scaled-inverse-Wishart family (O'Malley and Zaslavsky, 2005), whose redundant parameterization allows the user to supply different prior precisions for scale and correlation parameters.

### *11.1.5 Summary*

In conclusion, prior information is often what makes Bayesian inference work. I won't say it's always necessary–noniformative machine learning methods seem to work pretty well in classification problems with huge sample sizes and simple questions–but in the political science examples of which I'm aware, information needs to come in, whether as regression predictors or regularization (that is, prior distributions) on parameters. An important challenge for Jim Berger and his successors in the theory of Bayesian statistics is to study the mapping from prior to posterior in indirect-data settings such as hierarchical models, and thus to figure out which aspects of the prior distribution we need to be particularly careful to specify well. Such theory may indirectly inform our understanding of public opinion, elections, and international relations, by enabling us to study social and political phenomena with ever more realistic (and thus complicated and parameter-laden) models.

*Acknowledgments:* We thank the National Science Foundation, National Institutes of Health, Institute of Educational Sciences, and the Columbia University Applied Statistics Center for partial support of this work.

## 11.2 Bayesian Computation in Finance

*Satadru Hore, Michael Johannes, Hedibert Lopes, Robert E. McCulloch, and Nicholas G. Polson*

Modern-day finance uses arbitrage and equilibrium arguments to derive asset prices as a function of state variables and parameters of the underlying dynamics of the economy. Many applications require extracting information from asset returns and derivative prices such as options or to understand macro-finance models such as consumption-based asset pricing models. To do this the researcher needs to combine information from different sources, asset returns on the one hand and derivative prices on the other. A natural approach to provide inference is Bayesian (Berger, 1985; Bernardo and Smith, 1994; Gamerman and Lopes, 2007).

Our computational challenges arise from the inherent nonlinearities that arise in the pricing equation, in particular through the dependence on parameters. Duffie (1996) and Johannes and Polson (2009) show that empirical asset pricing problems can be viewed as a nonlinear state space models. These so-called affine models provide a natural framework for addressing the problem as well. Whilst affine pricing models in continuous time go a long way to describe the evolution of derivative prices, empirically extracting the latent state variables and parameters that drive prices has up until now received less attention due to computational challenges. In this paper, we address these challenges by using simulation-based methods, such as Markov chain Monte Carlo (MCMC), Forward filtering backward sampling (FFBS) and particle filter (PF). Hence we solve the inverse problem of filtering state variables and estimating parameters given empirical realizations on returns and derivative prices.

The statistical tools that we describe include MCMC methods, with particular emphasis on the FFBS algorithm of Carter and Kohn (1994) and Frühwirth-Schnatter (1994). For sequential methods we describe PF algorithms, with particular emphasis on the sequential importance sampling with resampling (SISR) filter of Gordon, Salmond, and Smith (1993) and the particle learning (PL) algorithm of Lopes et al. (2010). This current research shows how to also estimate parameters such as agents' preferences from empirical data. In many cases the agents will be given the underlying parameters and the problem becomes one of filtering the hidden states as conditioning information arrives.