Bayesian Methods in Election Forecasting

Michael Maltese

Pomona College, Department of Mathematics

April 4, 2014

Contents

1 Introduction

Nate Silver and other statistically-minded political forecasters garnered much acclaim in the 2012 presidential election cycle for their unashamedly mathematical approach to political punditry and for the resulting accuracy of their models. Silver went on to become something of a pop science celebrity, popularizing his idea of looking at the world through a Bayesian lens (Silver, 2012).

In this work we examine different political election forecasting models and the different statistical and mathematical tools behind them.

All models looked at in this work follow a similar strategy of incorporating opinion poll data at various dates before the election, and attempting to track the future possibilities and minimize the variance of their forecasting. In American presidential races, polling data is provided on both a national and state level, and by many different organizations (Gallup, Rasmussen, PPP, etc). Polling data is inherently both inaccurate (a single sample probably will not represent the entire country, and people's opinions can change over a few months) and biased (organizations have political bents).

Some models deal with the biased nature of polling data by simply averaging all available polls under the assumption that biases cancel out (Wang, 2012) and others specifically analyze each organization and assign a measure of bias (Silver, 2012). Strauss (2007) attempts to estimate biases directly from the data during a live campaign, requiring the use of a more computationally intensive model and Markov-Chain Monte Carlo methods.

Polls also suffer from the "early bird" problem: people might report their preferences early in the campaign, and then later change their minds. Models typically handle this by using an estimate of how much information polls carry at different dates in a campaign, and initialize their model using a prior forecast from first principles.

Lock and Gelman (2010), for example, use historical data on the relevance of polls at different dates in the campaign season (obtained with the help of mixed-effects linear regression, a technique which compensates for small factor samples inside a larger overall sample, described further by Gelman 2006) and combines it with a prior forecast based on economic and military events (Hibbs, 2008).

Strauss (2007) and Jackman (2005), on the other hand, develop more complicated models estimating poll variance from theoretical principles as a random walk.

Bayesian statistics, a variant of classical frequentist statistics, provides the theoretical underpinning for many of these models and tools. They allow a model to specify a prior belief, such as macro-level factors in a presidential election (like economic and military events, as interpreted by Hibbs 2008), and then update that belief based on new data, such as opinion polls. The specific technique of Bayesian inference is a way to mathematically solve the problem of weighting between different data based on how much information we think they contain, respectively.

The statistical tools and ideas used here have further applications in the political world. Lock and Gelman (2010) note that "an approach such as described here could be applied to study changes in public opinion and other phenomena with wide national swings and fairly stable spatial distributions relative to the national average," such as the application of these ideas by Lax and Phillips (2009) to opinions and state policies on gay rights.

2 Motivation: the Blotto game

This topic is relevant both to the news media, who stand to make money off of being able to day-after-day produce accurate and/or exciting predictions, and to political campaigns and organizations, who use the information to inform decision-making. The application to news media should be familiar to a reader who has heard of Nate Silver.

In campaign strategy, a campaign can gain an advantage over its opponents with a more accurate picture of the voting landscape. Many election cycles have a number of "battleground" or "swing" states which have no clear winner. Campaigns must strategically allocate resources to these states in order to win the majority of Electoral College votes.

Merolla et al. (2005) compare the election cycle to a Blotto process, where opponents have no pure equilibrium strategy. A Blotto process describes a game similar to the Electoral College, where opponents must allocate resources across multiple battlefields to win a majority of battles. In a symmetric game, where each side has equivalent resources, the best strategy is a mixed strategy. Players do not have a single best option to choose, and must instead choose moves randomly from a probability distribution which maximizes their chances of winning.

We present an example of the game: tomorrow morning, you battle Colonel Blotto. Each of you has 100 soldiers to deploy among three battlefields. Whoever wins the most battles wins the war. We see that if you and Blotto play, respectively,

$$(50, 50, 0) \text{ vs } (33, 33, 34),$$

then you win. However, if Blotto decides to play a different strategy,

$$(50, 50, 0) \text{ vs } (60, 0, 40),$$

then you lose. For every strategy, there is an opposing strategy that beats it.

Forecasting elections plays into this in the forecasting of "safe" and "unsafe" states—imagine playing a game of Blotto, where each battle has a (unknown) handicap towards one side or the other. Ideally, you want to deploy the minimum amount of resources necessary to win to a battle, so you have more resources for other battles. If you have more information than your opponent on what handicaps are where, then you can more effectively allocate your resources.

3 Before opinion polls

3.1 Fundamental determinants of voter preference

The models and techniques examined in this paper rely primarily on polling data for an up-to-date estimate of popular opinion. But even without opinion polling, presidential elections can be predicted ahead of time using various indicators.

For example, Hibbs (2008) presents a Bread-and-Peace model of presidential elections, where he attempts to explain persistent fundamental determinants of election outcomes in a linear regression model. He looks at two variables,

1. the weighted-average growth of per-capita real personal disposable income over the previous term;

2. and U.S. military fatalities in unprovoked, hostile deployments of American forces abroad,

and interpets elections as mostly referendums on the White House party's economic record, along with a substantial aversion to foreign wars.

States popular votes can also be estimated ahead of time, using similar indicators such as economic factors, and further intuitive ones such as general political leaning or home-state advantage (Campbell 1992; Campbell et al. 2006).

The models analyzed in this paper use these estimates as prior knowledge on election outcomes, to be incorporated with polling data.

3.2 State deviance between election cycles

For states, we can note that deviations from the national average are relatively consistent between election cycles Lock and Gelman (2010). This means that a state that leans more Democratic than the nation during one election, will probably lean the same way during the next election.

Lock and Gelman (2010) handle this using a multilevel model, to compensate for the small number of data points for each state (looking at the last eight election cycles gives only eight data points for each state). Multilevel modeling takes advantage of both within-group and between-group information. Groups with poor information will get pulled towards the mean, and groups with rich information will weight their own observations more highly. The technique is also used in education, demographics, and geographical data (Ghitza and Gelman, 2013; Gelman, 2006; Aitkin et al., 1981).

In this case, we want to estimate variance in state deviations from the national vote between election cycles,

$$\frac{1}{N}\sum_{y=1}^{N}(d_{s,y} - d_{s,y-1})^2,$$

where N denotes the total number of election cycles we're looking at and $d_{s,y}$ denotes the deviation from the national vote in election cycle y for state s.

The multilevel model, then, looks like

$$\left[(d_{s,y} - d_{s,y-1})^2\right]_i \sim \mu + \gamma_s + \epsilon_i, \quad \gamma_s \sim \mathrm{N}(0, \sigma_\gamma^2), \quad \epsilon_i \sim \mathrm{N}(0, \sigma_\epsilon^2),$$

where $\mu + \gamma_s$ will give us the estimated variance in deviations for state s. We describe this class of model further in Section 6.

4 Bayesian inference

4.1 Incorporating opinion polls

In the American presidential elections, opinion polls are provided on both the national and state level by an assortment of different organizations. The models in this paper use polling data to estimate voter opinion.

Polls come with a few problems that models have to deal with. Polls themselves are not as accurate as them claim to be. People change their minds over the election cycle, making early polls worse estimates of the election than later ones. Different organizations have different biases and inaccuracies, stemming from either political leanings or different methods of sampling. All organizations have to estimate who will end up voting to weight their polls appropriately.

We first examine the model described by Lock and Gelman (2010), which uses simple Bayesian inference to incorporate polls based on estimated accuracy some months before Election Day. It doesn't handle organizational biases, but it makes up for it with simplicity and the intuitiveness of the math.

After that, we examine the model by Strauss (2007), which uses Gibbs sampling to fit a more involved forecasting model, which estimates organizational biases on the fly.

4.2 Updating beliefs using additional data

To incorporate polling data with predictions from fundamentals, we use the method of Bayesian inference.

Given a prior distribution over some outcome, $\Pr(\theta)$, and data $\boldsymbol{x}$, we use Bayes' Rule

$$\Pr(A|B) = \frac{\Pr(B|A)\Pr(A)}{\Pr(B)},$$

where A and B are events in some probability space, to get our posterior distribution. We say that

$$p(\theta|\boldsymbol{x}) \propto p(\boldsymbol{x}|\theta)p(\theta),$$

where we drop the denominator because it doesn't include θ in any way and thus becomes part of the uniquely-determined normalization constant for the distribution.

Luckily, we don't have to necessarily write out probability density functions and "read off" what family the posterior distribution is. Conjugate families of distributions provide an easy method for calculating posterior distributions based on the parameters of the prior and the data. We show two useful conjugate families, the Normal distribution with unknown mean, and the Normal distribution with unknown precision (the inverse of variance).

Lemma (Normal with unknown mean)**.** *Given a distribution for data* $x \sim \mathrm{N}(\mu, \tau^{-1})$ *and a prior on the mean* $\mu \sim \mathrm{N}(\mu_0, \tau_0^{-1})$*, the posterior distribution is:*

$$\mu|x, \tau \sim \mathrm{N}\left(\frac{x\tau + \mu_0\tau_0}{\tau_0 + \tau},\ \tau_0 + \tau\right).$$

Proof. We see that:

$$\begin{aligned}
\Pr(\mu|x,\tau) &\propto \Pr(x|\mu,\tau)\Pr(\mu) \\
&\propto \exp\left(-\frac{\tau(x-\mu)^2}{2}\right)\exp\left(-\frac{\tau_0(\mu-\mu_0)^2}{2}\right) \\
&= \exp\left[-\frac{1}{2}\left(\tau x^2 - 2x\mu\tau + \tau\mu^2 + \tau_0\mu^2 - 2\mu\mu_0\tau_0 + \tau_0\mu_0^2\right)\right],
\end{aligned}$$

and all terms not involving μ can be dropped, as they'll work out in the normalization constant:

$$\begin{aligned}
&\propto \exp\left[-\frac{1}{2}(\mu(\tau+\tau_0) - 2\mu(x\tau + \mu_0\tau_0)\right] \\
&= \exp\left[-\frac{1}{2}(\tau_0+\tau)\left(\mu - \frac{x\tau+\mu_0\tau_0}{\tau_0+\tau}\right)^2\right].
\end{aligned}$$

□

Lemma (Normal with unknown precision)**.** *Given a a distribution for* n *i.i.d. data points* $x_i \sim \mathrm{N}(\mu, \tau)$*, and a prior on the precision* $\tau \sim \Gamma(\alpha, \beta)$*, the posterior distribution is*

$$\tau|x, \mu \sim \Gamma\left(\alpha + \frac{n}{2},\ \beta + \frac{\sum(x_i+\mu)^2}{2}\right).$$

Proof. We see that,

$$\begin{aligned}
\Pr(\tau|\boldsymbol{x},\mu) &\propto \Pr(\boldsymbol{x}|\tau,\mu)\Pr(\tau) \\
&\propto \prod\left(\tau^{\frac{1}{2}}\exp\left(-\frac{\tau(x_i-\mu)^2}{2}\right)\right)\tau^{\alpha-1}\exp(-\beta\tau) \\
&= \tau^{\alpha+\frac{n}{2}-1}\exp\left(-\tau\left(\beta + \frac{\sum(x_i+\mu)^2}{2}\right)\right).
\end{aligned}$$

□

4.3 Modeling variance in vote intention before an election

In this section we examine the model proposed by Lock and Gelman (2010). The model empirically estimates variance in voter intention at different times before Election Day, and uses these estimates to incorporate polling data into forecasts derived from fundamentals (Hibbs, 2008).

Let α_t denote the true national proportion of people who intend to vote for the Democratic candidate t months before the election, and let $\widehat{\alpha}_t$ denote an estimate of this value from a poll.

The poll variance is the variance of a mean of q Bernoulli samples, or of the ratio of a Binomial sample and q. We see this because, for a Binomial(q, μ) sample, qy,

$$\mathrm{Var}(y) = \frac{1}{q^2} qy(1-y) = \frac{y(1-y)}{q}.$$

Lemma (Law of total expectation)**.** $\mathrm{E}(X) = \mathrm{E}(\mathrm{E}(X|Y))$.

Proof.

$$\begin{aligned}
\mathrm{E}(\mathrm{E}(X|Y)) &= \int \left[\int x p(x|y)\, dx \right] p(y)\, dy \\
&= \iint x p(x,y)\, dx dy \\
&= \int x \int p(x,y)\, dy dx \\
&= \int x p(x)\, dx \\
&= \mathrm{E}(X).
\end{aligned}$$

□

Lemma (Law of total variance, or decomposition of variance)**.** $\mathrm{Var}(Y) = \mathrm{E}(\mathrm{Var}(Y|X)) + \mathrm{Var}(\mathrm{E}(Y|X))$.

Proof.

$$\begin{aligned}
\mathrm{Var}(Y) &= \mathrm{E}(Y^2) - \mathrm{E}(Y)^2 \\
&= \mathrm{E}(\mathrm{E}(Y^2|X)) - \mathrm{E}(\mathrm{E}(Y|X))^2 \\
&= \mathrm{E}\left(\mathrm{Var}(Y|X) + \mathrm{E}(Y|X)^2\right) - \mathrm{E}(\mathrm{E}(Y|X))^2 \\
&= \mathrm{E}(\mathrm{Var}(Y|X)) + \mathrm{Var}(\mathrm{E}(Y|X)).
\end{aligned}$$

□

So, by the laws of total variance and total expectation, we see that

$$\begin{aligned}
\mathrm{Var}(\widehat{\alpha_t}|\alpha_0) &= \mathrm{E}(\mathrm{Var}(\widehat{\alpha_t}|\alpha_t)|\alpha_0) + \mathrm{Var}(\mathrm{E}(\widehat{\alpha_t}|\alpha_t)|\alpha_0) \\
&= \mathrm{E}\left(\frac{\alpha_t(1-\alpha_t)}{n}|\alpha_0\right) + \mathrm{Var}(\alpha_t|\alpha_0) \\
&= \frac{\mathrm{E}(\alpha_t|\alpha_0) - \mathrm{E}(\alpha_t^2|\alpha_0)}{n} + Var(\alpha_t|\alpha_0) \\
&= \frac{\mathrm{E}(\alpha_t|\alpha_0) + \mathrm{E}(\alpha_t|\alpha_0)^2}{n} + \frac{n-1}{n}\mathrm{Var}(\alpha_t|\alpha_0) \\
&\approx \frac{\alpha_0(1+\alpha_0)}{n} + \mathrm{Var}(\alpha_t|\alpha_0).
\end{aligned}$$

We can estimate $\text{Var}(\alpha_t|\alpha_0)$ empirically using polling data and outcomes from past elections, by calculating the empirical variance of polls and subtracting $\alpha_0(1+\alpha_0)/n$. Let $\widehat{\alpha}_{t,i}$ and $n_{t,i}$ denote the estimated vote and sample size for the i-th poll in month t, and N_t the number of polls in month t. Then

$$\widehat{\text{Var}}(\alpha_t|\alpha_0) = \frac{1}{N_t}\sum_{i=1}^{N_t}\left[(\widehat{\alpha}_{t,i}-\alpha_0)^2 - \frac{\alpha_0(1-\alpha_0)}{n_{t,i}}\right].$$

We also want to know $\text{Var}(\widehat{d}_{s,t}|d_{s,0})$, the variance in the relative position t months before the election of state s. By a similar process, we estimate this using data from the last few election cycles,

$$\widehat{\text{Var}}(d_{s,t}|d_{s,0}) = \frac{1}{elections \cdot 50}\sum_{y}^{elections}\sum_{s=1}^{50}\left[\left(\widehat{d}_{s,y,t}-d_{s,y,0}\right)^2 - \frac{\alpha_{s,y,0}(1-\alpha_{s,y,0})}{n_{s,y,t}}\right].$$

Polls do not give us $\widehat{d}_{s,y,t}$, so we estimate it as $\widehat{\alpha}_{s,y,t} - \widehat{\alpha}_{y,t}$.

Now we have our estimates on variance, dependent only on time t.

From before, we have our poll data

$$\widehat{d_{s,t}}|d_{s,0} \sim \text{Normal}\left(d_{s,0}, \frac{\alpha_{s,0}(1-\alpha_{s,0})}{n_{s,t}} + \text{Var}(d_{s,t}|d_{s,0}\right),$$

where we justify normality by the size of the polls. We also have a prior on state deviance from the national outcome, from our multilevel model,

$$d_{s,0}|d_{s,previous} \sim \text{Normal}(d_{s,previous}|\,\text{Var}(d_{s,0}|d_{s,previous})).$$

Because these are both Normal distributions, we end up with a conjugate posterior for state deviation $d_{s,0}|data$.

We have a distribution for poll data,

$$\widehat{\alpha}_t|\alpha_0 \sim \text{Normal}\left(\alpha_0, \frac{\alpha_0(1-\alpha_0)}{n_t} + \text{Var}(\alpha_t|\alpha_0)\right),$$

and some prior estimated from fundamentals

$$\alpha_0 \sim \text{Normal}(\mu_0, \sigma_0^2),$$

so again we end up with a simple conjugate posterior.

Lock and Gelman (2010) complete the model by estimating the distribution of Electoral College outcomes. They simulate 100,000 elections by

1. Randomly drawing a national popular vote from the national posterior distribution;
2. Randomly drawing a deviation for each state, where $\widehat{d}_{s,t} = \widehat{\alpha}_{s,0} - \alpha_0$;
3. And calculating the winner of the Electoral College.

5 Markov-Chain Monte Carlo methods

5.1 Markov chains

Let X_t be the value of a random variable at time t. The random variable is a **Markov process** if the the distribution of X_{t+1} is dependent only on the current state, i.e.,

$$\Pr(X_{t+1}|X_t, \dots, X_0) = \Pr(X_{t+1}|X_t).$$

We call a sequence $(X_0, \dots, X_n)$ of the variables generated by a Markov process, a **Markov chain**.

A chain is defined by its transition probabilities between states. We denote the probability of switching from state i to state j by

$$\Pr(i \to j).$$

Let $\pi_j(t) = \Pr(X_t = j)$, and let $\boldsymbol{\pi}(t)$ be all state space probabilites at time t. Then by the **Chapman-Kolmogorov equation**, we see that $\pi_i(t+1) = \Pr(X_{t+1} = s_i) = \sum_k \Pr(X_{t+1} = s_i|X_t = s_k) \Pr(X_t = s_k) = \sum_k \Pr(k \to i)\pi_k(t)$.

This becomes useful in matrix form. Let $\boldsymbol{P}$ be the probability transition matrix whose (i, j)-th element is $\Pr(i \to j)$. Then

$$\boldsymbol{\pi}(t+1) = \pi(t)\boldsymbol{P},$$

and we see how to obtain subsequent steps of the Markov chain

$$\boldsymbol{\pi}(t) = \boldsymbol{\pi}(t-n)\boldsymbol{P}^n = \boldsymbol{\pi}(0)\boldsymbol{P}^t.$$

We call a Markov chain **irreducible** if all states communicate, i.e., we can go from any state to any other state.

We call a Markov chain **aperiodic** if the number of steps to move between any two states is not required to be a multiple of some integer other than one. In other words, there are no fixed-length cycles between states.

A finite Markov chain that is both irreducible and aperiodic has a stationary distribution.

Example. Suppose $\Pr(rain|rain) = 0.5$, $\Pr(sun|rain) = 0.25$, $\Pr(cloud|rain) = 0.25$, and

$$\boldsymbol{P} = \begin{pmatrix} 0.5 & 0.25 & 0.25 \\ 0.5 & 0 & 0.5 \\ 0.25 & 0.25 & 0.5 \end{pmatrix}.$$

Suppose today is sunny, or $\boldsymbol{\pi}(0) = (0, 1, 0)$. Then

$$\boldsymbol{\pi}(2) = \boldsymbol{\pi}(0)\boldsymbol{P}^2 = (0.375, 0.25, 0.375)$$
$$\boldsymbol{\pi}(7) = (0.4, 0.2, 0.4).$$

Or, suppose today is rainy, $\boldsymbol{\pi}(0) = (1, 0, 0)$. Then

$$\boldsymbol{\pi}(2) = (0.4375, 0.1875, 0.375) \text{ and}$$
$$\boldsymbol{\pi}(7) = (0.4, 0.2, 0.4),$$

which is converging to a stationary distribution.

A sufficient condition to have a stationary distribution is the **detailed balance equation**, or **reversibility condition**,

$$\Pr(j \to k)\pi_j^* = \Pr(k \to j)\pi_k^*,$$

where $\boldsymbol{\pi}^*$ is the stationary distribution.

We extend this to a continuous state space by

$$\int \Pr(x \to y)dy = 1,$$

and the Chapman-Kolmogrov equation becomes

$$\pi_t(y) = \int \pi_{t-1}(x) \Pr(x \to y)dy,$$

or

$$\boldsymbol{\pi}^*(y) = \int \boldsymbol{\pi}^* \Pr(x \to y)dy.$$

5.2 The Metropolis-Hastings algorithm

Suppose we want to sample from the probability distribution

$$p(\theta) = k \cdot f(\theta),$$

where K is some unknown normalizing constant.

We first present the **Metropolis** algorithm. Start with

$$\theta^{(0)} : f(\theta^{(0)}) > 0.$$

Let $q(\theta^{(1)} \to \theta^{(2)})$ be a **jumping distribution** (or **proposal**, or **candidate-generating** distibution) which is symmetric, $q(\theta_1 \to \theta_2) = q(\theta_2 \to \theta_1)$. Then:

1. Sample a candidate point θ^* from q, which denotes the probability of returning θ_2 given a previous value of θ_1;

2. Calculate the ratio of density,

$$\alpha = \frac{p(\theta^*)}{p(\theta^{(t-1)})} = \frac{f(\theta^*)}{f(\theta^{(t-1)})},$$

where the normalizing constant k cancels out;

3. If θ^* increases the density ($\alpha > 1$), then accept the candidate and set $\theta^{(t)} = \theta^*$. Otherwise, then accept the candidate with probability α, or pick a new candidate and try again.

Another way to look at it is, accept a candidate with probability

$$\alpha = \min\left[\frac{f(\theta^*)}{f(\theta^{(t-1)})}, 1\right].$$

The **Metropolis-Hastings** algorithm extends the above to work in the case where $q(\theta_1 \to \theta_2)$ is any distribution, symmetric or not. Let the density ratio be

$$\alpha = \min\left[\frac{f(\theta^*)}{f(\theta^{(t-1)})}\frac{q(\theta^* \to \theta^{(t-1)})}{q(\theta^{(t-1)} \to \theta^*)}, 1\right].$$

In the case where q is symmetric, this becomes simply Metropolis.

Theorem. *The Metropolis-Hastings algorithm generates a Markov chain with stationary distribution p.*

Proof. We show Metropolis-Hastings satisfies the detailed balance equation. Since we sample from $q(\theta_1 \to \theta_2)$, and accept with probability $\alpha(\theta_1, \theta_2)$,

$$\Pr(\theta_1 \to \theta_2) = q(\theta_1 \to \theta_2)\alpha(\theta_1, \theta_2) = q(\theta_1 \to \theta_2)\min\left[\frac{p(\theta_2)}{p(\theta_1)}\frac{q(\theta_2 \to \theta_1)}{q(\theta_1 \to \theta_2)}, 1\right].$$

We want to show that

$$\Pr(\theta_1 \to \theta_2)p(\theta_1) = \Pr(\theta_2 \to \theta_1)p(\theta_2),$$

or

$$q(\theta_1 \to \theta_2)\alpha(\theta_1, \theta_2)p(\theta_1) = q(\theta_2 \to \theta_1)\alpha(\theta_2, \theta_1)p(\theta_2).$$

Assume without loss of generality that $\alpha(\theta_1, \theta_2) < 1$. Then terms cancel and we see the equality. □

Example. Consider the distribution,

$$p(\theta) = C \cdot \theta^{-n/2} \cdot \exp\left(-\frac{a}{2\theta}\right),$$

with $n = 5$ and $a = 4$.

Take a Metropolis candidate distribution, the uniform distribution from 0 to 100. We know p has tails outside this range, but assume that they're negligible.

Now, run the Metropolis algorithm: Let $\theta^{(0)} = 1$, and suppose $\theta^* = 39.82$. Then

$$\alpha = \min\left[\frac{p(\theta^*)}{p(\theta^{(t-1)})}, 1\right] = \min\left[\frac{(\theta^*)^{-5/2}\exp\left(-\frac{2}{\theta^*}\right)}{(\theta^{(t-1)})^{-5/2}\exp\left(-\frac{2}{\theta^{(t-1)}}\right)}, 1\right] = 0.0007,$$

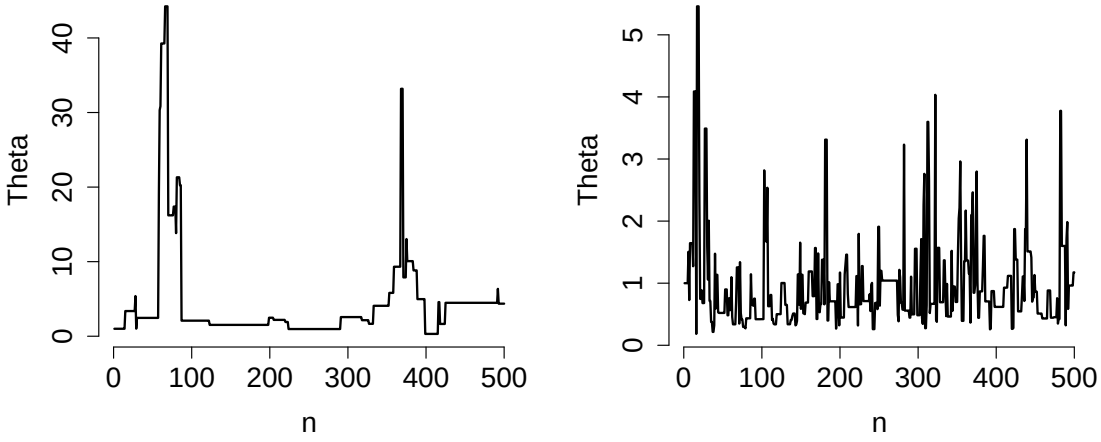

Figure 1: Poorly-mixing and well-mixing Metropolis chains

so we accept θ^* with probability 0.0007.

The first 500 steps of this algorithm are shown in the left side of Figure 1. Note the long flat periods, where all candidate values are rejected. This chain is called **poorly-mixing**.

Suppose that we instead use χ^2_1 as the candidate distribution. This distribution is no longer symmetric, so we must use Metropolis-Hastings instead of just Metropolis. A resulting sample run is shown in the right side of Figure 1. The series of samples looks like **white noise**, and we call this chain **well-mixing**.

5.3 Gibbs sampling

Consider a bivariate random variable (x, y), for which we want to find the joint distribution

$$p(x, y).$$

It's easier to calculate the conditional distributions $p(x|y)$ and $p(y|x)$ than to integrate to get the marginal, $p(x) = \int p(x, y)dy$.

The Gibbs sampling algorithm for two variables is as follows:

1. Start with some value for y, $y^{(0)}$;
2. Sample $x^{(0)} \sim p(x|y = y^{(0)})$;
3. Sample $y^{(1)} \sim p(y|x = x^{(0)})$;
4. Sample $x^{(1)} \sim p(x|y = y^{(1)})$;
5. And so on.

The general multivariate version is analagous.

In the next section we provide a concrete example of finding the conditional distributions for the Gibbs sampler. In general, we find the full joint distribution, which by definition is proportional to the conditional distribution of each variable. Then for each conditional distribution, we simply drop unneeded terms into the normalization constant and hope to get a conjugate posterior.

Theorem. *The Gibbs sampler is a Markov process with stationary distribution the true joint distribution.*

Proof. We show the detailed balance equations hold for sampling x with regards to the joint distribution $p(x, y)$,

$$p(x_2, y)p(x_1|y) = p(x_1, y)p(x_2|y).$$

By the definition of conditional probability, this is

$$p(x_2|y)p(y) \cdot p(x_1|y) = p(x_1|y)p(y) \cdot p(x_2|y),$$

so we're done. □

5.4 Fitting a random walk preference model with house biases

We model popular opinion at some time until Election Day t (where we denote Election Day by $t = 0$). We assume that each time period, popular opnion undergoes some random "shock" with variance ω^2. In other words,

$$\alpha_t \sim \text{Normal}(\alpha_{t-1}, \omega^2).$$

We then model opinion polls y_i as being sampled from popular opinion at that time, with some house bias δ_j, or

$$y_i \sim \text{Normal}(\alpha_t + \delta_j, s_i^2),$$

where s_i^2 is the poll variance $\frac{y_i(1-y_i)}{q_i}$ and q_i is the number of respondents in the poll. We assume that house biases cancel out, i.e., $\sum \delta_j = 0$.

Strauss (2007) models the national preference for a Republican president as a reverse random walk, taking α_0 as the election result. He builds off a model initially proposed by Jackman (2005), who uses a forward random walk to predict Australian parliamentary elections. The difference in nomenclature derives from where each assigns priors: Jackman (2005) assigns a prior at the earliest time step of his model, whereas Strauss (2007) assigns a prior on the final value of his model.

Strauss (2007) estimates priors on the shock as $\omega \sim \text{Uniform}(0, (0.01)^2)$ and on house bias as $\delta_j \sim \text{N}(0, (0.1)^2)$.

We obtain posterior distributions from this model using a Gibbs sampler. We first obtain the full joint posterior distribution for all parameters, given our polling

data. We then condition for each parameter, and find that our conditional posteriors are all simple conjugate distributions.

Given our model above, we see our joint distribution

$$\Pr(\alpha_0, \ldots, \alpha_n, \delta_i, \ldots, \delta_J, \omega | y_1, \ldots, y_P)) = \Pr(\Theta | D),$$

where n is our total number of time periods, J is the number of organizations, and P is the number of polls we have data for. For convenience, we let D denote all data, and Θ denote all parameters. By Bayes' Theorem, this is proportional to

$$\Pr(D|\Theta) \Pr(\Theta).$$

We assume that house biases are independent from popular opinion at each time period, and from the variance of shock to opinion (which may not be true, but is too complicated otherwise). By the definition of conditional probability,

$$\Pr(A, B) = \Pr(A|B) \Pr(B),$$

and our joint distribution becomes

$$\propto \Pr(D|\Theta) \left[\prod \Pr(\delta_j)\right] \Pr(\alpha_0, \ldots, \alpha_n | \omega) \Pr(\omega).$$

Again, by the definition of conditional probability, the distribution becomes

$$\propto \left[\prod_{i=1}^{P} \Pr(y_i|\Theta)\right] \left[\prod_{j=1}^{J} \Pr(\delta_j)\right] \left[\prod_{t=1}^{n} \Pr(\alpha_t | \alpha_{t-1}, \omega)\right] \Pr(\alpha_0) \Pr(\omega).$$

To find the conditional probabilites, we take the conditional as proportional to the joint distribution, and drop terms that don't matter.

For α_0, we see that

$$\Pr(\alpha_0 | \Theta_{-\alpha_0}, D) \propto \Pr(\alpha_0) \Pr(\alpha_1 | \alpha_0, \omega) \prod \Pr(y_i | \alpha_0, \delta_j),$$

for all polls y_i that were conducted at time period α_0. Note that this is simply a conjugate posterior, so we see that $\alpha_0 | \Theta_{-\alpha_0}, D$ has a Normal distribution with variance

$$\left(\frac{1}{\sigma_0^2} + \frac{1}{\omega^2} + \sum \frac{1}{s_i^2}\right)^{-1}$$

and mean

$$\left(\frac{\mu_0}{\sigma_0^2} + \frac{\alpha_1}{\omega^2} + \sum \frac{y_i - \delta_j}{s_i^2}\right) \mathrm{Var}(\alpha_0 | \Theta_{-\alpha_0}, D).$$

We see conditional distributions for the rest of the α_t similarly.

For ω, we see that

$$\Pr(\omega^2 | \Theta_{-\omega}, D) \propto \Pr(\omega) \prod_{t=1}^{n} \Pr(\alpha_t | \alpha_{t-1}, \omega),$$

or, letting $\mathbb{I}_\omega = \Pr(\omega)$,

$$\begin{aligned}
&\propto \mathbb{I}_\omega \cdot \prod_{t=1}^{n} (\omega^2)^{-1/2} \exp\left[-(\omega^2)^{-1} \frac{(\alpha_t - \alpha_{t-1})^2}{2}\right] \\
&= \mathbb{I}_\omega \cdot (\omega^2)^{-n/2} \exp\left[-(\omega^2)^{-1} \frac{\sum_{t=1}^{n} (\alpha_t - \alpha_{t-1})^2}{2}\right],
\end{aligned}$$

which is the probability density function of an Inverse Gamma distribution, restricted to some range. Thus we say

$$\omega^2 \sim \mathbb{I}_\omega \cdot \text{InvGamma}\left(\frac{n-2}{2}, \frac{\sum_{t=1}^{n} (\alpha_t - \alpha_{t-1})^2}{2}\right).$$

When sampling this, we can the assumption that the Inverse Gamma will not have much mass outside of the prior on ω, and can be approximated by dropping all samples outside of $\mathbb{I}_\omega$ and resampling, or we can use another Markov Chain Monte Carlo method such as Metropolis-Hastings or rejection sampling.

Finally, for δ_j, we see that

$$\Pr(\delta_j | \Theta_{-\delta_j}, D) \propto \Pr(\delta_j) \prod \Pr(y_i | \alpha_0, \delta_j)$$

for all polls y_i from the organization indexed by j. Given a prior $\delta_j \sim \text{Normal}(0, d^2)$, we see we have a conjugate posterior, so $\delta_j | \Theta_{-\delta_j}$ has a Normal distribution with variance

$$\left(\frac{1}{d^2} + \sum \frac{1}{s_i^2}\right)^{-1}$$

and mean

$$\left(\sum \frac{y_i - \alpha_t}{s_i^2}\right) \text{Var}(\delta_j | \Theta_{-\delta_j}).$$

We graph the results of this model on simulated election data in Figure 2.

6 Multilevel regression models

6.1 Fitting multilevel models with Gibbs samplers

Gelman and Hill (2007) recommend starting multilevel modeling by using fast computational methods for point estimates of variance parameters, and then moving to fitting fully Bayesian model using Gibbs sampling to obtain point and uncertainty estimates for all parameters in the model. In addition, in some cases, the direct computational methods are unstable in the presence of a small number of groups or a complicated model, and we must use the Bayesian approach. We describe first how to use Gibbs sampling to estimate the parameters of these models, as it's the more

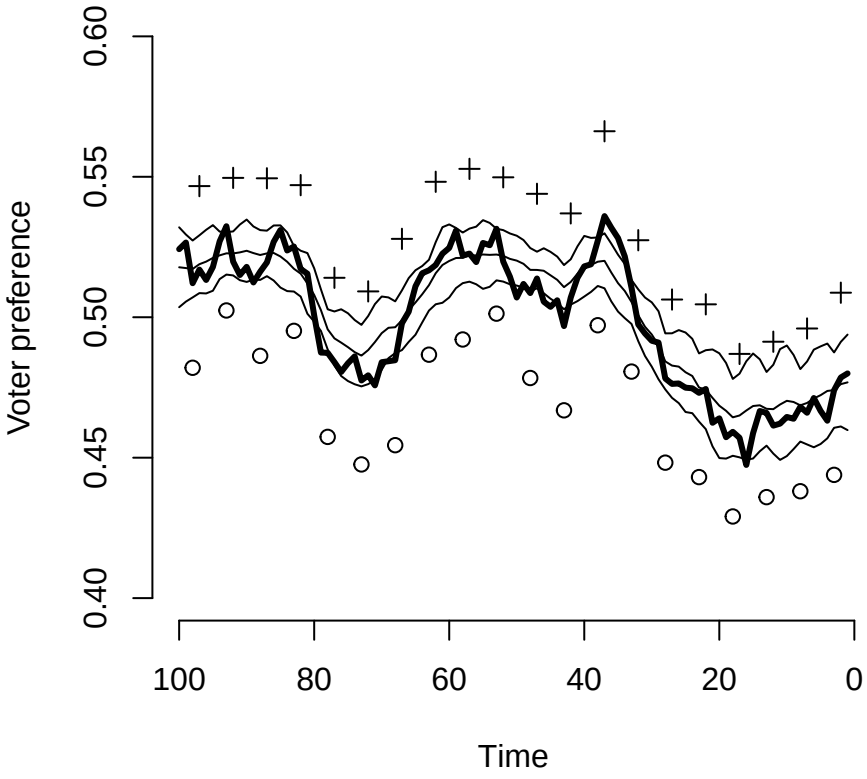

Figure 2: Gibbs sampler run on simulated election model. The bold line in the middle is true voter preference, the symbols above and below it are polls conducted by different organizations, and the thin lines are the point estimate of voter preference and its 90% CIs.

robust technique, and then provide a rough overview of how a computational method for point estimates can work.

We follow the method proposed by Gelman and Hill (2007). Given a model with group-level predictors,

$$y_i = \alpha_{j[i]} + \beta x_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma_y^2),$$
$$\alpha_j = \gamma_0 + \gamma_1 u_j + \eta_j, \quad \eta_j \sim N(0, \sigma_\alpha^2),$$

we assign priors and run a Gibbs sampler. For a group j, the group level variable a_j gets assigned a prior $N(\gamma_0 + \gamma_1 u_j, \sigma_\alpha^2)$. Each α_j differs in the predictor u_j, thus each α_j has a different prior. Hyperparameters are assigned noninformative uniform priors over the full domain, so the posterior distributions will be nicely distributed. Coefficients will have priors Uniform$(-\infty, \infty)$, and standard deviations will have priors Uniform$(0, \infty)$. Gelman and Hill (2007) justify using these noninformative

priors for building models, and suggest moving to the full Bayesian approach and defining informative prior beliefs later in the process.

We also describe a different instance of the model to illustrate a point: given a varying-intercept, varying-slope model,

$$y_i = \alpha_{j[i]} + \beta_{j[i]} x_i + \varepsilon_i,$$

we assign a bivariate normal prior to the (α_j, β_j) pairs to account for correlation of the parameters within each group.

6.2 The Expectation-Maximization algorithm: fitting models with unobserved data

If we only need point estimates, there are faster computational methods available. We describe the Expectation-Maximization algorithm, used for models with missing data or latent variables. This approach nicely fits multilevel modeling.

Given a model $Y \sim (X, Z)$, where we observe Y and X but not Z, we can define an iterative process to find the maximum likelihood estimate for the parameters θ. For a given previous estimate of the parameters, $\theta^{(t)}$, let

$$\theta^{(t+1)} = \arg\max_{\theta} \mathrm{E}_{z|x,\theta^{(t)}} \log L(\theta|x, z).$$

Maximizing the expected log likelihood results in maximizing $L(\theta|x)$, as follows: we write $L(\theta|x) = p(x|\theta)$, so

$$\log p(x|\theta) = \log p(x, z|\theta) - \log p(z|x, \theta).$$

Then, the expected value with respect to the unobserved data z is

$$\begin{aligned} \log p(x|\theta) &= E_{z|x,\theta^{(t)}}(\log p(x, z|\theta)) - E_{z|x,\theta^{(t)}}(\log p(z|x, \theta)) \\ &= Q(\theta|\theta^{(t)}) - H(\theta|\theta^{(t)}), \end{aligned}$$

where Q and H are defined by the terms they replace. The term $H(\theta|\theta^{(t)})$ is maximized by $\theta^{(t)}$, so any value of θ which increases Q also decreases H and thus increases the log likelihood. We show this by showing that $H(\theta^{(t)}|\theta^{(t)}) - H(\theta|\theta^{(t)})$ is always positive:

$$H(\theta^{(t)}|\theta^{(t)}) - H(\theta|\theta^{(t)}) = \mathrm{E}_{Z|X,\theta^{(t)}} \left(-\log p(Z|X, \theta)/p(Z|X, \theta^{(t)}) \right),$$

which by Jensen's inequality is

$$\begin{aligned} &= -\log \mathrm{E}_{Z|X,\theta^{(t)}} \left(p(Z|X, \theta)/p(Z|(X, \theta^{(t)}) \right) \\ &= 0. \end{aligned}$$

Lemma (Jensen's Inequality)**.** *If ϕ is a convex function, then $\phi(\mathrm{E}(x)) \leq \mathrm{E}(\phi(x))$.*

Proof. Let $x_0 = \mathrm{E}_X(g(x))$. Since ϕ is convex, the line tangent at x_0 is always less than or equal to ϕ. In other words, there exist real numbers a and b such that, for all $x \in \mathbb{R}$, $ax + b \leq \phi(x)$, and $ax_0 + b = \phi(x_0)$. Thus, $\mathrm{E}_X\left(\phi(g(x)) \geq \mathrm{E}_X\left(ag(x) + b\right) = a\,\mathrm{E}_X\left(g(x)\right) + b = ax_0 + b = \phi(x_0) = \phi\left(\mathrm{E}_X(g(x))\right)\right.$. □

6.3 Fast computational methods for point estimates

We describe roughly the algorithm from Bates and Pinheiro (1998), which is implemented in the R package `nlme`. Given a model like

$$y_i = x_i\beta + z_i b_i + \varepsilon_i, \quad b_i \sim N(0,\Sigma), \quad \varepsilon \sim N(0,\sigma^2 I),$$

we apply the E-M algorithm as follows.

Rewrite the covariance matrix of the random effects, Σ, as the scaled covariance matrix $D = \Sigma/\sigma^2$. This can be further decomposed into factors, $D^{-1} = \Delta'\Delta$. We then use the E-M algorithm to estimate Δ.

The likelihood of the parameters given the data is given by

$$\begin{aligned} L(\Delta|y,x,\beta,\sigma^2) &= \prod p(y_i|b_i,x_i,\beta,\Delta,\sigma^2)p(b_i|x_i,\beta,\Delta,\sigma^2) \\ &= \prod \frac{1}{\sqrt{(2\pi\sigma^2)^{n+q}|D|}} \exp\left(-\frac{1}{2\sigma^2}\left\|y_i - x_i\beta - z_i b_i\right\|^2 - \frac{1}{2\sigma^2} b_i' D^{-1} b_i\right). \end{aligned}$$

Thus, the log likelihood is

$$\log L \propto \sum \left(-\frac{1}{2}\log|D| - \frac{1}{2\sigma^2} b_i' D^{-1} b_i\right).$$

To find the expected value, we need the conditional distribution of $b|y$. Note that this is proportional to the full joint distribution, or likelihood of the parameters, as above. The exponential term can be viewed as a penalized least-squares term, and rewritten as

$$\left\|y_i - x_i\beta - z_i b_i\right\|^2 + b_i' D^{-1} b_i = \left\| \begin{bmatrix} y_i - x_i\beta \\ 0 \end{bmatrix} - \begin{bmatrix} z_i \\ \Delta \end{bmatrix} b_i \right\|^2.$$

Thus the conditional distribution is

$$b_i|y_i \sim N\left(\begin{bmatrix} z_i \\ \Delta \end{bmatrix}^{-1} \begin{bmatrix} y_i - x_i\beta \\ 0 \end{bmatrix}, \sigma^2 \begin{bmatrix} z_i \\ \Delta \end{bmatrix}^{-1} \begin{bmatrix} z_i \\ \Delta \end{bmatrix}^{-T}\right),$$

and the expected log likelihood is

$$E_{b|y} \log L = \sum -\frac{1}{2}\log|D| - \frac{1}{2\sigma^2} \mathrm{E}\left(b_i' D^{-1} b_i\right).$$

We claim that

$$\frac{1}{2\sigma^2} E\left(b_i' D^{-1} b_i\right) = tr\left(D^{-1}\operatorname{Var}(b_i/\sigma|y_i)\right) + E(b_i'/\sigma) D^{-1} E(b_i/\sigma),$$

since $E(b_i' D^{-1} b_i) = trE(b_i' D^{-1} b_i) = E(tr(b_i' D^{-1} b_i)) = E(tr(D^{-1} b_i b_i')) = tr(D^{-1}(\operatorname{Var}(b_i) + E(b_i)^2)) = tr(D^{-1}\operatorname{Var}(b_i)) + E(b_i)'\Lambda E(b_i)$. We also note that a trace and a squared

magnitude are both specific cases of the squared 2-norm, so we can combine the terms to get

$$-tr(A'D^{-1}A),$$

where

$$A = \begin{bmatrix} E(b_i|y_i)/\sigma \\ \begin{bmatrix} z_i \\ \Delta \end{bmatrix}^{-1} \end{bmatrix}.$$

Thus the expected log likelihood is

$$E_{b|y} \log L = -M \log |D| + tr(A'D^{-1}A),$$

which is maximized by $\Delta = A^{-1}/\sqrt{M}$.

Given Δ at each step, we can improve the likelihood further by maximizing the likelihood with respect to β and σ^2. We see that

$$\log L(\beta|y_i, x_i, b_i, \Delta) \propto \sum \left\| \left(\begin{bmatrix} y_i \\ 0 \end{bmatrix} - \begin{bmatrix} z_i \\ \Delta \end{bmatrix} b_i \right) - \begin{bmatrix} x_i \\ 0 \end{bmatrix} \beta \right\|^2,$$

which is a linear model, so

$$\widehat{\beta}(\Delta) = \left(\begin{bmatrix} x_1 \\ 0 \\ \cdots \end{bmatrix}' \begin{bmatrix} x_1 \\ 0 \\ \cdots \end{bmatrix} \right)^{-1} \begin{bmatrix} x_1 \\ 0 \\ \cdots \end{bmatrix} \left(\begin{bmatrix} y_i \\ 0 \\ \cdots \end{bmatrix} - \begin{bmatrix} z_i \\ \Delta \\ \cdots \end{bmatrix} b \right).$$

We also have that

$$\log L(\sigma^2|y_i, x_i, b_i, \Delta) \propto -N \log(2\pi\sigma^2) - \frac{\|K\|^2}{\sigma^2},$$

which is maximized by $\widehat{\sigma^2} = \|k\|^2/N$.

Bibliography

Aitkin, M., Anderson, D., and Hinde, J. (1981). Statistical modelling of data on teaching styles. *Journal of the Royal Statistical Society*, 144(4):419–461.

Bates, D. M. and Pinheiro, J. C. (1998). Computational methods for multilevel modelling. Bell Labs Technical Memorandum.

Campbell, J. E. (1992). Forecasting the presidential vote in the states. *American Journal of Political Science*, 36(2):386–407.

Campbell, J. E., Ali, S., and Jalalzai, F. (2006). Forecasting the presidential vote in the states, 1948–2004: An update, revision, and extension of a state-level presidential forecasting model. *Journal of Political Marketing*, 5(1–2):33–57.

Gelman, A. (2006). Multilevel (hierarchical) modeling: What it can and cannot do. *Technometrics*, 48:432–435.

Gelman, A. and Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

Ghitza, Y. and Gelman, A. (2013). Deep interactions with MRP: Election turnout and voting patterns among small electoral subgroups. *American Journal of Poltical Science*, 57:762–776.

Hibbs, D. (2008). Implications of the 'bread and peace' model for the 2008 US presidential election. *Public Choice*, 137:1–10.

Jackman, S. (2005). Pooling the polls over an election campaign. *Australian Journal of Political Science*, 40:499–517.

Lax, J. and Phillips, J. (2009). Gay rights in the states: Public opinion and policy responsiveness. *American Political Science Review*, 103:367–386.

Lock, K. and Gelman, A. (2010). Bayesian combination of state polls and election forecasts. *Political Analysis*, 18:337–348.

Merolla, J., Munger, M., and Tofias, M. (2005). In play: A commentary on strategies in the 2004 U.S. presidential election. *Public Choice*, 123:19–37.

Silver, N. (2012). *The Signal and the Noise: Why So Many Predictions Fail—but Some Don't*. Penguin Group.

Strauss, A. (2007). Florida or Ohio? Forecasting presidential state outcomes using reverse random walks. Princeton University Political Methodology Seminar.

Wang, S. (2012). About the meta-analysis (FAQ). Princeton Election Consortium.