Appendix to Accompany "Pooling the Polls Over an Election Campaign" *Australian Journal of Political Science*, 2005, V40(4):499-517.

Simon Jackman

October 2005

Let α_t be the coalition 2PP intended vote share at time t; here I let t index days, with June 18 2004, the first date in my analysis (corresponding to the field date of the first poll in my data set). Let $i = 1, \ldots, n$ index the polls available for analysis. Each poll result is assumed to be generated as follows:

$$y_i \sim N(\mu_i, \sigma_i^2) \qquad (1)$$

where y_i is the result of poll i. Each of the n polls is generated by organization j_i on field date t_i. σ_i is the standard error of the poll (a function of y_i and the poll's sample size) and

$$\mu_i = \alpha_{t_i} + \delta_{j_i} \qquad (2)$$

where δ_j is the bias of polling organization j, an unknown parameter to be estimated.

To model change in vote intentions, I use the following simple random walk model:

$$\alpha_t \sim N(\alpha_{t-1}, \omega^2), \;\; t = 2, \ldots, T \qquad (3)$$

with the distribution

$$\alpha_1 \sim \text{Uniform}(.4, .6) \qquad (4)$$

initializing the random walk (i.e., before we see any polling, I assume that coalition support is anywhere between 40% and 60%, bracketing the historical range reported earlier). In adopting this model I assume that on average, today's level of coalition support is the same as yesterday's, save for random shocks that come from a normal distribution with mean zero and standard deviation ω.

Prior Distributions

Prior distributions are a critical component of Bayesian modeling, a formal statement of the researcher's *a priori* beliefs about the model parameters. Equations 3 and 4 supply priors for the α_t parameters. For the house-effects parameters δ_j I use a vague normal prior centered at zero,

$$\delta_j \sim N(0, d^2) \tag{5}$$

with d an arbitrary large constant, so that the data will dominate inferences for these parameters. Since the δ_j are house-specific shifts on the scale of the observed survey proportions, it is reasonable to posit that δ_j = .15 would be a "very large" house effect (i.e., survey organization j is biased by 15 percentage points, on average). Accordingly, a prior distribution that had a 95% confidence interval spanning -.15 to .15 would formalize the *a priori* belief that we are reasonably ignorant about the magnitudes of house effects, but that are likely to be no larger than plus or minus 15 percentage points. In turn, this corresponds to $d^2 = (.15/2)^2 = .005625$. That is, given the scale of the data being modeled, a vague normal prior can be specified with any $d > .075$.

A prior distribution over ω completes the specification of the model: I use a uniform prior,

$$f(\omega) = \begin{cases} 100 & \text{if } 0 < \omega < .01 \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

That is, before seeing data, I presume that day-to-day changes are not massive, with the standard deviation of the daily changes no larger than one percentage point; even at this maximum prior value of $\omega = .01$ (one percentage point), the implication is that 95% of the daily changes in the α_t are no larger than plus or minus two percentage points. This seems a plausible constraint, allowing the occasional large day-to-day shift in voter sentiment, but consistent with the idea that on most days, the change in underlying voter sentiment is probably not this large.

Bayesian Inference via Markov chain Monte Carlo

Equations 1, 2 and 3 define the statistical model used for pooling and smoothing the polls, as well as estimating house effects. The unknown parameters are (1) α_t, the Coalition's vote share on day t, $t = 1, \ldots, T$ - 1; (2) δ_{j_i}, the house effect of polling organization j_i, where j indexes the set of 5 polling organizations analyzed here; (3) ω^2, a variance parameter tapping the magnitude of day-to-day variability in the α_t.

Inference for these parameters is via Bayesian methods, meaning that we compute the posterior density of the model parameters, conditional on the observed data, denoted as $\pi(\boldsymbol{\Theta}|\mathbf{y})$, where, in this case, $\boldsymbol{\Theta}$ is the set of the unknown parameters listed above, and **y** are the observed data (poll estimates and sample sizes). To compute the posterior density I rely on Markov chain Monte Carlo (MCMC) techniques, and in particular, the workhorse MCMC technique, Gibbs sampling. MCMC exploits two ideas

1. *the Monte Carlo principle*. If θ is a random variable (e.g., a parameter) then anything we wish to know about θ can be ascertained by sampling many times from its probability density $\pi(\theta)$. The sampled values of θ can be stored, summarized or plotted to produce estimates of the mean of θ, confidence intervals, etc. The quality of the estimates produced by this sampling approach is limited only by the number of samples one's computers can generate, store and summarize (and the researcher's patience).

2. *joint distributions are completely characterized by conditional distributions*. Suppose a model has several parameters, $\mathbf{\Theta} = \{\theta_1, \ldots, \theta_k\}$ which have the joint posterior density $\pi(\mathbf{\Theta}|\mathbf{y})$. Then the following iterative algorithm can be used for sampling from the joint posterior density $\pi(\mathbf{\Theta}|\mathbf{y})$. Let $\mathbf{\Theta}^{(p)}$ be the values of $\mathbf{\Theta}$ sampled at iteration p. Then $\mathbf{\Theta}^{(p+1)}$ can be generated as follows:

$$\begin{array}{ll} 1. & \text{sample } \theta_1^{(p+1)} \text{ from } f(\theta_1|\mathbf{\Theta}_{-1}^{(p)}, \mathbf{y}) \\ 2. & \text{sample } \theta_2^{(p+1)} \text{ from } f(\theta_2|\mathbf{\Theta}_{-2}^{(p)}, \mathbf{y}) \\ \vdots & \vdots \\ j. & \text{sample } \theta_j^{(p+1)} \text{ from } f(\theta_j|\mathbf{\Theta}_{-j}^{(p)}, \mathbf{y}) \end{array}$$

   This scheme "works" in a wide variety of conditions (e.g., Tierney 1996) and, remarkably, irrespective of where the iterative scheme is initialized (i.e., the value of $\mathbf{\Theta}^{(0)}$); the key result is as the number of iterations increases ($t \rightarrow \infty$) the samples can be validly regarded as samples from the joint posterior density $\pi(\mathbf{\Theta}|\mathbf{y})$.

Conditional Distributions for the Gibbs Sampler

Implementing this scheme for the model used here requires knowing the forms of the conditional distributions in the iterative scheme defined above. To derive these conditional distributions, it is helpful to recognize that if a statistical model can be expressed as a directed acyclic graph $\mathcal{G}$, the conditional distribution of node θ_j in the graph is

$$f(\theta_j|\mathcal{G}_{-\theta_j}) \propto f(\theta_j|\text{parents}[\theta_j]) \times \prod_{w \in \text{children}[\theta_j]} f(w|\text{parents}[w]), \tag{7}$$

where $\mathcal{G}_{\text{-}\theta_j}$ stands for all nodes in the graph $\mathcal{G}$ other than θ_j. See Spiegelhalter, Thomas and Best (1996) and Spiegelhalter and Lauritzen (1990) for defintions and proofs of these propositions.

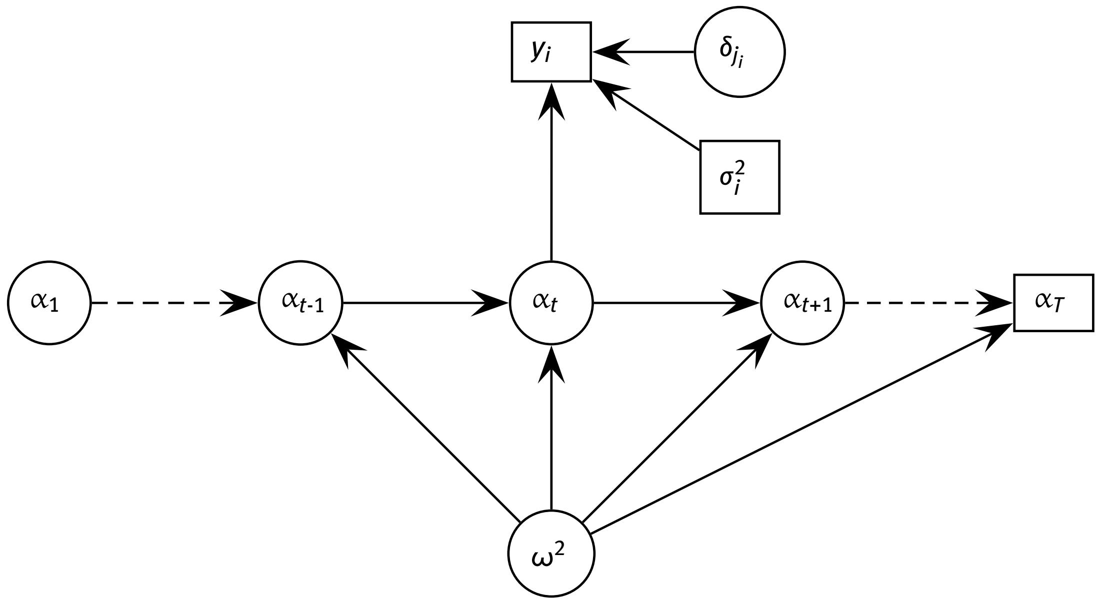

Figure 1: Directed Acyclic Graph (DAG) Corresponding to equations 1, 2 and 3. Arrows run from parent nodes to child nodes in the DAG.

The directed acyclic graph for the model in equations 1, 2 and 3 is shown in Figure 1. The graph is defined for an arbitrary three day sequence, with a poll result y_i available on day t, but not on days t - 1 or t + 1. Using a conventional notation, unknown parameters are represented as circles, and rectangles denote fixed or deterministic quantities. Parent-child relations are denoted by arrows: e.g., α_{t+1} is a child of α_t, α_t is a parent of y_i, as are δ_j and σ_t^2, and ω^2 is a parent of $\alpha_2, \ldots, \alpha_T$.

Given these parent-child relationships, and the result in equation 7, we can now proceed to derive the conditional distributions for the Gibbs sampler. Most of the relevant distributions are normal distributions (e.g., see equations 1 and 3). The following lemma, a standard result in Bayesian statistics, provides a result on normal distributions that will be used repeatedly in the derivations below:

Lemma 1 *If $z \sim N(\mu, \sigma^2)$ and $\mu \sim N(a, b^2)$, then $\mu|z \sim N(\tilde{a}, \tilde{b}^2)$, where*

$$\tilde{a} = \left[\frac{z}{\sigma^2} + \frac{a}{b^2}\right] \left[\frac{1}{\sigma^2} + \frac{1}{b^2}\right]^{-1}$$

and

$$\tilde{b} = \left[\frac{1}{\sigma^2} + \frac{1}{b^2}\right]^{-1}$$

Proof: Since z and μ have normal distributions,

$$f(z) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(z-\mu)^2}{2\sigma^2}\right] \tag{8}$$

$$f(\mu) = \frac{1}{\sqrt{2\pi b^2}} \exp\left[-\frac{(\mu-a)^2}{2b^2}\right] \tag{9}$$

Via Bayes Rule,

$$\begin{aligned}
f(\mu|z) &\propto f(z)\times f(\mu) \\
&\propto \exp\left[-\frac{(z-\mu)^2}{2\sigma^2}\right] \exp\left[-\frac{(\mu-a)^2}{2b^2}\right] \\
&= \exp\left[-\frac{(z-\mu)^2}{2\sigma^2} - \frac{(\mu-a)^2}{2b^2}\right] \\
&= \exp\left[\frac{-1}{2\,\sigma^2\,b^2}\left[(z-\mu)^2 b^2 + (\mu-a)^2\sigma^2\right]\right] \\
&= \exp\left[\frac{-1}{2\,\sigma^2\,b^2}\left(z^2b^2 + \mu^2b^2 - 2z\mu b^2 + \mu^2\sigma^2 + a^2\sigma^2 - 2\mu a\sigma^2\right)\right] \\
&\propto \exp\left[\frac{-1}{2\,\sigma^2\,b^2}\left[\mu^2(b^2+\sigma^2) - 2\mu(zb^2 + a\sigma^2)\right]\right] \\
&= \exp\left[\frac{-1}{2}\left[\mu^2\left(\frac{1}{b^2}+\frac{1}{\sigma^2}\right) - 2\mu\left(\frac{z}{\sigma^2}+\frac{a}{b^2}\right)\right]\right] \\
&= \exp\left[\frac{-1}{2}\left(\frac{1}{b^2}+\frac{1}{\sigma^2}\right)\left[\mu^2 - 2\mu\frac{\frac{z}{\sigma^2}+\frac{a}{b^2}}{\frac{1}{b^2}+\frac{1}{\sigma^2}}\right]\right] \\
&\propto \exp\left[\frac{-1}{2}\left(\frac{1}{b^2}+\frac{1}{\sigma^2}\right)\left(\mu - \frac{\frac{z}{\sigma^2}+\frac{a}{b^2}}{\frac{1}{b^2}+\frac{1}{\sigma^2}}\right)^2\right] \\
&\quad \text{(i.e., completing the square)} \\
&= \exp\left[\frac{-(\mu-\tilde{a})^2}{2\,\tilde{b}^2}\right]
\end{aligned}$$

which is proportional to a normal distribution with mean $\tilde{a}$ and variance $\tilde{b}^2$. Multiplying by the normalizing constant $(2\pi\tilde{b}^2)^{-1/2}$ gives the normal distribution $N(\tilde{a}, \tilde{b}^2)$. ■

One of the results of the lemma has a simple interpretation: a normal prior distribution over a parameter μ, when multiplied by a normal likelihood for some data z yields a normal posterior density $f(\mu|z)$ with mean equal to the precision-weighted average of the prior and the posterior.

With these results at hand, the conditional distributions that drive the Gibbs sampler are:

1. $f(\alpha_t|\mathcal{G}_{\text{-}\alpha_t})$. Via equation 3, the parents of α_t are $\alpha_{t\text{-}1}$ and ω^2. The children of α_t are (1) α_{t+1} with ω^2 appearing again as a parent (again, see equation 3) and (2) y_i, with co-parents δ_{j_i} and σ_i^2 (equation 1). Equation 7 shows that the conditional distribution for α_t is proportional to the product of the normal distributions in equations 3 (one for α_t and another for α_{t+1}) and equation 1. Application of the lemma shows that the conditional distribution for α_t is a normal distribution with mean
$$\left[\frac{y_i - \delta_{j_i}}{\sigma_i^2} + \frac{\alpha_{t\text{-}1} + \alpha_{t+1}}{\omega^2}\right]\left[\frac{1}{\sigma_i^2} + \frac{2}{\omega^2}\right]^{-1}$$
and variance
$$\left[\frac{1}{\sigma_i^2} + \frac{2}{\omega^2}\right]^{-1}.$$
The following special cases warrant elaboration:

   (a) *No surveys on day t.* Consider a sequence of three days (t - 1, t, t + 1) without any survey data available on day t. In this case the conditional distribution for α_t is simply the product of the normal distribution for α_t and the normal distribution for the child node α_{t+1}. Via the lemma,
$$\alpha_t|\mathcal{G}_{\text{-}\alpha_t} \sim N\left(\frac{\alpha_{t\text{-}1} + \alpha_{t+1}}{2}, \frac{\omega^2}{2}\right)$$
i.e, absent polling data, the model linearly interpolates between temporally adjacent observations.

   (b) *Multiple surveys on one day.* It is also possible for multiple surveys to be available on any single day. Letting s index days, define $\mathcal{P}_s = \{i : t_i = s\}$ as the subset of surveys with field date s. All $y_i, i \in \mathcal{P}_s$ are children of α_s, as is α_{s+1}. Repeated application of the lemma shows the conditional distribution of α_s to be a normal distribution with mean
$$\left[\left(\sum_{i\in\mathcal{P}_s} \frac{y_i - \delta_{j_i}}{\sigma_i^2}\right) + \frac{\alpha_{s\text{-}1} + \alpha_{s+1}}{\omega^2}\right] \cdot \left[\frac{1}{\sum_{i\in\mathcal{P}_s} \sigma_i^2} + \frac{2}{\omega^2}\right]^{-1}$$
and variance
$$\left[\frac{1}{\sum_{i\in\mathcal{P}_s} \sigma_i^2} + \frac{2}{\omega^2}\right]^{-1}.$$

Note that the effect of multiple polls is to relatively down-weight α_{t-1} in coming up with the current estimate of α_t, net of the effects of the house-effect parameters. In addition, more poll information generates more precision (smaller variances in the conditional distribution for α_t); e.g., compare the variance term obtained for the case of multiple polls with that of one poll or no polls.

(c) *The conditional distribution for α_1 (the first day).* The conditional distribution for α_1 is different from the conditional distributions for the other $\alpha_t, t > 1$. The parent of α_1 is not α_{t-1}, since there is no "day zero". Rather, the uniform prior distribution for α_1 (equation 4) serves as its parent. Since there is also a poll available for $t = 1$, the conditional distribution of α_1 is a normal distribution with mean

$$\left[\frac{y_1 - \delta_{j_1}}{\sigma_1^2} + \frac{\alpha_2}{\omega^2}\right] \cdot \left[\frac{1}{\sigma_1^2} + \frac{1}{\omega^2}\right]^{-1}$$

and variance

$$\left[\frac{1}{\sigma_1^2} + \frac{1}{\omega^2}\right]^{-1},$$

truncated to the interval (.4, .6).

2. $f(\delta_j|\mathcal{G}_{-\delta_j})$. The parent distribution of δ_j is simply its prior, $\delta_j \sim N(0, d^2)$ (introduced above as equation 5). The children of δ_j are all the poll estimates y_i published by polling organization j. Letting k index polling organizations, define $\mathcal{P}_k = \{i : j_i = k\}$ as the set of polls published by polling organization k. Then via equation 7 and applying the lemma, δ_k has a conditional distribution that is normal, with mean

$$\left[\sum_{i\in\mathcal{P}_k} \frac{y_i - \alpha_{t_i}}{\sigma_i^2}\right] \cdot \left[\frac{1}{\sum_{i\in\mathcal{P}_k} \sigma_i^2} + \frac{1}{d^2}\right]^{-1}$$

and variance

$$\left[\frac{1}{\sum_{i\in\mathcal{P}_k} \sigma_i^2} + \frac{1}{d^2}\right]^{-1}.$$

3. $f(\omega^2|\mathcal{G}_{-\omega^2})$. For tractability, I work with $\tau = g(\omega) = \omega^{-2}$; conversely, $\omega = g^{-1}(\tau) = \tau^{-1/2}$. The uniform prior over ω in equation 6 implies the restriction that $\omega^2 < .0001$ and in turn, the restriction that $\tau > 10,000$. Moreover, using

standard results on transformations of random variables (e.g., Gelman et al. 2004, 21), the uniform prior on ω implies the following prior for τ:

$$\begin{aligned} f(\tau) &= \left|\frac{\partial \omega}{\partial \tau}\right| f_\omega \left[g^{-1}(\tau)\right] \\ &= \frac{1}{2}\tau^{-3/2} p(\tau), \end{aligned}$$

where $p(\tau) = 100$ if $\tau > 10,000$ and 0 otherwise, and so

$$f(\tau) \propto \tau^{-3/2}, \tau > 10,000.$$

The children of τ are simply the children of ω^2, the α_t parameters, $t = 2, \ldots, T$. Thus, using the general formulation in equation 7, the conditional distribution of τ is

$$\begin{aligned} f(\tau|\mathcal{G}_{-\tau}) &\propto \tau^{-3/2} \prod_{t=2}^{T} f(\alpha_t; \alpha_{t-1}, \tau) \mathcal{I}(\tau > 10,000) \\ &= \tau^{-3/2} \prod_{t=2}^{T} \frac{1}{\sqrt{2\pi\omega^2}} \exp\left[\frac{-(\alpha_t - \alpha_{t-1})^2}{2\omega^2}\right] \mathcal{I}(\tau > 10,000) \\ &= \tau^{-3/2} \prod_{t=2}^{T} \frac{\tau^{1/2}}{\sqrt{2\pi}} \exp\left[-\tau\frac{1}{2}(\alpha_t - \alpha_{t-1})^2\right] \mathcal{I}(\tau > 10,000) \\ &\propto \tau^{(T-4)/2} \exp\left[-\tau\frac{1}{2}\sum_{t=2}^{T}(\alpha_t - \alpha_{t-1})^2\right] \mathcal{I}(\tau > 10,000) \end{aligned}$$

which is a Gamma distribution over τ with parameters $(T - 4)/2$ and $\frac{1}{2}\sum_{t=2}^{T}(\alpha_t - \alpha_{t-1})^2$, and where $\mathcal{I}()$ is an indicator function, evaluating to one if its argument is true, and zero otherwise (i.e., constraining to density to have support only where $\tau > 10,000$). The sampled τ can be simply transformed back into ω (i.e., $\omega = g^{-1}(\tau) = \tau^{-1/2}$, since that is the scale on which we wish to perform inference (i.e., prior beliefs about the magnitudes of day-to-day movements in α_t were formulated on the scale of the standard deviation of those movements, ω, not the inverse of the variance of the shocks τ).

The posterior density of the ω is not discussed at great length in the paper. Figure 2 shows the posterior density of ω, and the uniform prior (dotted line). The posterior has the usual χ^2 type skew one associates with a variance (or standard deviation) parameter, and is quite different from the prior. Quite clearly, given the model, the data are informative about this parameter.

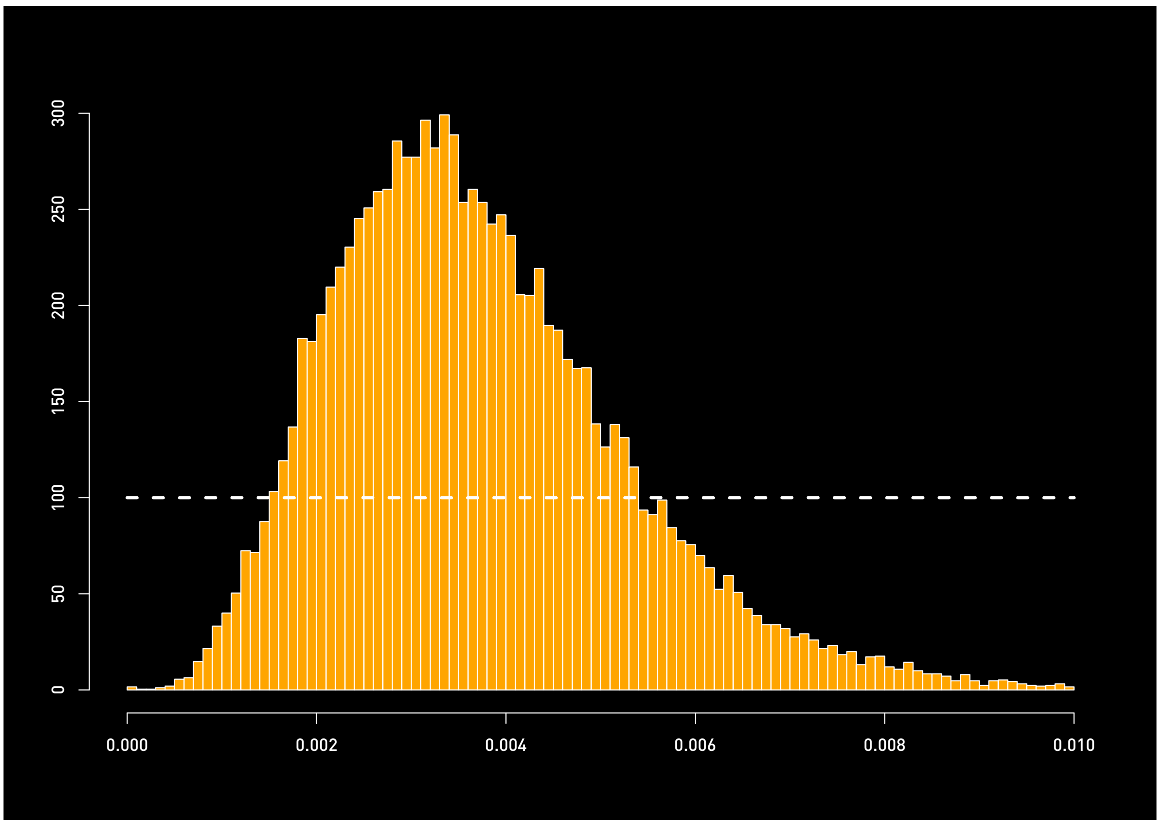

Figure 2: Posterior and Prior Density of ω.

Implementation

I use JAGS, a free, open-source program for Bayesian analysis via Gibbs sampling very similar to BUGS. The JAGS commands needed to perform the analysis described above are extremely simple:

```
model{
 ## measurement model
 for(i in 1:NPOLLS){
       mu[i] <- house[org[i]] + alpha[date[i]]
       y[i] ~ dnorm(mu[i],prec[i])
 }

 ## transition model (aka random walk prior)
 for(i in 2:NPERIODS){
        mu.alpha[i] <- alpha[i-1]
        alpha[i] ~ dnorm(mu.alpha[i],tau)
 }

 ## priors
 tau <- 1/pow(sigma,2)    ## deterministic transform to precision
 sigma ~ dunif(0,.01)     ## uniform prior on standard deviation

 alpha[1] ~ dunif(.4,.6) ## initialization of daily track

 for(i in 1:5){           ## vague normal priors for house effects
       house[i] ~ dnorm(0,.01)
 }

}
```

R files supplied in this replication archive read the raw data `read.r`. The R files `TwoPartyPreferred.r` and `firstPrefs.r`. run the JAGS jobs, for both the two-party preferred daily track and the first preferences daily track, respectively. Note that the file `jags.cmd` specifies a 1,000 iteration burn-in before a 25,000 iteration run, for testing purposes; note that the paper uses a 1,000,000 iteration burn-in followed by a 25,000,000 iteration run.

Some of the R files that generate plots may make reference to `family=dinfamily`; comment out this command, since it is a direction to use some proprietary typefaces I have installed on my machine and will generate an error on your machine.

References

Gelman, Andrew, John B. Carlin, Hal S. Stern and Donald B. Rubin. 2004. *Bayesian Data Analysis*. Second ed. Boca Raton, Florida: Chapman and Hall.

Spiegelhalter, David J., Andrew Thomas and Nicky G. Best. 1996. “Computation on Bayesian Graphical Models.” In *Bayesian Statistics 5*, ed. J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith. Oxford University Press pp. 407--425.

Spiegelhalter, David J. and S. L. Lauritzen. 1990. “Sequential updating of conditional probabilities on directed graphical structures.” *Networks* 20:579--605.

Tierney, Luke. 1996. “Introduction to general state-space Markov chain theory.” In *Markov Chain Monte Carlo in Practice*, ed. W. R. Gilks, S. Richardson and D. J. Spiegelhalter. London: Chapman and Hall pp. 59--74.